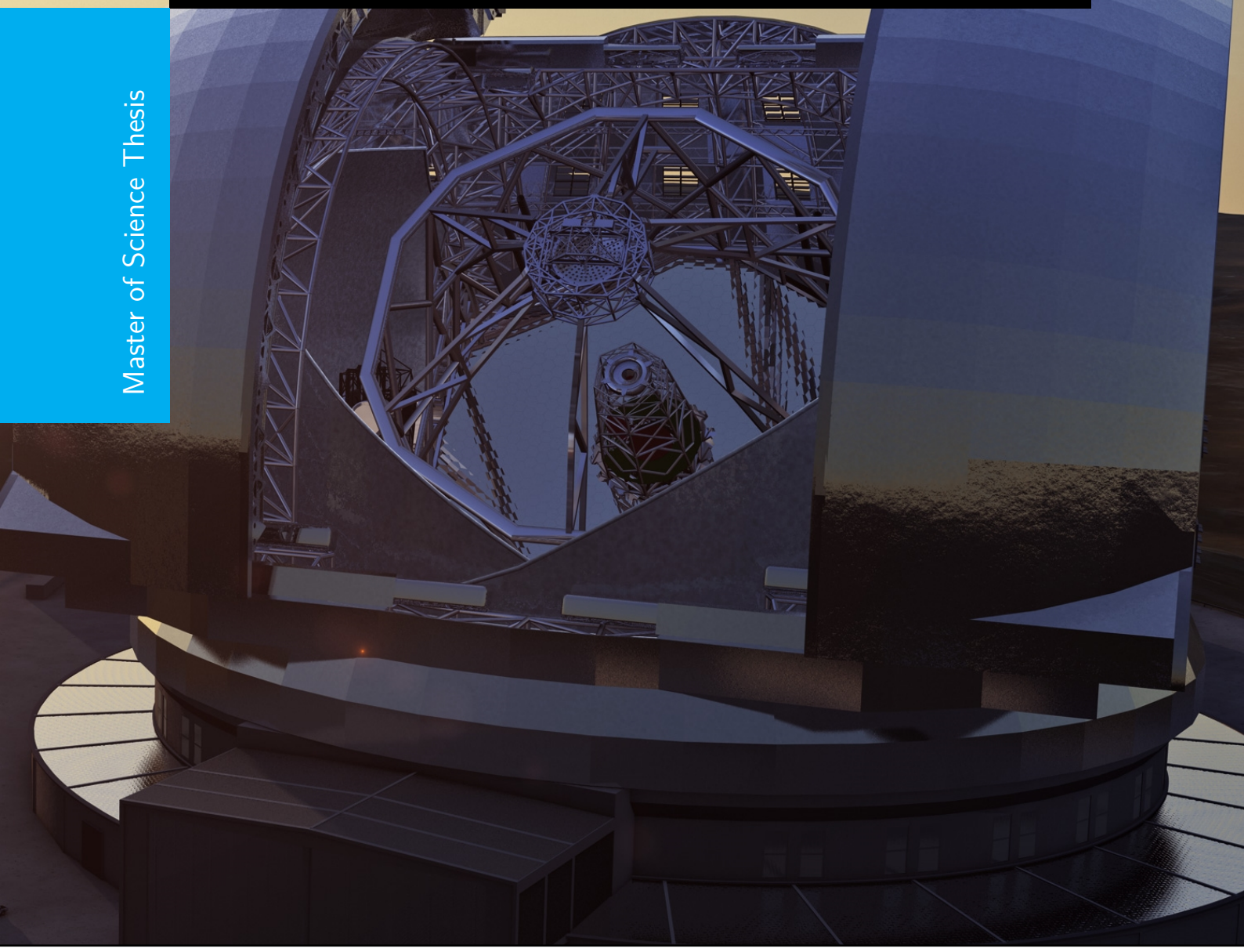
Sparse VARX Model Identification for Large-Scale Adaptive Optics

Pieter Piscaer

Master of Science Thesis



TU Delft Delft University of Technology

Delft Center for Systems and Control

Sparse VARX Model Identification for Large-Scale Adaptive Optics

Master of Science Thesis

For the degree of Master of Science in Systems and Control at Delft University of Technology

Pieter Piscaer

July 20, 2016

Faculty of Mechanical, Maritime and Materials Engineering (3mE) · Delft University of Technology

Abstract

Atmospheric turbulence is a major constraint on the resolution of ground-based astronomical telescopes. By the passage of light trough the atmosphere, a by origin flat wavefront is turned into a non-flat smooth surface. Depending on the weather conditions, turbulence limits the angular resolution of large telescopes to a telescope with a diameter of only 5 to 20 cm. Without any form of compensation, increasing the telescope diameter beyond this size will not improve the image quality any further. Adaptive Optics (AO) is a technique to compensate for wavefront aberrations introduced by turbulence in optical systems. A wavefront sensor (WFS) measures the distortions in the wavefront and a deformable mirror (DM) actively adds the opposite of the optical path length perturbations by changing the shape of its reflective surface.

Most AO systems use a simple control law consisting of a wavefront reconstruction step and a mapping onto the DM, thereby neglecting the dynamics of the turbulence. More modern optimal control strategies have been proposed lately with the ability to predict the evolution of the wavefront in the near future. These methods are an improvement over the classical approach in the compensation for temporal errors caused by the delay between measurement and correction. However, the size of AO systems is growing to the point that problems are occurring in terms of computational complexity and memory capacity. The purpose of this thesis is to find a data-driven procedure that maintains a similar performance to the optimal control methods in terms of accuracy, while having a scalable memory requirement and computational complexity.

The essential step of the strategy used to achieve this goal is an efficient identification routine of sparse VARX (Vector Auto-Regressive with eXogenous input) models for large-scale adaptive optics systems. Given the identified model, the control problem minimizes the mean squared error of the estimated wavefront distortions. The identification requires open-loop WFS data and fits the VARX coefficient matrices in a separable least-squares framework by solving it for each row of the coefficient matrices independently. The dimensionality of optimization problem is drastically reduced by assuming a rough estimation of the sparsity pattern in a graphical modelling framework based on the sensor geometry and Taylor's frozen turbulence assumption. Because of the decreased dimensionality, it was shown that the complexity of the VARX model identification algorithm scales linearly with the total number of

outputs. This complexity forms a large contrast with the standard identification of a state-space innovation model, solving a Riccati equation with a cubic complexity. The sparsity is optimized by adding an ℓ_1-norm regularization term and the resulting separable regularized least-squares problem is solved using the alternating direction method of multipliers (ADMM).

Given the identified sparse VARX model, the control strategy has been chosen to minimize the one-step-ahead prediction of the mean squared error of the phase slopes. It has been shown that the control law can be written as a large sparse least-squares problem. By exploiting the sparsity, an efficient solution is found, giving the new controller an advantage over the existing control methods. Moreover, for a minimum-variance estimator, a sparse approximation of the inverse covariance matrix of the stochastic input is required. An efficient solution to this estimation problem is found either via covariance selection or from solving a separable least-squares problem similar to the one in the identification routine.

The sparse VARX identification method is compared to a Kalman filter using accurate statistical knowledge of the system. A validation study has shown that a low-order sparse VARX approximation gives an accuracy similar to the Kalman filter. Furthermore, the added ℓ_1-regularization term is able to enhance the sparsity significantly with only a limited increase of the prediction error. However, increasing the measurement noise variance results in a decrease in performance of the VARX model with respect to the Kalman filter. The new controller has also been validated and was shown to outperform the conventional controller with a performance comparable to the Kalman filter based algorithm. Especially when the turbulence conditions get heavier and under mild noise conditions, the new method has shown to be very effective in reducing the temporal error. This combination of accuracy and scalability demonstrates the potential of the new method for large-scale AO systems.

Table of Contents

Preface

One of the main goals for me this year was to get an idea of the work and life of a researcher. The complete process of this thesis has been truly enlightening on all the obstacles and possibilities a scientist encounters during his research. Furthermore, it has given me the opportunity to learn more than I could imagine in such a short time, including a deeper understanding in topics already familiar to me as well as subjects that were completely new to me at first.

There are of course many people I would like to thank who have helped me over the past year. First of all, I would like to thank professor Verhaegen for all the help on finding the right subject for this graduation project and in particular for giving me the opportunity to be a part of his team from the day I started. I would also like to thank Chengpu Yu, for having the patience with answering the many questions and guiding me into the right directions. Likewise, my special thanks to Baptiste Sinquin. Your insights and the many useful discussions helped me to obtain the results of this thesis. Of course, I also want to thank everyone else from the CSI group for making me feel a part of the team. Last but not least, I would like to thank my parents, sister and friends who have helped, supported and motivated me in every possible way.

Chapter 1

Introduction

Atmospheric turbulence forms a major constraint on the resolution of ground-based telescopes. Mixing air of different temperatures results in local inhomogeneities in the refractive index, causing fluctuations in the optical path length. By the passage of light through the atmosphere, a by origin flat wavefront is turned into a non-flat smooth surface. When this distorted wavefront is observed, the image is blurred. Depending on the weather conditions, turbulence limits the angular resolution of large telescopes to a telescope with a diameter of only 5 to 20 cm. Without any form of compensation, increasing the telescope diameter beyond this size will not improve the image quality any further.

Adaptive Optics (AO) is a technique to compensate for wavefront aberrations in optical systems and is widely used in astronomy, microscopy and lithography. The distorted wavefront is directed by a system of lenses towards a deformable mirror (DM), able to actively add a phase correction by changing the shape of the reflective surface. The residual wavefront, the difference between the turbulence induced wavefront and the applied correction, is directed to a wavefront sensor (WFS) measuring the remaining phase aberrations. Based on the WFS measurements, the controller should determine the actuator commands to the DM in such a way that it removes the distortions.

Most AO systems use a simple control law consisting of a wavefront reconstruction step from WFS data and a mapping onto the DM. This reconstructed wavefront is assumed constant until the compensation by the DM is applied. Due to delays within the AO system and the dynamic nature of turbulence, this method tends to compensate for the distortions that occurred some time in the past rather than the actual aberration. There has been a large focus on improving this random-walk prediction model by including turbulence models either from first principles or from system identification (see e.g. Kulcsár et al., 2006; Hinnen et al., 2008). However, the new extremely large-scale generation of ground-based telescopes requires a scalable control law establishing a trade-off between a good performance in compensating for the phase distortion and a low computational complexity.

1-1 Research motivation

The scale of ground-based telescopes is growing to the point that problems are occurring in terms of computational complexity and memory capacity of conventional numerical algorithms. For example, the European extremely large telescope (E-ELT) will contain a 39.3 m primary mirror with a total number of actuators and sensors in the order of tens of thousands (Gilmozzi and Spyromilio, 2007). Existing control methods are no longer efficient enough for AO systems of these dimensions.

The main focus in the literature is either on creating a more efficient implementation of the static classical approach, or on approximating the Riccati equation which needs to be solved in the optimal estimation problem. However, the classical approach still remains static with much lower performance than the optimal control strategies. The estimation of the Riccati equation on the other hand is still relatively computationally demanding and is based on first principle models. First principle models do not provide the most accurate predictions since they are based on simplified physical laws and parameters that are difficult to determine exactly in practice. Therefore, it is preferred to use *data-driven* methods, where the model is derived from input and output data under the actual current atmospheric circumstances. By identifying the model from the input-output data, the dynamics of the model will match the one of the real system. A more detailed discussion on the limitations of classical control and the advantage of data-driven identification can be found in the study of Hinnen (2007). The main focus of this thesis will be on the extension of the data-driven optimal control framework to a scalable routine for extremely large-scale AO systems.

Typically, studying large-scale systems requires a simplification of the analysis by exploiting certain properties such as symmetry, structure or sparsity of the system matrices. When sparsity is induced, it can be exploited in later stages, drastically reducing both the computational complexity and memory requirements. One tool for creating sparsity is by estimating physical models as so-called *graphical models* (see e.g. Dahlhaus and Eichler, 2003). With an eye on recent advances in the estimation of graphical VAR (Vector Auto-Regressive) models from Gaussian time series, the idea originated to model turbulence in this framework.

In the graphical modelling framework, VAR and VARX (Vector Auto-Regressive with eXogenous input) models have the promising property that the graph topology has a one-to-one correspondence to the sparsity pattern of the VAR(X) coefficient matrices. In other words, by simple reasoning, it can be determined beforehand which elements in the matrices are surely zero and which are not. By only identifying the non-zero elements, the dimensionality of the identification problem can be reduced significantly. Furthermore, because of their relative high accuracy and low complexity, VAR models have been a popular choice in describing turbulence dynamics in the literature. This trade-off between simplicity and accuracy is one of the characteristics that the new method should have. With the aforementioned information, a concrete goal of this thesis can be presented.

1-2 Goal of the thesis

This thesis will develop a scalable data-driven control procedure that outperforms the classical methods significantly, while having a much lower computational complexity then existing

optimal control methods. The method will include an efficient routine for the identification of sparse VARX models by exploiting the spatio-temporal correlations in the wavefront in a graphical modelling approach. In essence, a trade-off is made between the loss of accuracy and the improvement in terms of computational complexity in comparison with both the classical and optimal approach.

1-3 Thesis synopsis

The remainder of the thesis is structured as follows. The theoretical background on system identification and graphical modelling is presented in Chapter 2. It can be read independently from the rest of the report and will be referred to when considered necessary. Chapter 3 contains the theory behind atmospheric turbulence and adaptive optics, but will also serve as support and motivation for the results presented afterwards. The first three sections are dedicated to the theory behind atmospheric turbulence and different modelling approaches. Afterwards, the principle of adaptive optics is introduced. The possibilities in representing sensor and actuator dynamics are discussed, providing a motivation for choosing VARX models. This chapter is concluded with the introduction of existing control strategies and a discussion on the obstacles in the control of large-scale AO systems. The novel scalable data-driven AO control method, forming the main contribution of this thesis, is presented in Chapter 4. The complete procedure, from the selection of the model to the identification algorithm and the control law, is discussed and the performance is supported in a validation study. Finally, a brief summary of the main conclusions, followed by a number of recommendations for future research is presented in Chapter 5.

Chapter 2

Theoretical Considerations

Before discussing the novel identification and control method for the atmospheric turbulence and adaptive optics application, there are a number of theoretical considerations that are important for the complete understanding of the method and results. This chapter will serve as a summary of some relevant theoretical background only and can be read separate from the rest of the report. A certain level of linear algebra and signal analysis assumed in reading this chapter. It includes a short introduction to system identification, in particular the identification of autoregressive models. Furthermore, the basics of graphical modelling is presented including the definitions of causality and conditional independence, the application in modelling multivariate time series and the covariance selection problem. All theory concerning the application, i.e. atmospheric turbulence and adaptive optics, is included in Chapter 3.

2-1 Introduction to system identification

System identification is the method of obtaining a model from given data. There exist multiple methods for system identification and it depends on many factors which method is the most suitable for a certain application. In the case of modelling turbulence, it is important that the method can address both the deterministic and stochastic part of the model. The objective is to determine a one-step-ahead predictor of the turbulence. The classical approach of solving these problems are the so called prediction error methods, that determines this one-step-ahead predictor without using knowledge of the system and stochastic disturbances. A completely different approach to this problem is referred to as subspace identification. These methods are based on retrieving subspaces that are related to the system matrices that have to be identified. The vast literature that has been written on the topic of system identification is too extensive to discuss in this thesis. A complete discussion is presented in Verhaegen and Verdult (2007). This section will be used to present the theory that is essential in understanding the remainder of this thesis.

A general formulation of a linear system with no directed feed-through is the *state-space*

model:

$$\begin{aligned} x(k+1) &= Ax(k) + Bu(k) + w(k) \\ y(k) &= Cx(k) + v(k) \end{aligned} \tag{2-1}$$

where $w(k) \sim \mathcal{N}(0, C_w)$ and $v(k) \sim \mathcal{N}(0, C_v)$ are the process noise and measurement noise respectively. Moreover, $x(k) \in \mathbb{R}^n$ is the state of the system, $u(k) \in \mathbb{R}^m$ is a (deterministic) input and $y(k) \in \mathbb{R}^p$ is the measured output. One different formulation of the state-space model that is of particular interest to system identification is the so-called *innovation form representation*. By writing the stochastic part in terms of the innovation sequence $e(k) = y(k) - C\hat{x}(k|k-1)$ and by introducing the *Kalman gain* K, the output $y(k)$ can be written as

$$\begin{aligned} \hat{x}(k+1|k) &= A\hat{x}(k|k-1) + Bu(k) + Ke(k) \\ y(k) &= C\hat{x}(k|k-1) + e(k) \end{aligned} \tag{2-2}$$

The identification of the innovation representation amounts to finding the matrices A, B, C and K from input-output data.

Another type of model that is central to this thesis is the *VARX (Vector Auto-Regressive with eXogenous input) model*

$$y(k) = \sum_{i=1}^{n_a} A_i y(k-i) + \sum_{i=1}^{n_b} B_i u(k-i) + w(k) \tag{2-3}$$

with stochastic input $w(k) \sim \mathcal{N}(0, C_w)$ and a deterministic input $u(k)$. In contrast to the innovation form, the VARX model only represents the input-output relations. An important property that relates the two models is that each stable model of type (2-1) can be represented as a (higher order) VARX model (2-3).

In the following paragraphs, the identification of these type of models is introduced. Two main classes of methods can be used for this purpose: the prediction error and subspace identification methods. Moreover, special attention will be paid to enforcing sparsity onto the system matrices. Another sub-problem in system identification is the identification of purely stochastic models, which are obtained when modelling atmospheric turbulence.

2-1-1 Prediction error methods

Prediction error (PE) methods form a class of methods estimating the parameters in an LTI model and estimates both the deterministic and the stochastic part of the model. If we assume a set of input-output sequences on a certain time interval and return to the innovation model (2-2), PE methods can be used to identify parametrized representations of the matrices A, B, C, and K. The main step is the minimization of a desired cost function with the parameters θ as optimization variables. A common way of doing this is inspired by the minimum-variance state-reconstruction property of the Kalman filter. The predictor $\hat{y}(k, \theta)$ follows directly from:

$$\begin{aligned} \hat{x}(k+1, \theta) &= A(\theta)\hat{x}(k, \theta) + B(\theta)u(k) + K(\theta)\left(y(k) - C(\theta)\hat{x}(k, \theta)\right) \\ \hat{y}(k, \theta) &= C(\theta)\hat{x}(k, \theta) \end{aligned}$$

A property of the Kalman filter is that the variance of the prediction error $y(k) - \hat{y}(k,\theta)$ is minimized. Hence, when we denote the optimal parameter θ_0, we know it should minimize

$$\theta_0 = \arg\min_{\theta} \ \text{trace}\left(E\left[(y(k) - \hat{y}(k,\theta))(y(k) - \hat{y}(k,\theta))^T\right]\right) \tag{2-4}$$

Under the assumption that the variance is constant, the prediction error is ergodic and we may approximate this by

$$\min_{\theta} \frac{1}{N} \sum_{k=0}^{N-1} \|y(k) - \hat{y}(k \mid k-1,\theta)\|_2^2 \tag{2-5}$$

This linear least-squares problem is referred to as the *prediction-error estimation problem* and forms the basis of PE methods. The relation (2-4) and in particular the approximation (2-5) is important for the remainder of this report.

2-1-2 Subspace identification

Subspace identification methods retrieve certain subspaces related to the system matrices of the state-space model. These subspaces are used to determine the system of matrices up to a similarity transformation. In subspace identification, the model is not parametrized and it can be obtained from simple linear algebra problems such as RQ factorization, SVD and linear least-squares, instead of optimization methods.

The relationship between the inputs, states and outputs forming the basis of subspace identification is the so-called *data equation*. For a model of the form (2-2), the data equation reads:

$$Y_{i,s,N} = \mathcal{O}_s X_{i,1,N} + \mathcal{T}_s U_{i,s,N} + \mathcal{S}_s E_{i,s,N}$$

Where $Y_{i,s,N}$, $U_{i,s,N}$ and $E_{i,s,N}$ are block Hankel matrices storing the identification data, defined as:

$$Y_{i,s,N} = \begin{bmatrix} y(i) & y(i+1) & \cdots & y(i+N-1) \\ y(i+1) & y(i+2) & \cdots & y(i+N) \\ \vdots & \vdots & \ddots & \vdots \\ y(i+s-1) & y(i+s) & \cdots & y(i+s+N-2) \end{bmatrix} \tag{2-6}$$

with $s > n$ and similar definitions for $U_{i,s,N}$, $E_{i,s,N}$ and $X_{i,1,N}$. Furthermore, $\mathcal{O}_s$ represents the extended observability matrix and $\mathcal{T}_s$ and $\mathcal{S}_s$ are two block lower-triangular Toeplitz matrices of the form

$$\mathcal{O}_s = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{s-1} \end{bmatrix} \quad \mathcal{S}_s = \begin{bmatrix} I & 0 & \cdots & 0 & 0 \\ CK & I & \cdots & 0 & 0 \\ CAK & CK & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ CA^{s-1}K & CA^{s-2}K & \cdots & CK & I \end{bmatrix} \quad \mathcal{T}_s = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ CB & 0 & \cdots & 0 & 0 \\ CAB & CB & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ CA^{s-1}B & CA^{s-2}B & \cdots & CB & 0 \end{bmatrix}$$

The solution to the subspace identification problem starts with the estimation of the column space of the matrix $\mathcal{O}_s$. From this subspace, the matrices C and A can directly be extracted from an SVD up to a similarity transformation. The matrix B follows from a least-squares problem or an RQ factorization and the Kalman gain is computed using a Riccati equation. The complete procedure is described in Chapter 9 of Verhaegen and Verdult (2007) and is too extensive to include in this chapter.

2-1-3 Enforcing sparsity

There are multiple ways of enforcing sparsity in a system during identification. When the sparsity pattern is known beforehand, the most straightforward method is to set all desired elements in the system matrices to zero as constraints in the optimization procedure. When the pattern is not known, adding an ℓ_1-norm ($\|x\|_1 = \sum_{i=1}^{n} |x_i|$) regularization term to the objective function is a common way of enforcing sparsity (Donoho, 2006). The regularized optimization problem becomes

$$\min_x \; f(x) + \lambda \|x\|_1$$

where $f(x)$ is the desired objective function and λ is a tuning parameter to enforce different degrees of sparsity. When λ is increased, the sparsity in the matrix x will be larger than for small λ. When a certain group sparsity needs to be enforced, i.e. a number of matrices that should have the same sparsity pattern, it was shown in Yuan and Lin (2006) that the regularization should be a sum of ℓ_2-norms ($\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$), better known as "Group lasso".

In the literature, there have been many applications of ℓ_1-regularization including soft-thresholding (Donoho, 1995), the lasso (Tibshirani, 1996), compressed sensing (Candès et al., 2006) and learning of sparse graphical models (Meinshausen and Bühlmann, 2006; Songsiri and Vandenberghe, 2010). Section 2-2 will present the basics of graph theory and how it can be exploited to identify sparse VAR models.

2-1-4 Stochastic realization

The stochastic realization problem is the problem of finding a stochastic linear model from the covariance matrices or time series measurements of the output only. This sub-problem in system identification can be solved using either prediction error or subspace identification methods. Since it is in general an extensive field, it is chosen not to go into detail on how to solve the problem in general. More information on linear stochastic systems can be found in many textbooks, for example the very recent book of Lindquist and Picci (2015). The next section will revolve around the identification of AR models. Since they do not have a deterministic input, this will be an example of stochastic realization. Moreover, the turbulence models that will be used to simulate turbulence dynamics in the numerical simulations are created following a stochastic realization approach since they are created from a certain theoretical autocovariance sequence, see for example Assémat et al. (2006) and Beghi et al. (2008). When there is no theoretical expression available of the autocovariance matrices, they can be approximated from time series measurements. Assuming that there is a time sequence of N output vectors available, the sample covariance matrix of $E[y(k+i)y(k)^T]$ is constructed via

$$S_i = \frac{1}{N-i} \sum_{k=1}^{N-i} y(k+i) y^T(k) \tag{2-7}$$

2-1-5 Identifying vector auto-regressive models

A type of models that is very suited for modelling turbulence (Massioni et al., 2015), is the *Vector Auto-Regressive (VAR) model*. This section will provide a brief overview of common

methods for estimating VAR models and its material can be found in many textbooks. A VAR model is defined as

$$y(k) + \sum_{i=1}^{q} A_i y(k-i) = w(k) \tag{2-8}$$

with output $y(k) \in \mathbb{R}^p$, stochastic input $w(k) \sim (0, C_w)$ and q the order of the VAR model. The matrices $A_i \in \mathbb{R}^{p\times p}$ are referred to as the *coefficient matrices* of the VAR model. Equivalently, the VAR model can be written as

$$\sum_{i=0}^{q} B_i y(k-i) = n(k)$$

with $n(k) \sim (0, I)$, which can, because of the identity noise covariance matrix, be more suitable for some applications. The coefficient matrices of both types are related via $B_0 = C_w^{-1/2}$ and $B_i = C_w^{-1/2} A_i$ for $i = 1, \ldots, q$.

Identifying VAR models is an example of the stochastic realization problem, i.e. the problem of finding A_i and C_w given a time series of $y(k)$ or its autocovariance sequence:

$$C_i = E\left[y(k+i)y^T(k)\right]$$

for $i = 0, \ldots, q$. When there is no theoretical expression available of the autocovariance matrices, they can be approximated using (2-7). A common way of solving the identification problem is using the so-called *Yule-Walker equations*. These equations are obtained if the transpose of (2-8) is post-multiplied by $y(k-i),\ i = 1, \ldots, q$ and after taking the expectation of both sides. When $y(k)$ is real, $C_{-i} = C_i^T$ and the following set of linear equations is found

$$\begin{bmatrix} C_0 & C_1 & \cdots & C_q \\ C_1^T & C_0 & \cdots & C_{q-1} \\ \vdots & \vdots & \ddots & \vdots \\ C_q^T & C_{q-1}^T & \cdots & C_0 \end{bmatrix} \begin{bmatrix} I \\ A_1^T \\ \vdots \\ A_q^T \end{bmatrix} = \begin{bmatrix} C_w \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{2-9}$$

These equations are usually referred to as Yule-Walker equations or normal equations. By estimating the auto-covariance sequence from time series data, the Yule-Walker equations can be used to derive the model parameters A_i and C_w.

Another way of finding the AR coefficients $A_1, \ldots, A_q$ is by minimizing the mean squared prediction error. The prediction error $e(k)$ is the difference between the measured output and the output according to the model, i.e.

$$e(k) = y(k) - \hat{y}(k) = y(k) + \sum_{i=1}^{q} A_i y(k-i)$$

The prediction-error estimation problem (2-5) for this example is given by

$$\min_{A_i} \quad \sum_{k=q+1}^{N} \|y(k) - \sum_{i=1}^{q} A_i y(k-i)\|_2^2$$

Note that this is completely equivalent to the following formulation

$$\min_{A_1,\ldots,A_q} \left\| \begin{bmatrix} I & A_1 & A_2 & \cdots & A_q \end{bmatrix} \begin{bmatrix} y(q+1) & y(q+2) & \cdots & y(N) \\ y(q) & y(q-1) & \cdots & y(N-1) \\ \vdots & \vdots & \ddots & \vdots \\ y(1) & y(2) & \cdots & y(N-q) \end{bmatrix} \right\|_2^2 \tag{2-10}$$

This problem can be solved for each row of $\begin{bmatrix} A_1 & A_2 & \cdots & A_q \end{bmatrix}$ separately. The mean squared error can also be expressed in terms of the autocovariance sequence. It is straightforward to show that the optimization problem (2-10) can be written as

$$\min_{A_1,\ldots,A_q} \operatorname{tr}\left(\begin{bmatrix} I & A_1 & \cdots & A_q \end{bmatrix} \begin{bmatrix} C_0 & C_1 & \cdots & C_q \\ C_1^T & C_0 & \cdots & C_{q-1} \\ \vdots & \vdots & \ddots & \vdots \\ C_q^T & C_{q-1}^T & \cdots & C_0 \end{bmatrix} \begin{bmatrix} I \\ A_1^T \\ \vdots \\ A_q^T \end{bmatrix} \right)$$

if the autocovariance matrices are replaced by the sample covariance matrices of (2-7). Moreover, it can be shown that with this representation of the covariance matrix, a similar least-squares problem in terms of $B_0, \ldots, B_q$ actually finds the maximum likelihood estimate. Both the maximum likelihood problem and the least-squares problem can easily be solved by setting the gradient of the cost function to zero. In both cases, it can be proven that this results in solving the Yule-Walker equations (see e.g. Songsiri et al., 2009).

2-2 Graphical modelling

This section will present a brief introduction to graphical modelling. Graphical models are widely used in a large variety of scientific fields. They offer insight in the structure of distributions and can exploit sparsity to improve the efficiency of statistical calculations. Throughout this report, a graph will be represented as the triplet $G = (V, A, E)$, where V is a finite set of *vertices*, $A \subseteq V \times V$ is the set of (directed) *arcs* $E \subseteq V \times V$ the set of (undirected) *edges*. Directed arcs represent a causality relation such as Granger causality, whereas undirected edges represent a conditional uncorrelatedness, often using the definitions of conditional orthogonality or conditional independence.

Often in graphical modelling only linear association and linear Granger non-causality is considered. This non-causality can be expressed in terms of conditional orthogonality. For random vectors x, y and z, x and y are conditionally orthogonal if x and y are uncorrelated after the linear effects of z have been removed. Conditional orthogonality is denoted as $x \perp y \mid z$. When non-linear relations are considered this can be extended by considering conditional independence instead. For Gaussian processes the two are identical. Since this thesis is restricted to Gaussian processes, the notation defined above will also be used to denote conditional independence.

Next, two important definitions for causality and conditional independence are discussed that form the basis of graphical models. More information on graphical models can be found in many textbooks, for example Lauritzen (1996).

2-2-1 Granger causality

Directed edges in graphical models denote *causality*. A general definition of causality was defined by Granger (Granger, 1980):

Considering the stationary processes $x(t)$ and $y(t)$, then $y(t)$ is said to cause $x(t+1)$ if

$$\Pr(x(t+1) \in A \mid \Omega(t)) \neq \Pr(x(t+1) \in A \mid \Omega(t) - y(t)) \text{ for some } A$$

where $\Omega(t)$ contains all knowledge at time t. In other words, $y(t)$ causes $x(t+1)$ only if the variable $y(t)$ has some unique information about what the value of $x(t+1)$ will be. Since it is impossible to know all knowledge in the universe at time t, this definition on itself cannot be tested on actual data. For this reason Granger introduced a number of constraints. Consider the proper information set $J(t)$ available at time t, consisting of terms of the vector series $z(t)$. Now suppose that $z(t)$ includes $x(t)$, but not $y(t)$ and let $J'(t)$ denote the augmented set $\{z(t-j), y(t-j), j \geq 0\}$. Furthermore, $F(x(t+1) \mid J'(t))$ is the conditional distribution function of $x(t+1)$ given $J'(t)$. Two operational definitions follow:

$y(t)$ does *not cause* $x(t+1)$ with respect to $J'(t)$ if

$$F(y(t+1) \mid J(t)) = F(x(t+1) \mid J'(t))$$

and if

$$F(x(t+1) \mid J(t)) \neq F(y(t+1) \mid J'(t))$$

then $y(t)$ is said to be a *prima facie cause* of $x(t+1)$ with respect to $J'(t)$.

This is clearly equivalent to the general definition when $J'(t) = \Omega(t)$. An important difference between the general and the operational definitions is that using the operational definitions, true causality is impossible to prove since there can always be data missing from $J(t)$ that influence the value of $x(t+1)$.

2-2-2 Conditional independence

A missing undirected edge between two vertices often represent *conditional independence*, which can be formalized as follows:

If x, y and z are random variables, we say that x is conditionally independent of y given z (denoted as $x \perp y \mid z$) if,

$$\Pr(x \cap y \mid z) = \Pr(x \mid z) \Pr(y \mid z)$$

In other words: given that z occurs, knowledge of y is irrelevant for the likelihood of x and vice versa.

Let $x \sim \mathcal{N}(0, \Sigma)$ be an n-dimensional Gaussian random variable. It can be shown that conditional independence of two elements of x (x_i and x_j), corresponds to zero entries in the inverse of the covariance matrix (Dempster, 1972), (Lauritzen, 1996). Suppose x is partitioned into components y and z such that its covariance matrix is

$$\Sigma_x = \begin{bmatrix} \Sigma_{yy} & \Sigma_{yz} \\ \Sigma_{zy} & \Sigma_{zz} \end{bmatrix}$$

where $y = (x_i, x_j)$ are the two components of interest and let $K := \Sigma_x^{-1}$. The conditional density is proportional to the joint density of y and z. Exploiting that z is fixed it can be found that

$$\begin{aligned} f(y \mid z) &\propto \exp\{\begin{bmatrix} y^T & z^T \end{bmatrix} \begin{bmatrix} \Sigma_{yy} & \Sigma_{yz} \\ \Sigma_{zy} & \Sigma_{zz} \end{bmatrix}^{-1} \begin{bmatrix} y \\ z \end{bmatrix}\} \qquad (2\text{-}11) \\ &\propto \exp\{-y^T K_{yy} y/2 - y^T K_{yz} z\} \\ &= \exp\{-y^T K_{yy} y/2 - y^T K_{yy} K_{yy}^{-1} K_{yz} z\} \\ &\propto \exp\{-(y - K_{yy}^{-1} K_{yz} z)^T K_{yy} (y - K_{yy}^{-1} K_{yz} z)/2\} \qquad (2\text{-}12) \end{aligned}$$

Which is a normal distribution function with mean $K_{yy}^{-1} K_{yz} z := \mu_{y|z}$ and covariance $K_{yy}^{-1} := \Sigma_{y|z}$. Furthermore, use that

$$\begin{bmatrix} \Sigma_{yy} & \Sigma_{yz} \\ \Sigma_{zy} & \Sigma_{zz} \end{bmatrix}^{-1} = \begin{bmatrix} K_{yy} & K_{yz} \\ K_{zy} & K_{zz} \end{bmatrix} = \begin{bmatrix} \left(\Sigma_{yy} - \Sigma_{yz}\Sigma_{zz}^{-1}\Sigma_{zy}\right)^{-1} & -\left(\Sigma_{yy} - \Sigma_{yz}\Sigma_{zz}^{-1}\Sigma_{zy}\right)^{-1} \Sigma_{yz}\Sigma_{zz}^{-1} \\ \star & \star \end{bmatrix}$$

such that

$$\begin{aligned} K_{yy}^{-1} &= \Sigma_{y|z} = \Sigma_{yy} - \Sigma_{yz}\Sigma_{zz}^{-1}\Sigma_{zy} \qquad (2\text{-}13) \\ K_{yy}^{-1} K_{yz} z &= \mu_{y|z} = -\Sigma_{yz}\Sigma_{zz}^{-1} z \end{aligned}$$

Restoring the original partition, it is found that

$$\Sigma_{y|z} = K_{yy}^{-1} = \begin{bmatrix} \left(\Sigma^{-1}\right)_{ii} & \left(\Sigma^{-1}\right)_{ij} \\ \left(\Sigma^{-1}\right)_{ji} & \left(\Sigma^{-1}\right)_{jj} \end{bmatrix}^{-1} \qquad (2\text{-}14)$$

where the subscript ij denotes the element on row i column j, proving that x_i and x_j are conditionally independent given x_z if and only if

$$\left(\Sigma^{-1}\right)_{ij} = \left(\Sigma^{-1}\right)_{ji} = 0 \qquad (2\text{-}15)$$

This property was first introduced by Dempster (1972) and forms the basis of the *covariance selection* problem. That is, the problem of finding the maximum likelihood estimate of the inverse covariance matrix (Σ^{-1}) of a multivariate Gaussian variable $\mathcal{N}(0, \Sigma)$, subject to conditional independence constraints as in (2-15). This problem is already very well-established in the literature and was more recently extended to the time series case (Dahlhaus, 2000). Further extensions to VAR model identification have been proposed by Songsiri et al. (2009). Section 2-2-4 will give a quick literature overview on efficient algorithms to find the maximum likelihood estimate of the inverse covariance matrix.

2-2-3 Graphical models for multivariate time series

In this paragraph, the methods proposed in Dahlhaus and Eichler (2003) will be discussed. In here, three different types of graphical models for multivariate time series are defined. The first class represents each time index of a certain variable as a separate vertex, resulting in a generalization of classical graphical models such as the *time series chain graph*. In the

second class the vertices do not consist of variables at different time instances, leading to mixed graphs termed *Granger causality graphs*. Furthermore, there are the so called *partial correlation* graphs, that are generalizations of the classical covariance selection models to the time series situation (Dahlhaus, 2000). These three types of graphs will be shortly discussed in the following. More information can be found in Dahlhaus and Eichler (2003) and the references therein.

The notation that will be used in this section is as follows. Let $x = \{x_i(t), t \in \mathbb{Z}, i = 1, \ldots, n\}$ be an n-variate stationary process and $V = \{1, \ldots, n\}$ the set of all indices. For any $A \subseteq V$ we define $x_A = \{x_A(t)\}$. Furthermore, $\bar{x}_A(t) = \{x_A(s), s < t\}$ denotes the past of x_A at time t. A backslash notation is used to exclude elements from certain sets, e.g. $x_{V\setminus i}(t)$ is used to exclude node i from the process $x_V(t)$.

Time series chain graphs

The first approach is an extension of the classical chain graph. In this type of graphs, each time sample of the stochastic process x are represented by separate vertices.

The mixed graph $G = (V_{TS}, A_{TS}, E_{TS})$ of a stationary process x with $V_{TS} = V \times \mathbb{Z}$ and

$$(i, t-\tau), (j, t) \notin A_{TS} \Leftrightarrow \tau \leq 0 \text{ or } x_i(t-\tau) \perp x_j(t) \mid \bar{x}_V(t) \backslash \{x_i(t-\tau)\}$$
$$(i, t), (j, t) \notin E_{TS} \Leftrightarrow x_i(t) \perp x_j(t) \mid \bar{x}_V(t) \cup \{x_{V\setminus\{i,j\}}(t)\}$$

is called a *time series chain graph* (TSC-graph).

In this type of graphs, the undirected edges represent the conditional dependence of the vertices at the same time instance. The directed edges represent a causal relation between nodes at different time instances. Both edges are shift invariant with respect to time.

In the case of VAR models, each causality relation is represented by one entry in one of the matrices. Denoting i, j-th element of the q-th order VAR model coefficient matrices A_τ, $\tau = 1, \ldots, q$ by $(A_\tau)_{ij}$, it can be found that

$$(i, t-\tau), (j, t) \in A_{TS} \Leftrightarrow \tau \in \{1, \ldots, q\} \text{ and } (A_\tau)_{ji} \neq 0$$

and the undirected edges are seen back in the inverse of the covariance of the noise K:

$$(i, t), (j, t) \in E_{TS} \Leftrightarrow K_{ij} \neq 0$$

with K_{ij} denoting the i, j-th element in the matrix K. Only when there are missing arcs or edges the entries above will be zero (Dahlhaus and Eichler, 2003).

Granger causality graphs

This type of graph uses the notion of Granger causality (Granger, 1969) for directed arcs and the same definition of undirected edges as for time series chain graphs.

The mixed graph $G = (V, A_C, E_C)$ of a stationary process x such that for all $(i, j) \in V$

$$(i, j) \notin A_C \Leftrightarrow x_j(t) \perp \bar{x}_i(t) \mid \bar{x}_{V\setminus\{i\}}(t)$$
$$(i, j) \notin E_C \Leftrightarrow x_i(t) \perp x_j(t) \mid \bar{x}(t), x_{V\setminus\{i,j\}}(t)$$

is called a *Granger causality (GC) graph.*

A missing directed edge form i to j thus means that $x_i(t)$ and $x_j(t-k)$ are conditionally independent (or conditionally orthogonal) for all $k \geq 1$ given all the other relevant past information.

Applying this definition and using the same notation as before, it can be derived that for VAR models

$$(i,j) \notin A_C \Leftrightarrow (A_k)_{ij} = 0 \quad \forall k \in \{1,\dots,p\}$$
$$(i,j) \notin E_C \Leftrightarrow (K)_{ij} = 0$$

Hence a missing directed edge between two nodes j and i, i.e. x_i is not Granger-caused by x_j, creates a zero entry in all the coefficient matrices and thus eventually will provide a sparse model when the topology is sparse. The main difference in the VAR example between this definition and the TSC-graph is the fact that TSC-graphs can have a different sparsity pattern for each matrix A_i, while GC-graphs have a certain group sparsity, i.e. all matrices A_i have the same sparsity pattern.

Conditional independence graphs

Conditional independence graphs (or partial correlation graphs) only contain undirected edges. The idea was introduced first in Dahlhaus (2000).

An undirected graph $G = (V, E)$ is termed a *partial correlation graph* of a multivariate stationary process x when

$$(i,j) \notin E \quad \Leftrightarrow \quad x_i \perp x_j \mid x_{V\setminus\{i,j\}}$$

More precisely, an edge between i and j is missing if and only if $x_i(t)$ and $x_j(s)$ are uncorrelated for all $t, s \in \mathbb{Z}$ after removing all the linear effects of all other components $x_{V\setminus\{i,j\}}$.

Previously, it was shown that for Gaussian signals, conditional independence corresponds to zeros in the inverse of the covariance matrix of that signal. For partial correlation graphs there exists a similar characterization in terms of the inverse of the spectral matrix $S(\omega)$ of the process. In Dahlhaus (2000) it is proven that for $S(\omega)$ full rank for all ω, the conditional independence $x_i \perp x_j \mid x_{V\setminus\{i,j\}}$ holds if and only if the i,j-th element of the spectral matrix, $S^{-1}_{ij}(\omega) = 0$ for all ω. With this relation mind, the inverse covariance selection problems can be extended to the time-series case, by enforcing the sparsity pattern on the inverse spectrum instead of the inverse covariance.

2-2-4 The covariance selection problem

Consider a the problem of estimating the covariance matrix C of a certain n-variate Gaussian distributed sample dataset:

$$x(1),\dots,x(N) \sim \mathcal{N}(0, C)$$

The maximum likelihood estimation of this unknown covariance matrix from its sample covariance matrix, $S := \frac{1}{N}\sum_{k=1}^{N} x(k)x^T(k)$, follows from the following optimization problem:

$$\hat{C} := \arg\max_{C\succ 0} \; -\log\det C - \mathrm{tr}(SC^{-1})$$

By a change of variable: $X = C^{-1}$, the problem becomes:

$$\hat{C} = \arg\min_{X\succ 0} \; \log\det X + \mathrm{tr}(SX)$$

Now, assume that the signal $x(k)$ is represented by a *Gaussian graphical model*, an undirected graph describing conditional (in)dependence relations. As was shown in Section 2-2-2, defining the topology of a Gaussian graphical model, is equivalent to enforcing a certain sparsity on the inverse of the covariance matrix. This sparsity can be enforced on X either directly as equality constraints or by adding an ℓ_1-regularization term to the objective function. The problem of finding the topology of the graphical model, or equivalently the sparsity pattern in C^{-1}, is called the *covariance selection* problem and it can be formulated as:

$$\min_{X\succ 0} \; -\log\det X + \mathrm{tr}(CX) + \lambda\|X\|_1 \tag{2-16}$$

This notation was first presented by Banerjee et al. (2008). In this work, two new methods were introduced that can solve Gaussian processes with at least one thousand nodes in which a block coordinate descent method is applied to the dual and Nesterov's optimal gradient methods too a smoothed approximation of the ML objective. In Lu (2010), Nesterov's method is applied to the dual problem and is compared to a projected spectral gradient method. Yuan and Lin (2007) and Li and Toh (2010) solve the problem using interior point algorithms. Duchi et al. (2008) uses the gradient projection method to solve the dual problem. Friedman et al. (2008) uses the fact that block coordinate descent algorithms can be interpreted as an iterative penalized regression, called graphical LASSO (GLASSO). Mazumder and Hastie (2012) propose a primal algorithm, DP-GLASSO, that also operates by block coordinate descent. Furthermore, in Scheinberg et al. (2010) and Yuan (2012), ADMM is used to solve the covariance selection problem. It is shown that ADMM outperforms a number other algorithms discussed above. In Wiesel and Hero III (2012), a distributed estimation of the inverse covariance matrix is considered that is implemented using ADMM. Another recent approach is via second order Newton methods and is studied in Wang et al. (2010), Hsieh et al. (2011) and Dinh et al. (2013). In Hsieh et al. (2011), a second order algorithm is introduced that performs Newton steps using iterative quadratic approximations of the Gaussian negative log-likelihood, called QUIC. This algorithm has been shown to be able to solve problems up to $n = 20.000$. In Hsieh et al. (2013) an adaptation to this algorithm, named BIG-QUIC, is presented and is used solve 1 million dimensional ℓ_1-regularized Gaussian MLE problems. Other extremely high-dimensional algorithms have been proposed in Wang et al. (2013) and Treister and Turek (2014). Wang et al. (2013) presents a generalization of ADMM to multiple blocks, called PDMM, that randomly updates some blocks in parallel, behaving like randomized block coordinate descent. In Treister and Turek (2014), a new block coordinate descent approach is presented that defines the blocks as subsets of columns of the inverse covariance function and solves each block sub-problem by a quadratic approximation.

Chapter 3

Atmospheric Turbulence and Adaptive Optics

Adaptive optics is a well established technique to compensate for wavefront aberrations introduced by light propagation through a turbulent medium. Before commencing a study of designing a new adaptive optics (AO) control method, a basic understanding of the underlying statistical properties of the turbulence is essential. As the turbulence is largely a random process and changes quickly, it is a challenging phenomenon to model and correct.

In general, all physical phenomena, hence including turbulence, are represented by physical laws that are by nature essentially deterministic. This means that in principle we may be able to model it analytically, provided that we have at our disposal all relevant information of the atmosphere. It is however imaginable that this model would require an infeasibly large number of equations and prior knowledge. The stochastics of turbulence on the other hand are much easier to describe. In 1941, Kolmogorov proposed a theory to describe these properties (Kolmogorov, 1941a,b).

Kolmogorov's theory and everything it has produced will be considered in this chapter. Afterwards, the principle of AO and currently accepted control strategies are reviewed. In this way, the main purpose of this chapter is to serve as a theoretical support and motivation of Chapter 4, in which the main contribution of this thesis, a novel data-driven scalable control method, is proposed.

3-1 The need of AO in astronomy

Nowadays, ground-based telescopes are probably the most important tools in astronomy and many discoveries can be appointed to improving resolutions. Over the years, the telescopes have become bigger and bigger and will grow to even larger scales. The European Extremely Large Telescope (Gilmozzi and Spyromilio, 2007), which is currently being built, will be the largest telescope on earth with a primary mirror of 39.3 meter and tens of thousands actuators

and sensors. There are two main reasons for this drive of increasing the telescope size, namely the light collecting power and angular resolution. In principle, the resolution of each optical system is limited by diffraction caused by the aperture of the telescope. An incoming plane wave that is focussed by a circular lens will not be focussed to a spot, but forms a so-called airy-disk with a certain diameter d. When we would consider two point-sources, the images will become two airy-disks. When the sources are moved closer to each other, the disks will start to overlap to the point that it becomes one. The Rayleigh criterion,

$$\sin\theta \approx 1.22\frac{\lambda}{D}$$

describes the *angular resolution* θ at which the two sources still can be resolved. Hence increasing the diameter will lead to a smaller resolution.

However, as the diameter gets larger and larger, the limiting factor will be the aberrations introduced by atmospheric turbulence rather than diffraction. Turbulence is ultimately caused by the heat from solar radiation, which causes movements within the atmosphere. By mixing air of different temperatures, local inhomogeneities in the refractive index are created. A fluctuation in refractive index causes a so called *optical path difference*:

$$\Delta l = \int n(z)dz$$

where z represents the direction of the light propagation and $n(z)$ the refractive index along this line. When the function $n(z)$ is not equal for all paths that are collected with our telescope, some parts will be delayed more than others. A by origin perfectly flat wavefront hence gets distorted and will no longer be flat after it passes the earth's atmosphere. In general turbulence conditions, the angular resolution is limited to approximately one *arcsecond* ($\approx 5\mu$rad). Returning to the Rayleigh criterion, this corresponds to a telescope diameter in the order of 10cm. Increasing the diameter of the aperture on ground-based telescopes to larger dimensions will not improve the image quality any further. This limiting factor stresses the need for adaptive optics. Only by taking counter measures in the form of adaptive optics to compensate for the atmospheric wavefront distortions, increasing the diameter towards extremely-large telescopes will actually improve the image quality. In order to design such an adaptive optics system, a good understanding of atmospheric turbulence is crucial. The next section will discuss the statistical properties following from Kolmogorov's insights.

3-2 Kolmogorov theory

Kolmogorov (Kolmogorov, 1941a,b) explains turbulence using an energy cascade principle. He suggested that in turbulent flow, the energy in the large inhomogeneities (or *eddies*) is transferred into smaller and smaller inhomogeneities. The characteristic size of the largest structures is defined as the *outer scale* L_0. If the size of the smallest eddies get smaller than a certain *inner scale* l_o, the energy is dissipated as friction between molecules. The range between the inner and outer scale is called the inertial subrange. Typical values for the inner scale are 1mm to 10mm, the outer scale ranges from 10m up to 100m.

3-2-1 Spatial structure of atmospheric turbulence

In Kolmogorov's analysis, he shows that from dimensional analysis, the average speed of turbulent eddies v must be related to the size of the eddies r via

$$v \propto r^{1/3}$$

If we assume turbulence to be homogeneous and isotropic, we are only interested in the relative properties. To this intent, a *structure function* is defined describing the variance of the difference in a certain function $f(x)$ between two points separated by a distance r:

$$D(r) = \langle |f(x) - f(x+r)|^2 \rangle \tag{3-1}$$

with $\langle \dots \rangle$ representing the ensemble average of over different realizations of $f(x)$ and is related to the covariance $C(r) = \langle f(r_1), f(r_2) \rangle$ via

$$D(r) = 2(\sigma^2 - C(r)) \tag{3-2}$$

where σ^2 is the variance. A particularly interesting property is the refractive index within the medium. It was shown before how fluctuations in the refractive index inflict an optical path difference resulting in aberrations. These fluctuations are related to the velocity fluctuations such that the structure function that describes the difference in refractive index is of the form

$$D_n(r) = \langle |n(x) - n(x+r)|^2 \rangle = c_n^2 r^{2/3}$$

with c_n the refractive index structure coefficient (Roddier, 1981). This structure function is a measure of the strength of the turbulence and is valid only within the inertial subrange $l_o \leq r \leq L_0$.

From the structure function, the power spectral density (PSD) can be derived. It is usually expressed as a function of the spatial wavenumber $\kappa = 2\pi/l$, where l is the size of the fluctuations. Tatarski et al. (1961) showed that the Kolmogorov power spectrum can be found to be

$$\Phi_n^K(\kappa) = 0.033 c_n^2 \kappa^{-11/3} \tag{3-3}$$

and is only valid within the inertial subrange. There are however other models that are also accurate at scales smaller than the inner scale and larger than the outer scale. For example, the Von Kármán spectrum accounts for the outer scale

$$\Phi_n^{VK}(\kappa) = \frac{0.033 c_n^2}{(\kappa^2 + \kappa_0^2)^{-11/6}} \tag{3-4}$$

and the modified Von Kármán spectrum also for the inner scale

$$\Phi_n^{mVK}(\kappa) = \frac{0.033 c_n^2 \exp(-\kappa^2/\kappa_m^2)}{(\kappa^2 + \kappa_0^2)^{-11/6}} \tag{3-5}$$

where $\kappa_m = 5.92/l_0$ and $\kappa_0 = 2\pi/L_0$ (Schmidt, 2010). Note that when $l_0 = 0$ and $L_0 = \infty$, (3-4) and (3-5) collapse to the Kolmogorov PSD, (3-3).

3-2-2 Phase distortion through turbulence

To go from fluctuations in the refractive index to an aberration in terms of phase delays, the wavefront phase deformation can be expressed in terms of the integral over the optical path length z in the direction of the light propagation and the wave number:

$$\phi = k\Delta l = k \int n(z)dz \tag{3-6}$$

where $k = 2\pi/\lambda$ is the wavenumber of light with wavelength λ. In the same way as before, a structure function can be constructed as

$$D_\phi(r) = \langle |\phi(x) - \phi(x+r)|^2 \rangle \tag{3-7}$$

Assuming Kolmogorov's structure function, it has been found that

$$D_\phi(r) = 6.88 \left(\frac{r}{r_0}\right)^{5/3} \tag{3-8}$$

(see e.g. Roddier, 1999), where r_0 is the *Fried parameter* (Fried, 1965)

$$r_0 = \left[0.42 \frac{k^2}{\cos(\gamma)} \int c_n^2(h)dh\right]^{-3/5} \tag{3-9}$$

with γ the zenith angle of the source and c_n the refractive index structure coefficient as function of the height h above the ground. The Fried parameter gives an intuitive measure of turbulence strength. The ratio of the telescope diameter and the Fried parameter D/r_0 represents the severity of the distortions due to turbulence. The Fried parameter can be interpreted as the diameter over which turbulent effects begin to decrease the resolution: a diameter $D > r_0$ has a similar effective resolution as a telescope with $D = r_0$. In other words if $D < r_0$, the resolution is limited by diffraction according to Rayleigh's criterion and if $D > r_0$, it will be limited by the atmospheric turbulence. Another important property of r_0 is that the root mean square phase distortions over an area with diameter r_0 is about 1 radian. The Fried parameter is depending heavily on the weather conditions and ranges from 5cm in heavy turbulence during daytime to 20cm under quiet circumstances at night.

From the structure function (3-8), the phase PSD can be calculated to be

$$\Phi_\phi^K(\kappa) = 0.49 r_0^{-5/3} \kappa^{-11/3} \tag{3-10}$$

according to Kolmogorov's theory and the (modified) Von Kármán PSDs are

$$\Phi_\phi^{VK}(\kappa) = 0.49 r_0^{-5/3} (\kappa^2 + \kappa_0^2)^{-11/6} \tag{3-11}$$

and

$$\Phi_\phi^{mVK}(\kappa) = 0.49 r_0^{-5/3} (\kappa^2 + \kappa_0^2)^{-11/6} \exp(-\kappa^2/\kappa_m^2) \tag{3-12}$$

respectively. A visualization of the Kolmogorov, Von Kármán and modified Von Kármán phase PSDs is given in Figure 3-1. The influence of the inner and outer scales are clearly visible. Moreover, note that the highest energy is in the lower frequency range.

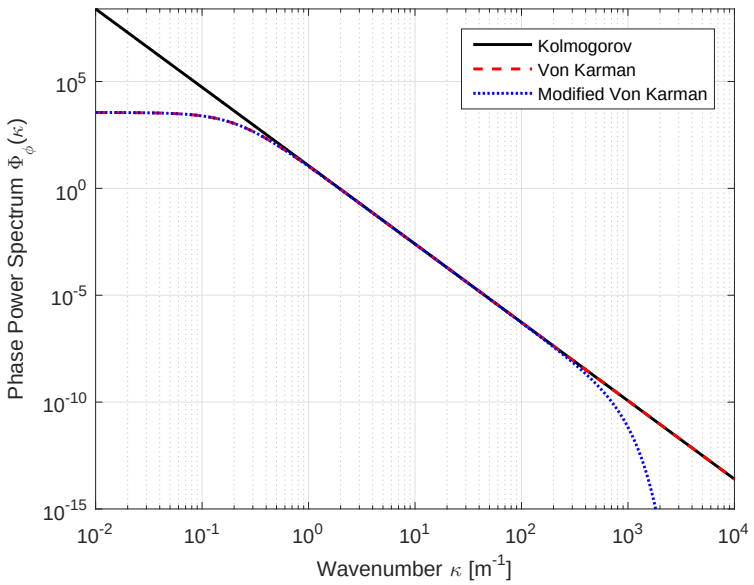


Figure 3-1: Kolmogorov, Von Kármán and modified Von Kármán phase PSDs

Throughout this thesis, we assume the Von Kármán spectrum as the underlying turbulence spectrum. The Von Kármán structure function can according to (3-2) equivalently be written as the difference between the variance and covariance of the two points as a function of the distance between them. This can be computed from the analytical expression for the spatial covariance and has been derived in Conan (2008):

$$C_\phi(r) = \frac{\Gamma(11/6)}{2^{5/6}\pi^{8/3}} \left(\frac{48\pi r \Gamma(6/5)}{5L_0}\right)^{5/6} \left(\frac{r_0}{L_0}\right)^{-5/3} \mathcal{K}_{5/6}\left(2\pi\frac{r}{L_0}\right) \tag{3-13}$$

and

$$\sigma_\phi^2 = \frac{\Gamma(11/6)\Gamma(5/6)}{\pi^{8/3}} \left(\frac{24}{5}\Gamma(6/5)\right)^{5/6} \left(\frac{r_0}{L_0}\right)^{-5/3} \tag{3-14}$$

with $\mathcal{K}(\cdot)$ the modified Bessel function of the second kind and $\Gamma(\cdot)$ the gamma function. The equations (3-13) and (3-14) will later be used to generate dynamic turbulence in simulation by exploiting the spatio-temporal correlations.

3-2-3 Layered turbulence model

If we consider more complex scenarios, it might not be possible to accurately describe the statistics of the wavefront in closed form (Schmidt, 2010). A common technique for mathematical simplification is to model the turbulence by a superposition of a number of discrete layers. Each phase screen can be seen as a model for the atmosphere at a certain height. The i^{th} phase screen is the model of a part of the atmosphere from a distance z_{i-1} to z_i above the ground. Let such a phase screen be denoted by $\psi_i(x, y, t)$, where (x, y) represents a certain

point on a plane and t the time instance. The total phase aberration equals

$$\phi(x,y,t) = \sum_{i}^{l} \psi_i(x,y,t) \tag{3-15}$$

When all layers are assumed to have the same spatial statistical properties, i.e. they are described by the same covariance function, it can be assumed that $l = 1$ without loss of generality (Beghi et al., 2008). Throughout this report, this assumption is assumed to hold such that we only need to consider one layer. The next section will discuss turbulence dynamics and how the spatio-temporal correlations can be exploited to model it.

3-3 Temporal behaviour of turbulence

The Kolmogorov model describes the spatial properties of the turbulence only. The temporal evolution is usually described by the Taylor hypothesis of frozen turbulence (Taylor, 1938). This section will present a number of methods used in the literature to model and simulate turbulence dynamics.

3-3-1 The frozen flow assumption

Taylor's frozen flow hypothesis is based on the layered turbulence model and assumes that each layer moves with a constant wind speed and in a certain direction. Since the change of the turbulent refractive index is assumed to be much slower than the the time needed for the layers to cross the telescope aperture, all layers can be seen as a frozen phase screens. The temporal behaviour of the wavefront of a single layer can thus be described entirely by the wind transport. This makes it possible to express the temporal relation of the phase at a point (x,y) as a spatial one:

$$\phi_i(x,y,t+\tau) = \phi_i(x - v_x\tau, y - v_y\tau, t) \tag{3-16}$$

where v_x and v_y represent the velocity of the wind in x- and y-direction and τ is a certain time (see also Figure 3-2). A typical windspeed $v = \sqrt{v_x^2 + v_y^2}$ is around 10 m/s, with peak values up to 50 m/s. By using (3-8) after substituting $r = v\tau$, the temporal phase structure function is obtained for Kolmogorov turbulence. Moreover, Greenwood (1977) showed that the temporal error due to the turbulence movements caused the the bandwidth specification of the AO system is a function of the so called *Greenwood frequency*:

$$f_G = 0.427\frac{v}{r_0} \tag{3-17}$$

f_G typically has a value around 25 Hz under normal turbulence circumstances (e.g. $r_0 = 0.15$m and $v = 10$m/s). The Greenwood frequency, and in particular the ratio Greenwood to sampling frequency, is an important measure in adaptive optics to describe the influence of the turbulence dynamics on telescope image and will be used in control performance validation experiments.

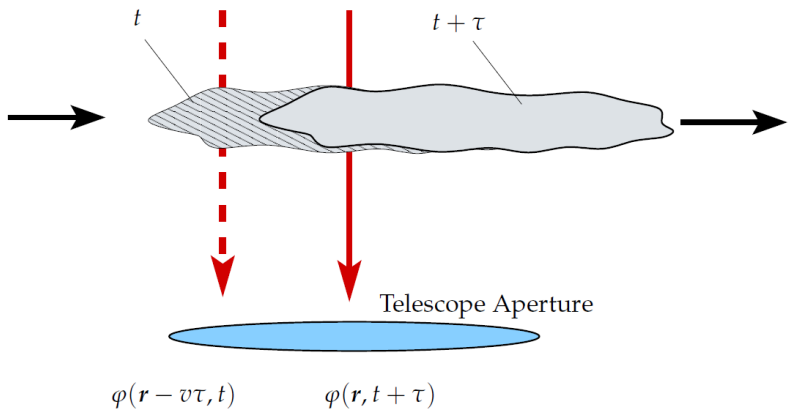


Figure 3-2: Visualization of the Taylor frozen flow hypothesis (Hinnen, 2007).

3-3-2 Simulating atmospheric turbulence

Atmospheric turbulence is a random process and so is the phase distortion that it causes. Consequently, turbulence models only give statistical properties like the power spectrum, making the simulation of atmospheric turbulence the problem of drawing individual realizations from a random process. In other words, generating a phase screen on a two dimensional grid that has the same PSD, structure function and covariance as the theory.

Usually, the phase is reconstructed using a weighted sum of basis functions. Common basis functions include Zernike polynomials and Fourier series. Other methods first generate a dynamic turbulence model and use this model to simulate the turbulence. Many methods for generating atmospheric phase screens can be found in the literature that focus either on accuracy or computational efficiency. Since the main focus of this thesis is on large-scale AO systems, the computational efficient methods are of most interest. A number of recent examples in the literature include Massioni et al. (2015), Assémat et al. (2006), Beghi et al. (2008) and Sriram and Kearney (2007). For other methods that focus on accuracy or flexibility, see (Schmidt, 2010, Sec. 9.3) and the references therein. In this paragraph, a number of these methods are discussed and the choice of simulation method that will be used in Chapter 4 is motivated.

Moving frozen phase screen realization

Probably the most straightforward approach is to construct a very large realization of the phase screen according to a theoretical power spectrum (e.g. the Von Kármán spectrum (3-11)) and dragging it over the aperture in a certain wind direction and speed. The realization can be constructed by writing the optical phase as a Fourier series

$$\phi(x,y) = \sum_{n=\infty}^{\infty} \sum_{m=\infty}^{\infty} c_{n,m} \exp\left(i2\pi(f_{x_n}x + f_{y_m}y)\right) \tag{3-18}$$

where f_{x_n} and f_{y_m} are discrete spatial frequencies in $x-$ and $y-$direction and $c_{n,m}$ are the Fourier coefficients. By randomly drawing the coefficients from a Gaussian distribution and using the inverse Fourier transform, the phase screen is synthesized.

A pleasant feature of this method is the fact that it creates a phase screen that is periodic and continuous across its opposite edges. This makes it possible to "roll" the screen around its opposite edges to continue simulation when the end of the screen is reached. By simulating multiple layers moving in different speeds and directions, the screens can be rolled around without relapsing into a turbulence sequence that has occurred before in the same simulation.

A drawback of this method is that the FFT method is often not able to represent low-order frequencies accurately enough, while the PSDs such as (3-11) have much of their power in the low spatial frequencies (recall Figure 3-1). A solution that has been proposed in Schmidt (2010) is to add so-called sub-harmonics to the FFT solution. This is, considering only a 3-by-3 grid of frequencies and repeat the above FT procedure a number of times, adding up each of these low-frequency realizations. Afterwards, this sum of screens is added to the higher-frequency solution to represent the turbulence at both higher and low-order spatial frequencies. Nevertheless, since the continuity across the edges is lost, the solution without sub-harmonics sometimes the more convenient alternative for turbulence simulators (for example in Reeves (2015)).

Turbulence as a first order VAR model

In a number of recent works (see Massioni et al. (2015) and the references therein), the dynamics of each turbulence layer is modelled by a first (or sometimes second) order VAR (Vector Auto-Regressive) model. For a first order VAR model, the turbulence vector $\phi(k)$ evolves in time according to

$$\phi(k+1) = A\phi(k) + w(k) \tag{3-19}$$

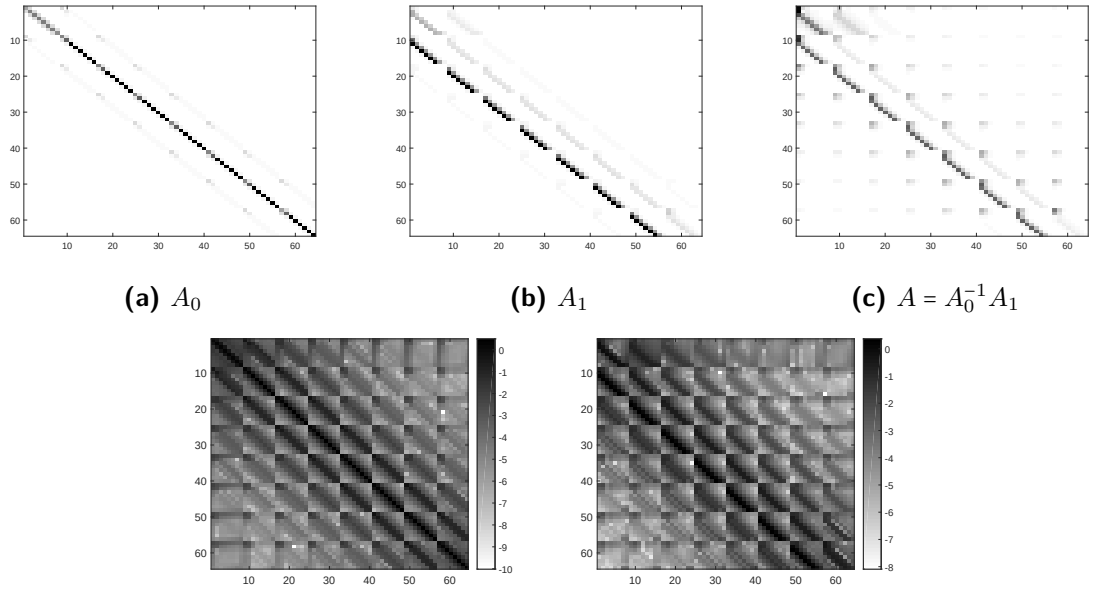
with $w(k) \sim \mathcal{N}(0, C_w)$ a white process noise, which should be consistent with the chosen spatial covariance matrix $C_\phi = E[\phi(k)\phi^T(k)]$, via the relation

$$C_w = C_\phi - AC_\phi A^T \tag{3-20}$$

where C_ϕ is defined according to the Kolmogorov or Von Kármán's theory and the sample grid geometry, e.g. (3-13). So by ensuring (3-20), the turbulence realization $\phi(k)$ is in accordance with the theoretical spatial properties (in our case the Von Karman PSD (3-11)).

The matrix A will describe the dynamics of the turbulence. In Massioni et al. (2015), the matrix A is chosen as diagonal $A = aI$ (with $|a| < 1$, usually around 0.99). The parameter a can be chosen different for each turbulence layer, simulating a different dynamics such as wind speed. This approach assumes a decoupling of the spatial and temporal dynamics where the frozen flow assumption showed otherwise. Therefore, the dynamics are not very satisfying in the sense that it does not follow the Taylor frozen flow hypothesis and only considers for each element of $\phi(k)$ the dependence on itself rather than all its surrounding neighbours. This assumption is only relatively accurate for very slow wind speeds. For larger wind speeds, the spatio-temporal correlation cause A to become multi-banded rather than diagonal. A method that will obtain such a model is discussed next.

As an alternative to defining A diagonal, one could estimate a VAR model of the form (3-19) by solving the stochastic realization problem given the theoretical covariance matrix. As discussed in Chapter 2, we can solve the Yule-Walker equations (2-9) to obtain A and C_w. In



(a) A_0 **(b)** A_1 **(c)** $A = A_0^{-1}A_1$

(d) $\log_{10}(A_0)$ **(e)** $\log_{10}(A_1)$

Figure 3-3: Example of the matrices obtained from (3-22) for the simulation of a small 8-by-8 grid with an arbitrary wind direction. Darker pixels indicate a higher absolute value of the corresponding entry in the matrix. (d) and (e) are the same matrices as (a) and (b), but now on a logarithmic scale ($\log_{10}$)

this way, the dynamics can be simulated more accurately compared to assuming A diagonal. The Yule-Walker equations for this example read

$$\begin{bmatrix} C_\phi & C_{\phi,1} \\ C_{\phi,1}^T & C_\phi \end{bmatrix} \begin{bmatrix} I \\ -A^T \end{bmatrix} = \begin{bmatrix} C_w \\ 0 \end{bmatrix}$$

where C_ϕ is the covariance matrix $E[\phi(k)\phi^T(k)]$ that follows from (3-13) and the grid geometry and $C_{\phi,1} = E[\phi(k+1)\phi^T(k)]$ can be derived using the relation (3-16). First, the second block row is solved such that

$$A = C_{\phi,1}C_\phi^{-1} \tag{3-21}$$

Next, the first block row together with (3-21) becomes (3-20) and is used to compute C_w. Moreover, we could write the model (3-19) as

$$A_0\phi(k+1) = A_1\phi(k) + n(k) \tag{3-22}$$

with $n(k) \sim \mathcal{N}(0, I)$, $A_0 = C_w^{-1/2}$ and $A_1 = C_w^{-1/2}A$ such that (3-22) and (3-19) are equivalent. An example of the matrices for a certain layer and wind properties A_0 and A_1 are depicted in Figure 3-3(a)-(b). Evidently, $A = A_0^{-1}A_1$ in Figure 3-3(c) is in general not a diagonal matrix.

Looking at the matrices A_0 and A_1 it seems that the we have already obtained a high sparsity. However, the white pixels in Figure 3-3(a)-(c) are not exactly zero. When we would plot the logarithm of the absolute values of the entries, we can see this very clearly, as illustrated

in Figure 3-3(d)-(e). Moreover, truncation might lead to a small shift in the poles of the system. Turbulence models tend to have poles close to the unit circle such that truncation might result in instability. Hence, in order to receive an exact sparse matrix, it has to be implemented within the identification method, e.g. by estimating the VAR model via least squares as discussed in Chapter 2 with an ℓ_1-norm regularization term.

Identifying a general VAR model

The methodology of the previous paragraph can be extended to identify a general VAR model of an arbitrary order and this essentially has been done in the work of Assémat et al. (2006). By interpreting this as a special form of the stochastic realization problem (Beghi et al., 2008), it can be seen as the identification of the following state-space formulation:

$$\begin{aligned} x(k+1) &= Ax(k) + Bn(k) \\ y(k) &= Cx(k) \end{aligned} \tag{3-23}$$

where $n(k) \sim \mathcal{N}(0, I)$ and $x(k)$ is obtained by piling the vectors $\{\phi(k), \dots, \phi(k-q+1)\}$, with q the order of the VAR model to be identified and

$$A = \begin{bmatrix} A_{1:q-1} & A_q \\ I_{(q-1)n} & 0 \end{bmatrix} \qquad B = \begin{bmatrix} \tilde{B}^T & 0 & \cdots & 0 \end{bmatrix}^T \qquad C = \begin{bmatrix} I_n & 0 & \cdots & 0 \end{bmatrix}$$

with $A_{1:q-1} = [A_1 \ \dots \ A_{q-1}]$ and A_i the i^{th} VAR coefficient matrix. If the top block row of A is denoted by $\tilde{A}$, it can be solved via:

$$\tilde{A} = \begin{bmatrix} C_\phi & C_{\phi,1} & \cdots & C_{\phi,q} \end{bmatrix} \bar{C}_\phi^{-1} \tag{3-24}$$

where $C_{\phi,i} = E[\phi(k+i)\phi^T(k)]$ and $\bar{C}_\phi$ denotes the large $q \times q$ block-Toeplitz matrix:

$$\bar{C}_\phi = \begin{bmatrix} C_\phi & C_{\phi,1} & \cdots & C_{\phi,q-1} \\ C_{\phi,1} & C_\phi & \cdots & C_{\phi,q-2} \\ \vdots & \vdots & \ddots & \vdots \\ C_{\phi,q-1} & C_{\phi,q-2} & \cdots & C_\phi \end{bmatrix}$$

Note the equivalence of (3-24) as an extension of (3-21). The matrix $\tilde{B}$ follows from solving

$$\tilde{B}\tilde{B}^T = C_\phi - \tilde{A}\bar{C}_\phi\tilde{A}^T \tag{3-25}$$

and can be extracted for example via SVD. This systematic approach can efficiently create a q^{th}-order VAR model that satisfies our desired statistical properties. Because of its high efficiency and sufficient accuracy, this method is used in the simulations of Chapter 4 to generate the turbulence realizations. However, it should be briefly noted that this method is not suitable for the efficient identification of a sparse VAR model (the main goal of Chapter 4). The matrix C_ϕ is generally dense and cannot be enforced to be sparse. Furthermore, the large centralized dense inverse will become an obstacle at larger dimensions.

Stochastic realization of state-space innovation model

For heavy and more complex turbulence (i.e. for many layers with diverse speeds and directions), a state-space model might be more suitable as the VAR model would be of a too large order. One method that identifies a state-space innovation model is the method of Beghi et al. (2008). This approach is again based on the fact that the spatial covariance according to the Von Kármán theory (3-13), together with the Taylor assumption (3-16), can be used to compute $E[\phi(k+i)\phi^T(k)],\ i = 0, 1, 2, \ldots$. Next, it identifies a dynamic model of the form:

$$\begin{aligned} x(k+1) &= Ax(k) + Ke(k) \\ \phi(k) &= Cx(k) + e(k) \end{aligned} \tag{3-26}$$

which can be used to generate the turbulence afterwards. The method proposed by Beghi that identifies the matricess A, B, C and K is a form subspace identification. By exploiting the fact that a large low-rank Hankel matrix can be constructed from the temporal output covariance matrices, the system matrices are extracted in closed form using SVD.

With the obtained model (3-26), a realization of the turbulence can be derived by generating a random white noise $e(k)$ and simulating the system. Advantages of this simulation technique is that it is much more efficient than the moving frozen phase screen and can describe more accurately complex turbulence models compared the identified VAR models of Assémat et al. (2006). However, compared to the method of Assémat, the complexity and memory requirements are still high. To avoid computational problems with the generation of turbulence during the simulations of the next chapter, the method of Assémat is preferred to Beghi's method.

3-4 Introduction to adaptive optics

Section 3-1 already discussed the need of adaptive optics in astronomy. Since atmospheric turbulence will even in quiet weather conditions become the limiting factor on the resolution, large telescopes will not work without adaptive optics to compensate for the turbulence effects. This section will discuss the principle of adaptive optics and its main components. The control problem will be treated separately in Section 3-6.

Adaptive Optics (AO) is a technique to compensate for wavefront aberrations in optical systems. A schematic representation of an AO system containing its main components is shown in Fig. 3-4. By the passage of light through the atmosphere, a by origin flat wavefront is turned into a non-flat smooth surface. This distorted wavefront is directed by a system of lenses towards a *deformable mirror* (DM) that actively adds the optical path difference of opposite phase by changing the shape of its reflective surface. The corrected (residual) beam is split by a beam splitter such that one beam is focused in a science camera, while the other is directed to a *wavefront sensor* (WFS). The measurements from the WFS are used as input for the controller that regulates the shape of the DM. One fundamental complication in AO control is that the WFS measures the slope of the residual wavefront rather than the residual wavefront itself. Since the performance is measured in terms of the mean squared error of the phase, there is usually an extra *wavefront reconstruction* step.

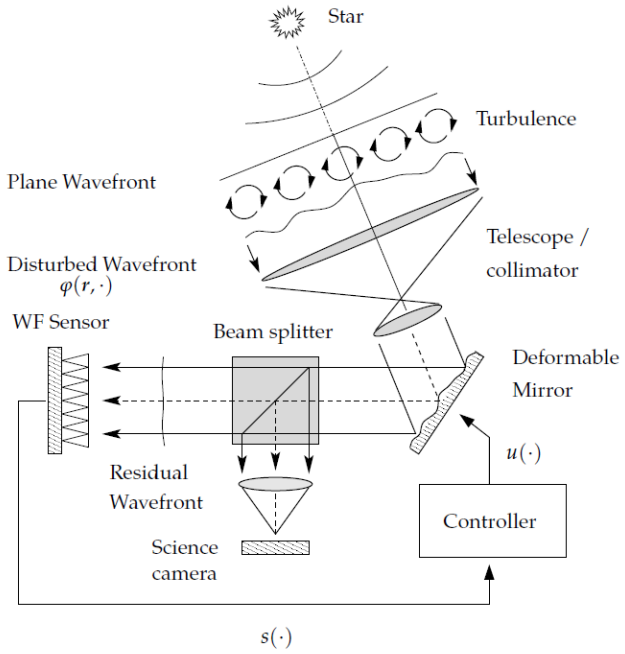


Figure 3-4: Schematic representation of an AO system (Verhaegen et al., 2015)

3-4-1 Wavefront sensor

An often used WFS in AO is the Shack-Hartmann sensor. This type of sensor consists of small lenses that partition the light on a regular grid. Each lens focusses a part of the total wavefront on a photon sensor. The position of this image can be used to calculate the local gradient at this point. Figure 3-5 illustrates this concept. The distance between the measured centroid and the reference position in x- and y-direction (Δx and Δy) are linearly related to the local slopes of the wavefront in x- and y-direction respectively. Consequently, the WFS gives information on the gradient of the wavefront at a number of sample points rather than measuring the phase directly. The link between the gradient and the phase sample points at the corners of the subaperture is illustrated in Figure 3-6 and can be written as

$$s = \frac{1}{2}\begin{bmatrix}(\phi_b + \phi_d) - (\phi_a + \phi_c)\\ (\phi_a + \phi_b) - (\phi_c + \phi_d)\end{bmatrix} \tag{3-27}$$

By following this reasoning for each subaperture, the relation between the phase $\phi(k) \in \mathbb{R}^n$, or strictly speaking the residual wavefront denoted by $\epsilon(k)$, and slopes $s(k) \in \mathbb{R}^p$ can be formulated in one linear expression:

$$s(k) = G\epsilon(k) + v(k) \tag{3-28}$$

where $G \in \mathbb{R}^{p\times n}$ is a matrix representing the relations (3-27) for all spatial sample coordinates and $v(k) \sim \mathcal{N}(0, C_v)$ is the measurement noise. Note that there are two modes, the piston mode ($\phi_a = \phi_b = \phi_c = \phi_d$ in Figure 3-6) and the waffle mode ($\phi_a = -\phi_b = -\phi_c = \phi_d$), that are invisible to the sensor since both result in a zero output of (3-27). For this reason, G will have a rank deficiency of 2 and the null-space of this matrix will consist of both unobservable modes.

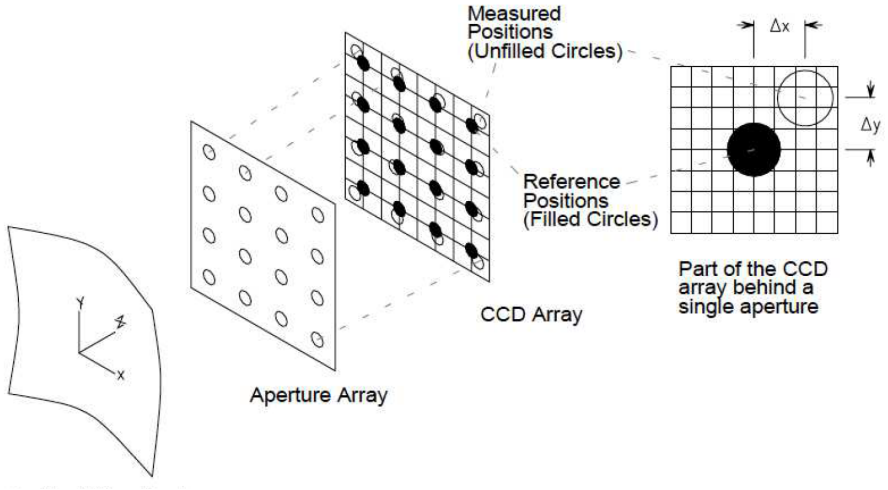


Figure 3-5: The principle of a WFS explained. (Hinnen, 2007)

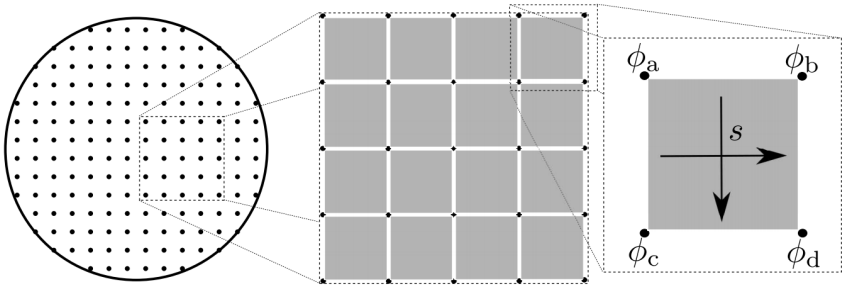


Figure 3-6: Schematic representation of the WFS sensor geometry. The black dots represent the sampling points. The grey areas are the subapertures of the sensor. (Massioni et al., 2015)

3-4-2 Deformable mirror

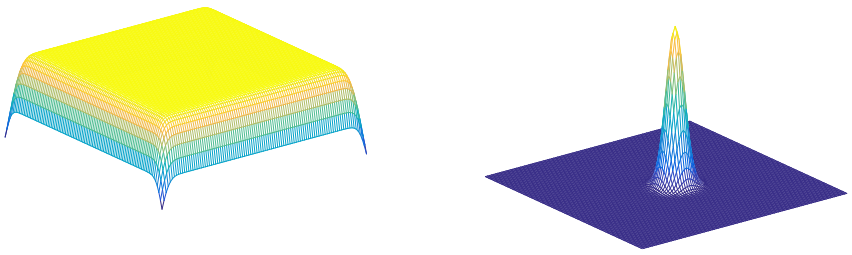
The deformable mirror (DM) is the actuator in the AO loop. It is responsible for the correction of the distorted wavefront by introducing the opposite of the optical path length differences induced by turbulence. Most DMs are a polished surface with a system of actuators behind it to control the shape. When activating each actuator individually, the obtained surface is called the *influence function* for that actuator. Hence, the set of shapes the mirror can realize can be seen as all feasible linear combinations of the influence functions. By writing the phase compensation as a matrix $H \in \mathbb{R}^{n \times m}$ times the actuator commands $u(k) \in \mathbb{R}^m$,

$$\phi_m(k+1) = Hu(k) \tag{3-29}$$

the influence functions will form the columns of the matrix H. A standard influence function is the exponential peak:

$$h(r) = e^{-\left(\frac{r}{\sigma}\right)^2} \tag{3-30}$$

where r represents the distance from the actuator location and σ is a parameter defining the width of the peak. The value $h(\delta_{act})$, where δ_{act} is the inter-actuator spacing, is called the



(a) Fully actuated mirror **(b)** Mirror with only one actuator poked.

Figure 3-7: Example of a fully actuated DM shape and a single influence function

coupling of the mirrors since it is the height of the neighbouring actuator's influence function at the location of one of the other actuators. When the coupling is too small or too large, the mirror might not be properly controllable.

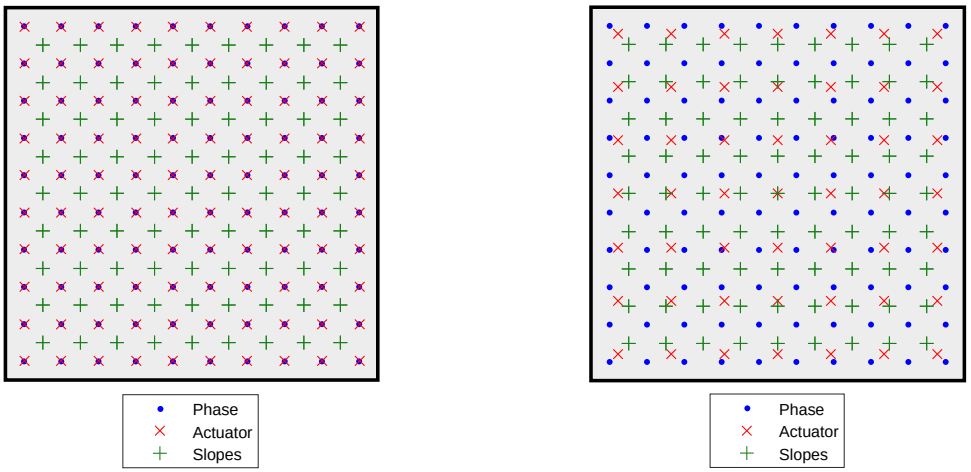
An example of a fully actuated mirror and one influence function can be seen in Figure 3-7. Figure 3-8 shows an example of the complete sensor and actuator configuration for a square mirror. The configuration in Figure 3-8a is called *Fried geometry*, since it was proposed in Fried (1977). The Fried geometry is also used in wavefront reconstruction to relate the slope measurements and reconstructed phase locations, including the configuration of Figure 3-6. In this thesis it is assumed that actuators are always evenly spread on a square grid. Except from Fried geometry, this means that there can also be less actuators than measurement locations as shown in Figure 3-8b.

This ideal static DM in Fried geometry is hardly ever satisfied for realistic systems, but still a very interesting situation in simulations. If the influence function matrix is of full row-rank, the DM will be able to compensate for any measured residual wavefront, i.e. there will be no DM fitting error that cannot be removed by the controller. Therefore, the residual phase only contains the prediction error.

3-4-3 Wavefront reconstruction

The process of estimating the residual phase distortion $\epsilon(k)$ form wavefront sensor measurements is called *wavefront reconstruction* (WFR). Classical wavefront reconstruction techniques were proposed in the works of Fried (1977), Hudgin (1977) and Southwell (1980). These methods can be categorized as *zonal* (local) wavefront reconstruction methods. Another type of wavefront reconstruction are the *modal* (global) methods, which are usually based on polynomials (such as Zernike polynomials and Karhoenen-Loève functions).

The most widely used zonal WFR methods are the so called finite difference methods. In these methods, the wavefront is defined on a rectangular grid with linear functions interpolating between the grid points. The methods differ in where the slopes are defined with respect to the reconstructed phase points. As was already discussed before, the Fried geometry (Fried, 1977) is used throughout this thesis (Figure 3-8a).



(a) Example of Fried geometry configuration. **(b)** Example with less actuators than phase points.

Figure 3-8: The DM and WFS geometry that will be used throughout this report. Left shows the special case of Fried geometry. However, the number of actuators can also be chosen smaller as shown on the right.

Usually, wavefront reconstruction assumes a static relation of the form $\hat{\epsilon}(k) = Fs(k)$ and aims at finding the matrix F that minimizes the mean squared error, or equivalently, the variance of the wavefront estimation error:

$$\min_F E\left[\|Fs(k) - \epsilon(k)\|_2^2\right]$$

With relation (3-28), a standard solution to retrieve F is to solve this problem neglecting the stochastic nature of the turbulence. The resulting optimal reconstructor clearly reduces to $F = G^T(GG^T)^{-1}$. This method however is very sensitive for measurement noise, but can give acceptable results for high signal to noise ratios.

By taking the stochastic properties into account, the reconstruction can be improved. Since the covariance of the residual wavefront C_ϵ is usually unknown, it is often approximated by the covariance of the turbulent phase C_ϕ. Using the covariance matrices of $\phi(k)$ and $v(k)$ as *a priori* information, the so called minimum-variance of maximum a posteriori (MAP) estimator (see e.g. Wallner, 1983) can be derived:

$$F = C_\phi G^T \left(GC_\phi G^T + C_v\right)^{-1} \tag{3-31}$$

One interpretation is to see $\hat{\epsilon}(k) = Fs(k)$ as the solution to the following regularized least-squares problem:

$$\min_{\epsilon(k)} \|s(k) - G\epsilon(k)\|_{C_v^{-1}}^2 + \|\epsilon(k)\|_{C_\phi^{-1}}^2 \tag{3-32}$$

where $\|x\|_W = x^T W x$ denotes the weighted 2-norm such that the first term is a weighted least-squares problem to the sensor equation and the regularization term reduces the sensitivity to

the unobservable modes. The optimal reconstruction $\hat{\epsilon}(k)$ follows indeed the static relation specified by (3-31), i.e.

$$\hat{\epsilon}(k) = C_\phi G^T \left(GC_\phi G^T + C_v\right)^{-1} s(k) \tag{3-33}$$

This static reconstruction method plays an important role in classical AO control, which will be discussed in Section 3-6.

Besides this classical way of wavefront reconstruction there are other more recent algorithms such as Cure-D (Rosensteiner, 2012) and D-SABRE (de Visser et al., 2016) that are more efficient. Optimizing the wavefront reconstruction is one way of reducing the complexity of AO control. However, it is not the path that is taken in this thesis and these methods are therefore out of the scope of this research.

3-4-4 Unobservable modes

Only the part of the wavefront that is in the row-space of G can be reconstructed from measurements. The unobservable modes (which according to the definition in Section 3-4-1 are the piston and waffle modes) lie in the null-space of G. This means that the observable and unobservable part can be split using the SVD of G:

$$G = U\Sigma V = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix}$$

$$GV = U \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} C & 0 \end{bmatrix}$$

where U and V are orthonormal matrices and Σ is the diagonal matrix containing all singular values. The partitioning is chosen such that Σ_1 contains only non-zero singular values. By substituting this representation into (3-28) and by defining

$$\bar{\phi}(k) = V_1^T \phi(k)$$
$$\phi_{pw}(k) = V_2^T \phi(k)$$

it can be concluded that $V_1\bar{\phi}(k)$ is the part of the turbulence containing only the observable modes and $V_2\phi_{pw}(k)$ is a sum of the unobservable piston and waffle mode present in $\phi(k)$. The WFS measurement equation (3-28) can now be written as a function of $\bar{\phi}(k)$ only

$$\begin{aligned} s(k) &= G\phi(k) + v(k) = GVV^T\phi(k) = \begin{bmatrix} C & 0 \end{bmatrix} \begin{bmatrix} \bar{\phi}(k) \\ \phi_{pw}(k) \end{bmatrix} + v(k) \\ &= C\bar{\phi}(k) + v(k) \end{aligned}$$

In this representation, $s(k)$ describes only the informative part of the measurements and it filters out the noise that cannot be caused by the wavefront. Furthermore, because of the orthonormality of V_1, $\phi(k)$ and $\bar{\phi}(k)$ will have the same 2-norm. To obtain one reduced formulation of the complete AO system, these definitions have to be extended to the DM. The reduced representation of the DM phase is defined as

$$\bar{\phi}_{dm}(k) = V_1^T H u(k-1)$$

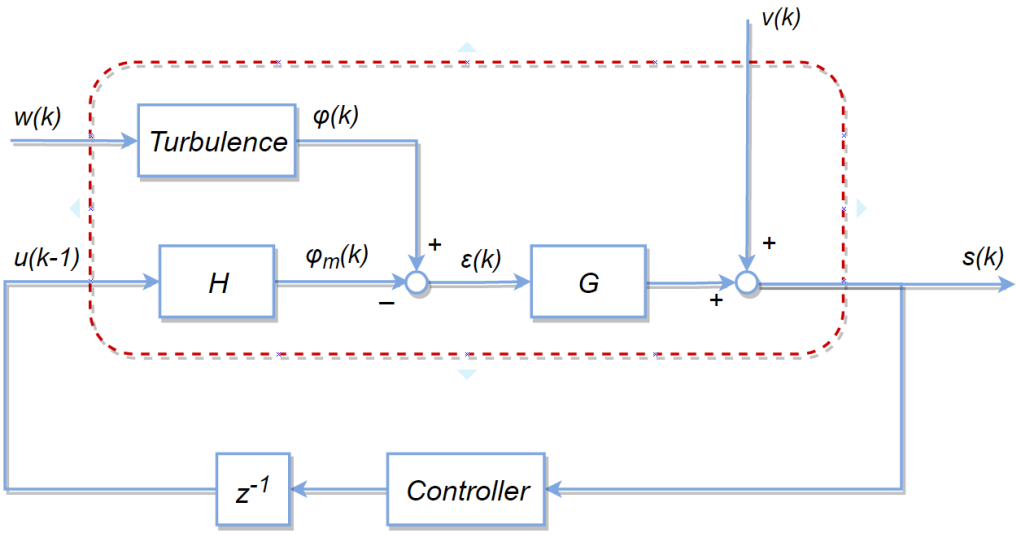


Figure 3-9: Schematic representation of an AO system as a closed control loop. The system within the red dashed box contains the open loop AO model to be identified.

The main advantages of the reduced model over the full model is the dimensionality reduction and the better numerical conditioning. On the other hand, the SVD will erase any structures or sparsity that are naturally present in the system. Therefore, if the structure and sparsity is crucial in obtaining an efficient method, this reduced formulation might not be a good option. One interpretation of the signal $\bar{\phi}(k)$ is to see it as a coefficient belonging to a column of V_1 which is nothing more than one of the observable global modes. Intuitively, the structures are lost because after the reduction the zonal formulation is changed into a modal one.

Dealing with the unobservable modes, while remaining in a zonal representation of the phase is an interesting subject. However, the unobservable modes do not necessarily have to be removed from the model to achieve satisfiable results. The piston mode does not affect the image quality, hence it can just be removed afterwards without consequences. The most direct method is to simply remove the mean of the obtained wavefront before processing it. The waffle mode however does influence the image quality. However, as it is the mode with the highest frequency that can be measured, it is expected to have only little influence since the Kolmogorov spectrum has the most power at the lower frequencies (see Figure 3-1). One possible danger is the appearance of the unobservable modes in the input signal. Since they are not measured, they will often also not influence the control objective and might end up in the phase of the DM. A random piston mode in the input might cause the saturation of the actuators and a random waffle mode directly decreases the sharpness of the image. Therefore, if the modes are not removed from the model, they should be removed from the input.

3-4-5 The AO system in closed loop

After the wavefront reconstruction step, the loop is closed by relating the reconstructed residual wavefront $\hat{\epsilon}(k)$ to the DM input $u(k)$. A scheme of this loop is displayed in Figure 3-9. The red dashed box contains all elements to be modelled and identified.

In classical AO control, the controller exists of the wavefront reconstructor, a projection of this residual onto the actuator space and an integrator. The wavefront reconstructor has been discussed in Section 3-4-3 and is often represented by the relation (3-33). The projection for the simple DM relation of (3-29) is usually also described by a static linear operation: $u(k) = M\hat{\epsilon}(k)$. This projection of the phase onto the actuators can be found by minimizing the mean squared fitting error

$$\min_M \|\hat{\epsilon}(k) - HM\hat{\epsilon}(k)\|_2^2$$

In contrast to the WFR problem, the mapping onto the mirror is a standard deterministic least-squares problem and for a tall matrix H, the projection on the mirror is given by

$$u(k) = M\hat{\epsilon}(k) = (H^T H)^{-1} H^T \hat{\epsilon}(k) \tag{3-34}$$

This combined with the wavefront reconstructor (3-33) actually is the classical static control approach that will be discussed in more detail in Section 3-6. The next section will proceed on modelling the open loop AO system.

3-5 Modelling and identification of AO systems

One of the primary goals of this thesis is the identification of the open-loop AO system. As was visualized in the control loop of Figure 3-9, the physical system has both a deterministic input $u(k)$ and a process noise $w(k)$ due to the stochastic nature of the turbulence. There are various possible model structures able to represent systems with both a deterministic and stochastic part. One general separation can be made between identifying only a transfer function from $u(k)$ and $w(k)$ to $s(k)$ or identifying a state-space model with deterministic input $u(k)$ and process noise $w(k)$.

In this thesis, the focus will lie mainly on identifying ARX (Auto-Regressive with eXogenous input) models, or strictly speaking VARX models with "V" standing for vector. The identification of a state-space innovation model has been investigated, but was discovered to have difficulties scaling to large dimensions. VARX models have the advantage that they are simple, while containing all sufficient information. This section will provide a quick motivation for choosing the VARX structure to represent the AO system in Chapter 4. Recall from Chapter 2 that VARX models are models of the form

$$\sum_{i=0}^{n_a} A_i y(k-i) = \sum_{i=1}^{n_b} B_i u(k-i) + r(k) \tag{3-35}$$

where the Gaussian noise $r(k) \sim \mathcal{N}(0, C_r)$ is independent from u. A more general description of the AO system is given by a state-space model of the form

$$\begin{aligned} x(k+1) &= Ax(k) + Bu(k) + w(k) \\ y(k) &= Cx(k) + v(k) \end{aligned} \tag{3-36}$$

where $w(k) \sim \mathcal{N}(0, C_w)$ and $v(k) \sim \mathcal{N}(0, C_v)$ are uncorrelated with the input $u(k)$. The input-output transfer function of this system also follows by formulating (3-36) in innovation form (Section 2-1).

$$\begin{aligned} \hat{x}(k+1) &= \bar{A}\hat{x}(k) + Bu(k) + Ky(k) \\ y(k) &= C\hat{x}(k) + e(k) \end{aligned} \tag{3-37}$$

where $\bar{A} = A - KC$ and $e(k) = y(k) - C\hat{x}(k)$ is the innovation sequence. The above model can be seen as the Kalman predictor model of (3-36). Since $e(k)$ is a zero mean white noise sequence, a predictor of (3-37) now can be written as

$$\hat{y}(k+1 \mid k) = C\bar{A}\hat{x}(k \mid k-1) + CBu(k) + CKy(k) \tag{3-38}$$

which by recursively substituting in the state becomes

$$\begin{aligned}\hat{y}(k+1 \mid k) = {} & CBu(k) + C\bar{A}Bu(k-1) + \ldots + C\bar{A}^k Bu(0) + \ldots \\ & CKy(k) + C\bar{A}Ky(k-1) + \ldots + C\bar{A}^k Ky(0)\end{aligned} \tag{3-39}$$

The obtained predictor can be interpreted as the predictor of a high-order VARX model of the form (3-35). This shows how all stable models of the form (3-37) can be represented by a higher order VARX model. Especially when there exists a relatively small integer q such that $\bar{A}^i \approx 0,\ \forall i > q$, VARX models might be a very good representation.

Next, it is shown how the AO system fits in the general representations discussed above.

3-5-1 Modelling of the adaptive optics system

Under a number of assumptions, the AO system will be represented into a state-space form. First, it is assumed that all layers are assumed to have the same spatial statistical properties such that it can be reduced to one layer without loss of generality. Secondly, the Taylor frozen flow is supposed to hold. Finally, the DM and WFS dynamics are neglected. Under these conditions, the turbulence model can be accurately described by a first order VAR model as in (3-19) and the DM is specified by (3-29), i.e.

$$\begin{cases} \phi(k+1) & = A\phi(k) + w(k) \\ \phi_m(k+1) & = Hu(k) \end{cases}$$

with $w(k) \sim \mathcal{N}(0, C_w)$. However, the derivation introduced in this section can straightforwardly be extended to more complex higher order VAR turbulence models (see also Massioni et al., 2015). The wavefront residual, $\epsilon(k) = \phi(k) - \phi_m(k)$, assuming the first order VAR turbulence model simply follows from the difference of the relations above, that is:

$$\begin{aligned}\epsilon(k+1) &= A\phi(k) - Hu(k) + w(k) \\ &= A(\epsilon(k) + \phi_m(k)) - Hu(k) + w(k) \\ &= A\epsilon(k) - Hu(k) + AHu(k-1) + w(k)\end{aligned}$$

A state-space model is constructed with state $\epsilon(k)$ and output $s(k)$ using the above equation and output equation (3-28):

$$\begin{aligned}\epsilon(k+1) &= A\epsilon(k) - Hu(k) + AHu(k-1) + w(k) \\ s(k) &= G\epsilon(k) + v(k)\end{aligned} \tag{3-40}$$

The innovation form is obtained by the derivation of the Kalman filter:

$$\hat{\epsilon}(k+1 \mid k) = (A - KG)\hat{\epsilon}(k \mid k-1) - Hu(k) + AHu(k-1) + Ks(k) \tag{3-41}$$

where the Kalman gain K is the outcome of a Riccati equation. However, solving the Riccati equation might cause numerical problems when the measurement or process covariance matrices get very small. For a more robust computation of the Kalman gain, it is computed using the so-called square root covariance filter (Verhaegen and Verdult, 2007). The matrix A can be derived by solving the Yule-Walker equations (2-9) when statistical knowledge of the turbulence $\phi(k)$ is available. Alternatively, it could be derived from measurements $s(k)$, but this would mean that the model does not include the same unobservable modes as the sensor.

3-5-2 Identifying the model

The next step is the identification of either the innovation model given by (3-41), or an approximation in the form of a VARX model. The standard case has been briefly analysed in Chapter 2 and the reader is referred to Section 2-1 for the basic theory on system identification. Moreover, with an eye on the scalability of the control algorithm, it is crucial to enforce sparsity within the system matrices. If it is assumed for now that the sparsity pattern is not precisely known, such that adding ℓ_1-regularization to the identification problem seems the most suitable tool for this purpose. Both a prediction error and subspace identification method have been investigated in the process of this research.

The prediction error method is based on a VARX approximation of (3-41) and follows the same approach as the VAR identification via least-squares from Section 2-1, but in a graphical modelling framework (Section 2-2). It should be noted that (3-41) cannot be exactly represented by a VARX model in general. Only for the special case that $A - KC$ is nilpotent a VARX model is obtained after substituting the state equation into the output equation. By approximating the model as a VARX model, a trade-off between accuracy and complexity has been made. Chapter 4 will present the identification routine step-by-step and the performance of this method will be validated and compared to an accurate innovation model of the form (3-41).

As an alternative to the prediction error method, a sparse state-space model has been derived using subspace identification. A nuclear norm approach, similar to the N2SID method of Verhaegen and Hansson (2014) or the SVD-free system identification approach of Signoretto et al. (2013), was shortly investigated in its capability to identify a sparse model during the process of this thesis. Although the method could be promising for small systems with only a few inputs and outputs, it became clear very quickly that is had no potential of scaling up to even small-scale adaptive optics applications. Since the main focus of this thesis lies on a solution to the large-scale problem it was chosen not to pursue the method any further. Nevertheless, for the sake of the completeness of this thesis, a concise summary of the method is included in Appendix B.

3-6 Control methods for AO

Over the past decades, various methods have been proposed to control the DM. Each of them establish a different trade-off between the effectiveness of minimizing the phase variance and computational complexity. In this section, two main categories are discussed. The first one

is the classical way of AO control, which assumes stochastic knowledge of the turbulence as *a priori* information and a random-walk prediction model. The second more modern type of control finds an optimal minimum-variance predictor for the turbulence phase and uses for example the framework of LQG regulators (Kulcsár et al., 2006) or $\mathcal{H}_2$ optimal control (Hinnen et al., 2008). The control problem for which a novel method will be proposed in Chapter 4 will be outlined below, followed by its solution using the classical and optimal control framework.

3-6-1 Control problem formulation

The goal of adaptive optics is to cancel out wavefront aberrations in optical systems by actively adding a change in optical path length of opposite phase. In other words, the residual wavefront has to be estimated and its inverse should be mapped onto the deformable mirror. Because it is assumed in this thesis that there are no WFS or DM dynamics and that there is only one single time delay, the computation of $u(k)$ requires a one-step-ahead prediction of the residual wavefront $\hat{\epsilon}(k+1)$. Furthermore, it is assumed that at time instance k, all prior control actions and WFS measurements (at $k-1, k-2, \ldots$) are known and that we can read the WFS to get the current sensor measurement $s(k)$ before computing the DM voltages.

Under these assumptions, the control problem reduces to an optimal estimation problem finding $\hat{\epsilon}(k+1)$ and the computation of the input $u(k)$ that flattens this predicted residual wavefront. As a measure of the flatness of the wavefront, usually the mean squared error is considered, such that the control law can be written as the following simple minimization problem

$$\min_{u(k)} \quad \|\hat{\epsilon}(k+1)\|_2^2 \tag{3-42}$$

where $\hat{\epsilon}(k+1)$ is a function of the input $u(k)$. Furthermore, to decrease the control effort, a regularization term on the input might be added.

$$\min_{u(k)} \quad \|\hat{\epsilon}(k+1)\|_2^2 + \lambda \|u(k)\|_2^2 \tag{3-43}$$

Increasing λ will decrease the control effort and the sensitivity to model uncertainties.

To have a real-time implementation of this controller, both the sensor reading and the mapping onto the mirror have to be within one sampling period. With a sparse model, problem (3-42) can be computed faster and therefore the computations can be completed within a shorter period of time or in the same period for a larger system, again emphasizing the advantage of retrieving a sparse model.

Next, two different solutions to the AO control problem are introduced. First, the classical static approach that maps the reconstructed wavefront on the mirror followed by a sequence of separate single input single output temporal filters. The second method constructs the optimal estimator (3-41) and minimizes the corresponding one-step-ahead predictor at each time instance.

3-6-2 Classical control for AO systems

In classical AO control methods, there is no turbulence model available for predicting $\epsilon(k+1)$. Without a model, the temporal dynamics of the wavefront disturbance is neglected and the next wavefront is estimated by the current reconstructed wavefront $\hat{\epsilon}(k)$. Since this concerns the residual wavefront, it implies that by using the measurements $s(k)$, the obtained control action will not be the control action $u(k)$ but its increment $\Delta u(k) = u(k) - u(k-1)$. In other words, the currently applied DM phase compensation is measured in the signal $s(k)$ along with the turbulence and the computed control signal has to be applied to the current actuator commands.

The classical control approach furthermore assumes a static relation between actuator input and the measurements given by $\Delta u(k) = Rs(k)$ such that the problem of finding $\Delta u(k)$ is equivalent to finding the matrix R. Because of this simple relation, the method is often referred to as *MVM*, standing for Matrix-Vector Multiplication. According to the above definitions, minimizing the mean-square error of $\hat{\epsilon}(k+1)$ boils down to finding the mapping R that optimizes

$$\min_R E\left[\|\hat{\epsilon}(k) - HRs(k)\|_2^2\right] \tag{3-44}$$

where the expectation $E[\cdot]$ is introduced because of the stochastic nature of $\hat{\epsilon}(k)$. Under the additional assumption that the difference between C_ϵ and C_ϕ is negligible, the maximum a posteriori estimate of R is given by

$$R = (H^T H)^{-1} H^T C_\phi G^T \left(GC_\phi G^T + C_v\right)^{-1} \tag{3-45}$$

The matrix R is a combination of wavefront reconstruction (3-31) and a projection onto the DM (3-34) as $R = MF$, i.e. it projects the reconstructed wavefront onto the mirror.

The control update $u(k)$ will be the sum of the previous value plus the optimal increment, i.e.

$$u(k) = u(k-1) + Rs(k) \tag{3-46}$$

This implies that the temporal filters should posses integrating action. Usually, two extra tuning parameters are introduced in the parallel feedback loops

$$u(k) = R\frac{c_1}{1 - c_2 z^{-1}} s(k) \tag{3-47}$$

where c_1 is the integrator gain and c_2 the loss factor. In this thesis, both parameters are however fixed to $c_1 = c_2 = 1$, such that the control action follows from (3-46).

3-6-3 Optimal control for AO

The random-walk prediction model in MVM control neglects the turbulence dynamics. Instead of this simplification, one could try to find a one-step-ahead predictor $\hat{\epsilon}(k+1 \mid k)$ to get an accurate estimation of what the wavefront residual is going to be in the next time step. Examples of this more modern control point of view can be found in Kulcsár et al. (2006),

Looze (2006) and Hinnen et al. (2008) amongst others.

In order to have a proper performance comparison for the new predictor introduced in Chapter 4, a Kalman filter based equivalent of that controller is used besides the classical MVM method. The method solves the control problem (3-42) with the predictor $\hat{\epsilon}(k+1 \mid k)$ obtained from the Kalman filter (3-41). Substituting the estimator into the control problem gives the following optimization problem

$$\min_{u(k)} \|(A - KG)\hat{\epsilon}(k|k-1) - Hu(k) + AHu(k-1) + Ks(k)\|_2^2 \tag{3-48}$$

where $\hat{\epsilon}(k|k-1)$ is the known previous one-step-ahead predictor. This standard least-squares problem is solved by

$$u(k) = (H^T H)^{-1} H^T \left((A - KG)\hat{\epsilon}(k \mid k-1) + AHu(k-1) + Ks(k)\right) \tag{3-49}$$

Because of its predictive capability, this method will improve the results compared to the MVM method, provided that the identified system matrices (A, B, C, K) are accurate. The main drawback of this approach is its lack of scalability, since for large systems it would require solving a large Riccati equation to retrieve K. The next paragraph will proceed on the obstacles of deriving a control method for large-scale AO systems.

3-6-4 Control for large-scale adaptive optics

As indicated before, the derivation of the Kalman filter (3-41) requires the solution to a Riccati equation. When the number of actuators and sensors are in the order of n, the number of operations required to solve the Riccati equation scales cubically with n. This exponential increase in complexity with the systems dimension will eventually cause problems when the number of actuators and sensors is increased. To achieve a scalable method, this complexity should be brought down significantly.

There are two main directions of improving the efficiency of adaptive optics control for extremely large-scale AO applications such as the E-ELT (Gilmozzi and Spyromilio, 2007) in the literature. First, the static MVM method can be made more efficient. This is not the path taken in this thesis since it still neglects the turbulence dynamics. Secondly, the complexity of the optimal control method can be reduced. However, the literature mainly focusses on a more efficient solution to the Riccati equation (see e.g. Massioni et al., 2015). An additional drawback of many techniques are that they rely on physical laws and cannot compensate for discrepancies between the theoretical model and real system. Especially for modelling the disturbance dynamics, data-driven methods that fit the model to the dynamics of the real system are preferred above first principle models.

The next chapter will introduce a novel data-driven optimal control method that consists of a scalable identification routine and a minimum-variance control law in terms of a sparse least-squares problem. The trade-off between performance and efficiency is central to the design of this method. Therefore, the open-loop AO system is approximated by a sparse VARX model. The approximation should only have a slightly lower performance than the optimal method discussed in Section 3-6-3, so still significantly outperforming MVM. Finally, the computational complexity of the identification should preferably scale close to linear to the number of sensors and the sparsity should be exploitable in the computation of the actuator commands in order for the method to be actually scalable.

Chapter 4

Sparse VARX model identification for large-scale AO

This chapter introduces a novel data-driven optimal control method that consists of a scalable identification routine and a minimum-variance control law in terms of a sparse least-squares problem. To decrease the complexity of the identification step, the open-loop AO system is approximated by a sparse VARX (Vector Auto-Regressive with eXogenous input) model. The advantage of this approach over the innovation model is that the VARX model framework can easily exploit the sparse nature of the physical system.

First, the validity of the VARX approximation is discussed in Section 4-1. The new data-driven identification of a sparse VARX model for AO is presented in Section 4-2. This section will include a routine to decrease the number of parameters in the identification problem by following a graphical modelling framework. The final sparsity in the coefficient matrices is determined by adding an ℓ_1-regularization term to the original prediction error estimation problem. The performance of the identified model is validated in Section 4-3, where it is compared to an innovation model. Finally, the new control law is discussed in Section 4-4, followed by numerical simulations comparing the new method with MVM (Section 3-6-2) and the controller based on the Kalman filter (Section 3-6-3).

4-1 AO system as a VARX model

Before the identification algorithm is presented, a more profound analysis of representing AO systems as VARX models will be considered. Also recall that any stable model of the form (3-37) can be described by a (higher order) VARX model. This implies that the expression

$$s(k) = A_1 s(k-1) + \ldots + A_{n_a} s(k-n_a) + B_1 u(k-1) + \ldots + B_{n_b} u(k-n_b) + e(k) \tag{4-1}$$

with $e(k) \sim \mathcal{N}(0, C_e)$ should in theory be able to represent the model (3-37) for high enough orders n_a and n_b. By introducing the matrix $A_0 = C_e^{-1/2}$, (4-1) can be rewritten in an

equivalent formulation that is more convenient for some applications, that is

$$A_0 s(k) = \bar{A}_1 s(k-1) + \ldots + \bar{A}_{n_a} s(k-n_a) + \bar{B}_1 u(k-1) + \ldots + \bar{B}_{n_b} u(k-n_b) + n(k) \tag{4-2}$$

with $n(k) \sim (0, I)$, $\bar{A}_i = A_0 A_i$ and $\bar{B}_i = A_0 B_i$. Depending on the complexity of the turbulence and system dynamics, the best model order to describe the VARX model might change.

However, when the turbulence is simulated using Taylor's frozen flow assumption and the DM and WFS dynamics are neglected, Section 3-5 has determined the fact that the AO system can be represented by

$$\begin{aligned} \epsilon(k+1) &= A\epsilon(k) - Hu(k) + AHu(k-1) + w(k) \\ s(k) &= G\epsilon(k) + v(k) \end{aligned} \tag{4-3}$$

In practice, it appears that the VARX model already accurately approximates (4-3) for $n_a = 1$ and $n_b = 2$, i.e.

$$s(k+1) = A_1 s(k) + B_1 u(k) + B_2 u(k-1) + e(k) \tag{4-4}$$

Or equivalently, by introducing the matrix $A_0 = C_e^{-1/2}$:

$$A_0 s(k+1) = \bar{A}_1 s(k) + \bar{B}_1 u(k) + \bar{B}_2 u(k-1) + n(k) \tag{4-5}$$

An interesting interpretation that explains this result follows when it is supposed that the linear relation $\epsilon(k) = Fs(k)$ of (3-33) holds. An expression for $s(k)$ can straightforwardly be derived by substituting the state update equation into the output equation of (4-3) and together with $\epsilon(k) = Fs(k)$, this can be converted into

$$s(k+1) \approx GAFs(k) - GHu(k) + GAHu(k-1) + r(k)$$

with a certain white noise sequence $r(k) \sim \mathcal{N}(0, C_r)$. Besides this simple reasoning, the low-order VARX approximation and the influence of increasing the order on the model's accuracy will be validated with numerical simulations in Section 4-3.

It should be stressed that the general VARX model of (4-1) should be able to represent also more complicated turbulence and system dynamics. However, this would require a more general turbulence simulation method (e.g. the method of Beghi et al., 2008). Since these simulators are problematic in terms of computational complexity, they are not used in this thesis.

4-2 Sparse VARX model identification

In order to be able to have an accurate predictor of the residual wavefront, it is a necessity to identify a model in the form of (4-1) or (4-2). The approach taken in this thesis follows the prediction error methods presented in Section 2-1. In this framework, one possible way of identifying (4-1) as sparse as possible would be to minimize the following regularized least-squares problem

$$\min_X \|Y - XZ\|_2^2 + \lambda \|X\|_1 \tag{4-6}$$

where

$$X = \begin{bmatrix} A_1 & \cdots & A_{n_a} & B_1 & \cdots & B_{n_b} \end{bmatrix}$$

$$Y = \begin{bmatrix} s(q+1) & s(q+2) & \cdots & s(N) \end{bmatrix} \qquad Z = \begin{bmatrix} s(q) & s(q+1) & \cdots & s(N-1) \\ s(q-1) & s(q) & \cdots & s(N-2) \\ \vdots & \vdots & \ddots & \vdots \\ s(q-n_a+1) & s(q-n_a+2) & \cdots & s(N-n_a-1) \\ u(q) & u(q+1) & \cdots & u(N-1) \\ u(q-1) & u(q) & \cdots & u(N-2) \\ \vdots & \vdots & \ddots & \vdots \\ u(q-n_b+1) & u(q-n_b+2) & \cdots & u(N-n_b-1) \end{bmatrix}$$

with $q = \max(n_a, n_b)$. This problem can easily be decentralized by solving the problem for each row of X separately, i.e. we get p distinct optimization problems. By denoting the i-th row of X as $X_{(i,\star)}$ and using similar definitions for the other matrices, the identification of the i-th row of matrix X is reduced to

$$\min_{X_{(i,\star)}} \|Y_{(i,\star)} - X_{(i,\star)} Z\|_F^2 + \lambda \|X_{(i,\star)}\|_1 \tag{4-7}$$

Although it is decentralized, the scalability of (4-7) is prohibited by the large dense matrix Z containing all identification data. One way of avoiding this difficulty is by considering a sparsity pattern *a priori*. If only a small number of elements in each row are known to be non-zero, the number of parameters in the prediction error estimation problem will drop significantly. The following paragraph will discuss a framework to determine such a pattern using graphical modelling theory to exploit the spatio-temporal correlations.

4-2-1 Enforcing a sparsity pattern via graph theory

By expressing the WFS signal $s(k)$ as (4-1), it implies that each element of $s(k)$ depends on past and current values of all other measurement points. However, based on the frozen flow assumption, we know that for a high enough sampling frequency, the turbulence is just a slightly shifted version of the turbulence in the previous time instance. Therefore, the turbulence that is now above a certain location on our grid, was one time instance back (approximately) above a point on the grid relatively close to this location. In other words, we do not need to consider the influence of all the other measurements, but only of a small number of neighbouring point. Graphical modelling, which was introduced in Chapter 2, formalizes this kind of reasoning such that it can be exploited to increase the efficiency of statistical calculations.

There are two main relations to be described in the graphical modelling framework. The first group contains the relations between the slopes $s(k)$ and its own past values $s(k-i)$. The underlying physical cause of these relations is mainly due to the turbulence dynamics. The second set describes the influence of the past DM voltages $u(k-i)$ on the current wavefront $s(k)$, caused by the coupling of the actuator influence functions. The WFS signal $s(k)$ consists of the slopes in the two directions of the measurement grid. The two slopes in each locations are influenced by the same elements and thus are represented by one "measurement location" in the square configuration grid of Figure 4-1. Moreover, a property of the new method should

be that it is completely data-driven. Therefore, it is assumed that there is no knowledge on the wind speed and wind direction available. This suggests that we have to look for a certain circular "neighbourhood of influence" for both the turbulence movement and DM coupling.

An example of such a neighbourhood is shown in Figure 4-1. The black asterisk represents the "measurement location" to be predicted, from now on denoted by $s_i(k+1)$, and only the actuator or sensor locations within this circle are considered to contribute to its estimation. For each separate relation between $s_i(k+1)$ and $s(k)$, $u(k)$, $s(k-1)$, $u(k-1)$, etc., such a neighbourhood has to be defined. To this intent, a graphical model framework is imposed.

Referring to the theory presented in Section 2-2, the relations between the elements of $s(k)$ and its past own or control action values are described by causality relations. Because there is no knowledge of the wind direction, each causality relation is assumed to be in both directions. It would be possible to define certain relations as one-directional arcs if the wind direction would be specified. Either the Granger causality graph framework or the time series chain graph seems to be most suited for the application, where the latter can achieve a more specific sparsity because of its more general structure. Furthermore, graphical VARX modelling in these frameworks has the property that a missing arc means a zero in one of the coefficient matrices. In the following, each measurement location and actuator location will be referred to as *nodes* or *vertices*, with node s_i the measurement location to be predicted, and a causality relation between two nodes is described by a directed *arc*. The set of vertices from which an arc starts towards node s_i is denoted as V_i which hence can be interpreted as the columns of $X_{(i,\star)}$ which are possibly non-zero. The row $X_{(i,\star)}$ with all zero values removed, will be denoted as $X_{(i,V_i)}$. Using similar notations for the other matrices, the reduced identification problem can be written as

$$\min_{X_{(i,\star)}} \|Y_{(i,V_i)} - X_{(i,V_i)} Z_{(V_i,\star)}\|_F^2 + \lambda \|X_{(i,\star)}\|_1 \tag{4-8}$$

For the sake of the comprehensiveness of this section, we will consider only the low-order approximation of the VARX model $n_a = 1$ and $n_b = 2$ (4-4). By defining $z(k-1) = [s^T(k-1)\ u^T(k-1)\ u^T(k-2)]^T \in \mathbb{R}^{p+2m}$, it can be concluded that

$$j \notin V_i \quad \text{if} \quad s_i(k) \perp z_j(k-1) \mid z_{\setminus j}(k-1)$$

where $z_j(k)$ denotes the j-th element in $z(k)$ and $z_{\setminus j}(k)$ all elements except for the j-th element. The conditional independence relations are directly linked to the distance between the location of each sensor and actuator on the grid, i.e. we can define three radii (r_{A_1}, r_{B_1} and r_{B_2}) such that

$$\begin{aligned} s_i(k) \perp z_j(k-1) \mid z_{\setminus j}(k-1), \quad & \text{for} \ \ 1 \le j \le p && \text{if} \ \ d(s_i, z_j) > r_{A_1} \\ & \text{for} \ \ p+1 \le j \le p+m && \text{if} \ \ d(s_i, z_j) > r_{B_1} \\ & \text{for} \ \ p+m+1 \le j \le p+2m && \text{if} \ \ d(s_i, z_j) > r_{B_2} \end{aligned} \tag{4-9}$$

with $d(\cdot,\cdot)$ denoting the spatial separation in meters between the two specified nodes on the grid and p and m are the dimensions of $s(k)$ and $u(k)$ respectively. Furthermore, z_j denotes the location of the actuator or sensor on the grid corresponding to the j-th element in $z(k)$. The following paragraphs will focus on finding the radii. It should be noted that this discussion can straightforwardly be extended to the case of a higher-order VARX model.

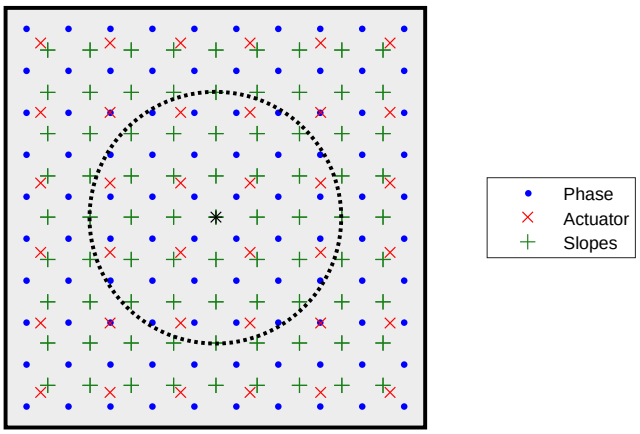


Figure 4-1: An example of a configuration of the reconstructed phase, slope measurements and actuator locations. Only the elements within the circle will influence the prediction of the location at the center (black *)

Topology of the AO graphical model

This paragraph will present a routine to roughly estimate the topology and hence the sparsity pattern of the matrices A_1, B_1 and B_2. It is stressed again that the same reasoning can also be applied to higher order VARX approximations.

Note that in general we cannot simply shift the measured $s(k-1)$ to predict the influence of the turbulence on $s(k)$ since it also contains the DM shapes at $k-1$. Therefore, to predict $s(k)$ we also need the value of $u(k-2)$ to exclude its influence from $s(k-1)$. This suggests that B_2 includes the DM information within $s(k-1)$ which makes it both dependent on the mirror influence function and the turbulence graph. The matrices A_1 and B_1 are not coupled and completely independent from each other, describing the turbulence dynamics and DM influence functions respectively.

Since it is assumed that $n_a = 1$, only the relation between $s_i(k)$ and $s(k-1)$ is considered regarding the turbulence dynamics. That is, $s_i(k) = A_{1,(i,\star)}s(k-1)$ after removing all effects of the DM from $s(\cdot)$. Due to the frozen flow dynamics, the turbulence will move vT_s meter per sampling period, where v is the wind speed in m/s and T_s the sampling time in seconds. Because the direction is assumed to be unknown and only an (over)estimation of the speed v is available, the neighbourhood that might influence element i is a circle with a radius of at least $r_{A_1} = vT_s$ meter. Therefore

$$s_i(k) \perp s_j(k-1) \mid z_{\setminus j}(k-1) \quad \text{if} \quad d(s_i, s_j) > vT_s$$

and since each conditional independence relation results in one zero in the coefficient matrix, each of the following i,j-th elements of A_1, denoted by $A_{1,(i,j)}$, will become zero

$$A_{1,(i,j)} = 0 \quad \text{if} \quad d(s_i, s_j) > vT_s$$

The topology of the graph belonging to the deformable mirror, i.e. the relation from u to s, is very clearly described by a circle with a radius equal to the width of the influence functions. So assuming the influence functions of (3-30), one can set a certain threshold ε under which the height can be neglected, e.g. when it gets below 0.1% of the maximum height. The radius of the circle on the spatial grid where the exponential function is equal to this height, given by $r_{B_1} = \sqrt{-\sigma^2 \ln \varepsilon}$, forms directly our radius to describe the graph. Therefore, the following conditional independence relations can be specified

$$s_i(k) \perp u_j(k-1) \mid z_{\setminus j+p}(k-1) \quad \text{if} \quad d(s_i, u_j) > \sqrt{-\sigma^2 \ln \varepsilon}$$

since $z_{j+p}(k-1)$ corresponds to $u_j(k-1)$, such that the elements of B_1 that will become zero are

$$B_{1,(i,j)} = 0 \quad \text{if} \quad d(s_i, u_j) > \sqrt{-\sigma^2 \ln \varepsilon}$$

Finally, the coupling between the DM input $u(k-2)$ and the measurement $s(k-1)$ is included in the matrix B_2. Its neighbourhood will consist of all actuator locations that influence the region defined by r_{A_1} somehow, which implies that $r_{B_2} = r_{A_1} + r_{B_1}$. Writing it again in terms of conditional independence relations gives

$$s_i(k) \perp u_j(k-2) \mid z_{\setminus j+p+m}(k-1) \quad \text{if} \quad d(s_i, u_j) > \sqrt{-\sigma^2 \ln \varepsilon} + vT_s$$

and the corresponding sparsity in the coefficient matrix B_2 follows directly from

$$B_{2,(i,j)} = 0 \quad \text{if} \quad d(s_i, u_j) > \sqrt{-\sigma^2 \ln \varepsilon} + vT_s$$

with these radii, the set V_i can be found from all elements j in (4-9) that are not conditional independent of $s_i(k)$.

Besides the causality relations, the time series chain graphs also contain undirected edges that represent the conditional dependence of the nodes at the same time instance. In terms of the AO system, these correlations appear in the process noise $w(k)$. Recall that in Chapter (2-2) it was shown how each missing edge corresponded to a zero value in the inverse of the covariance of the VARX model's stochastic signal ($e(k)$ in (4-1)). With an eye on the minimum variance control problem, where the inverse covariance matrix C_e^{-1} is used as a weighing matrix, a sparse estimation of this matrix is crucial. Following the same reasoning as above, the dimensionality of this estimation problem can be significantly reduced by removing all elements definitely corresponding to conditional independence relations.

When other turbulence models or multiple layers are considered, it is possible that a higher order VARX model is necessary. When the order gets higher, the radius will get larger and larger, as the uncertainties and overestimations in the radii will accumulate and more coupling terms will appear. From this we can also conclude that when more terms are needed in the VARX model to describe the AO systems' dynamics, the new matrices get more dense if we try to describe the sparsity using a circle of influence.

Finally, it should be underlined that the method described in this paragraph does not find the most sparse solution to the problem, but rather serves as a rough tool to remove the obvious zero values from the identification problem. Mainly because there are many parameters that are highly dependent on the exact circumstances and that a circle of influence is not always

the best way to describe the graphical structure, it is impossible to find the most optimal sparsity by hand. When retrieving the sparsest solution is paramount to the time it takes to solve the identification problem itself, an ℓ_1-regularization term should be added as done in (4-7). The corresponding problem can be solved for example with the Alternating Direction Method of Multipliers (ADMM) and will be discussed below.

4-2-2 Sparse VARX identification using ADMM

The next step of the new method is to solve the new reduced identification problem (4-8). Since it is an ℓ_1-norm regularized problem, it can be solved by the Alternating Direction Method of Multipliers (Boyd et al., 2011). The required background knowledge on ADMM is summarized in Appendix A. For the ease of notation, all subscripts will be dropped, and (4-8) is reformulated as

$$\min_{x} \ \frac{1}{2}\|y - \tilde{Z}x\|_F^2 + \lambda\|x\|_1 \tag{4-10}$$

such that $y = Y^T_{(i,\star)} \in \mathbb{R}^N$, $x = X^T_{(i,V_i)} \in \mathbb{R}^{v_i}$ and $\tilde{Z} = Z^T_{(V_i,\star)} \in \mathbb{R}^{N\times v_i}$, where v_i denotes the number of elements in V_i . The problem (4-10) is also known as the LASSO (Tibshirani, 1996). In ADMM form, the LASSO problem is

$$\begin{aligned} \min_{x_1,x_2} \quad & \tfrac{1}{2}\|y - \tilde{Z}x_1\|_F^2 + \lambda\|x_2\|_1 \\ \text{s.t.} \quad & x_1 - x_2 = 0 \end{aligned} \tag{4-11}$$

The ADMM updating steps for the LASSO can be found in Appendix A. The closed-form expressions of this particular example are given by

$$x_1^{k+1} = (\tilde{Z}^T\tilde{Z} + \rho I)^{-1}(\tilde{Z}^T y + \rho(x_2^k - u^k)) \tag{4-12}$$

$$x_2^{k+1} = S_{\lambda/\rho}(x_1^{k+1} + u^k) \tag{4-13}$$

$$u^{k+1} = u^k + x_1^{k+1} - x_2^{k+1} \tag{4-14}$$

where $S_a(b) = \max(0, 1 - a/|b|)\, b$ is the element-wise soft-thresholding operator for a vector b. In addition, the step with the highest complexity (4-12) shows the drastic improvement to the computational efficiency that the graph approach has realized by decimating the size of the matrix Z. When the matrix Z would be left complete, the large inverse would become a computational obstacle at large dimensions, scaling cubically with the number of sensors.

The stopping criteria of the algorithm are defined following the approach of Boyd et al. (2011). In this work, an absolute and relative criterion are used to find feasibility tolerances for the primal and dual feasibility conditions. For (4-11), the optimality conditions are

$$\|r^k\|_2 = \|x_1^k - x_2^k\|_2 \leq \sqrt{n_x}\epsilon^{abs} + \epsilon^{rel}\max(\|x_1^k\|_2, \|x_2^k\|_2) \tag{4-15}$$

$$\|s^k\|_2 = \|x_2^k - x_2^{k-1}\|_2 \leq \sqrt{n_x}\epsilon^{abs} + \epsilon^{rel}\|\rho u^k\|_2 \tag{4-16}$$

where n_x is the size of the vectors x_1 and x_2 and ϵ^{abs} and ϵ^{rel} represent the absolute and relative tolerance with typical values around 10^{-3} or 10^{-4}.

The algorithm (4-12)-(4-14) has two important parameters: ρ and λ, that have to be tuned for the desired sparsity and speed of convergence. Usually, λ is kept constant, while ρ can be updated during the iterations. A simple scheme that often works well is Boyd et al. (2011)

$$\rho^{k+1} = \begin{cases} \tau\rho^k & \text{if } \|r^k\|_2 > \mu\|s^k\|_2 \\ \rho^k/\tau & \text{if } \|s^k\|_2 > \mu\|r^k\|_2 \\ \rho^k & \text{otherwise} \end{cases}$$

such that the primal and dual residual norms remain within a factor μ of one another. For the proof of convergence of the ADMM algorithm for the LASSO problem the reader is referred to Boyd et al. (2011). In general, the ADMM algorithm converges quickly to a modest accuracy, but can be very slow to converge to high accuracy.

4-2-3 Computational complexity

To quantify the decrease in computational complexity, usually the big O notation ($\mathcal{O}(\cdot)$) is used to describe the limiting behaviour of the algorithm when the dimensions tend towards infinity. Let us assume that in all sets V_i there are on average v elements, then step (4-12) scales with $\mathcal{O}(Nv^2 + v^3)$ operations per iteration. The second and third ADMM update scale linearly with v, independent of N, and can thus be neglected. Assuming an average number of M ADMM iterations, the total number of operations becomes $\mathcal{O}(NMv^2p + Mv^3p)$ over all p separate optimization problems. In addition, because all p problems are completely independent, they can be distributed and computed in parallel over D different processors. Hence, the average number of computations per processor equals approximately $\mathcal{O}(\frac{NMv^2+Mv^3}{D}p)$. Furthermore, since v is independent of the number of outputs p, it can be concluded that the VARX method scales linearly with the number of sensors of the system. Or alternatively, considering a WFS measurement grid of $n \times n$ lenslets, the method scales with $\mathcal{O}(n^2)$. This forms a large contrast compared to the derivation of a dense state-space innovation model which solves a Riccati equation that scales with $\mathcal{O}(n^6)$.

4-3 Numerical validation of sparse VARX identification

In this section, the presented identification algorithm is tested in a validation study comparing the VARX approximation with the original Kalman filter. First, the simulation procedure and the performance metrics are discussed, followed by the main results In theory, the accuracy of the identified VARX model will always be lower than the innovation model. Nevertheless, it has the advantage that it is completely data-driven, has a linear computational complexity and results into a highly sparse model.

4-3-1 Simulation procedure

The simulated turbulence is based on a single layer satisfying the Von Kármán spectrum for $r_0 = 0.16m$ and $L_0 = 10m$. The turbulence dynamics is modelled by an identified second order VAR model following Assémat et al. (2006), based on the theoretical spatial covariance (3-13) and Taylor's frozen turbulence hypothesis (3-16). Furthermore, the wind direction is perfectly

aligned with the x-direction of the grid and it moves with a speed of $v_x = 10m/s$, resulting in a Greenwood frequency of $f_G = 26.7Hz$. The WFS is represented by a square of 16-by-16 lenslets and the actuators are placed in Fried geometry. The sampling frequency $f_s = 250Hz$ and the inter-lenslet/actuator spacing $\delta = v_x/f_s = 0.04m$ such that for each sampling period, the turbulence moves exactly a distance of one δ. Moreover, the DM influence functions are described by (3-30) for a coupling of 30%. The measurement noise in the WFS adds a white noise signal with $C_v = \sigma I$, where σ is such that the SNR is 20 dB. The algorithm parameters that need to be taken into consideration are the regularization parameter λ and the VARX orders n_a and n_b. The model order is based on the assumptions in Section 4-1, i.e. $n_a = 1$, $n_b = 2$. The regularization parameter has to be tuned such that an optimal trade-off is achieved between accuracy and sparsity.

Each simulation runs over $N = 5000$ time samples with an input drawn randomly from a normal distribution with unitary variance and a certain realization of the turbulence under the aforementioned conditions. The validation data are newly drawn realizations of the input and turbulence from the same spectra as used in the identification. In addition, each experiment (identification plus validation) is repeated for n_r different realizations. The results will be presented in terms of mean values and standard deviations over these n_r different realizations.

One way of representing the accuracy of a model, assumed that in simulation the real signals are known, is the variance accounted for (VAF). The variance accounted for is defined as

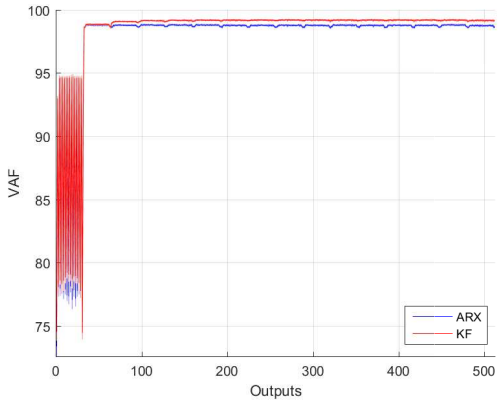
$$\mathrm{VAF}(y(k), \hat{y}(k)) = \max\left(0, \left(1 - \frac{\frac{1}{N}\sum_{k=1}^{N} \|y(k) - \hat{y}(k)\|_2^2}{\frac{1}{N}\sum_{k=1}^{N} \|y(k)\|_2^2}\right) \cdot 100\%\right) \tag{4-17}$$

where y is the real output and $\hat{y}$ the output predicted by the model. This definition will give a value between 0% and 100%, with 100% meaning a perfect fit. The VAF can be computed both for each output separately, or for all outputs together. Normally, the average value over all signal elements is taken such that only a scalar value describes the performance of the whole estimator.

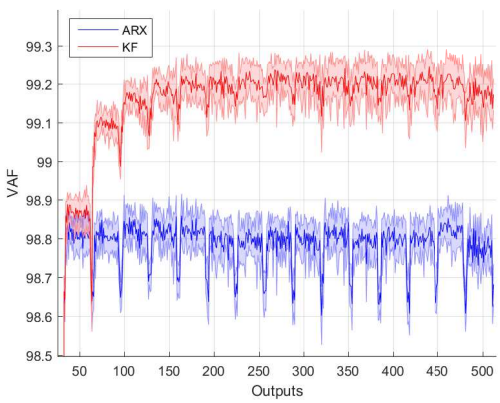
The sparsity of the model is described by the number of non-zero elements in each of the matrices A_i and B_i. By dividing it by the total number of elements in the matrix, a normalized sparsity is obtained that represents the fraction of non-zero elements in each matrix.

4-3-2 Numerical validation

First, the standard case discussed above is tested for 20 different realizations with the non-regularized least-squares solution (i.e. $\lambda = 0$). The results in terms of VAF for each sensor output separately is shown in Figure 4-2. Clearly, the VAF of the first 34 outputs, corresponding to the edge of the grid where the new unknown turbulence enters, is much lower than the rest of the outputs. The turbulence on this edge cannot be predicted accurately as there only is statistical knowledge available, while at the rest of the grid the turbulence is just a shifted version of the turbulence measured one time step before. Moreover, the VARX model has an average VAF slightly smaller than the Kalman filter. The normalized sparsity fraction under these circumstances are 0.038, 0.052 and 0.173 for A_1, B_1 and B_2 respectively.



(a) VAF for each output for for the Kalman filter and VARX predictor on the validation data
.



(b) Detail of the same figure. The thicker lines represent the mean and the transparent field the standard deviation over the different realizations.

Figure 4-2

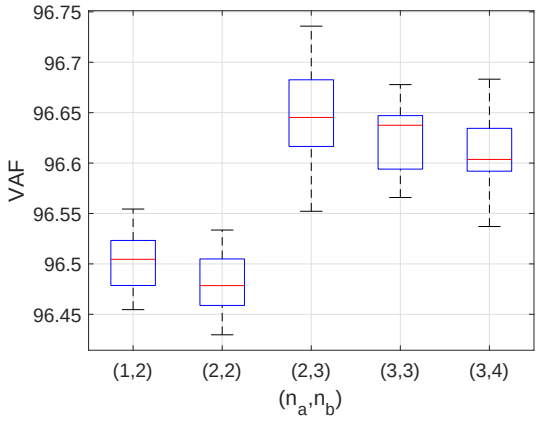


Figure 4-3: Boxplot of the accuracy of the VARX model for different combinations of n_a and n_b over 20 different realizations.

This small difference in accuracy might be caused by the low-order VARX approximation. However, since the difference is very marginal, it is already clear that increasing the VARX order will not improve the accuracy of the model significantly. To illustrate the trade-off between model order and accuracy, a number of combinations between n_a and n_b are tested in Figure 4-3. Even according to the approximation, $n_a = 1$ and $n_b = 1$ does not contain enough terms to become an accurate model and therefore is not considered. It is clear that orders larger than $n_a = 1$, $n_b = 2$ only improve the accuracy with 0.1%, supporting the approximation proposed in Section 4-1. When the order is increased even further, the increasing number of parameters even seem to cause a decrease in VAF. In conclusion, the accuracy varies only slightly and with an eye on the trade-off between computational complexity and accuracy, increasing the model order is not useful. However, it is important to note that this low order approximation only holds since the turbulence simulation used in this validation study behaves perfectly frozen. When the turbulence gets more complex, it is expected that the

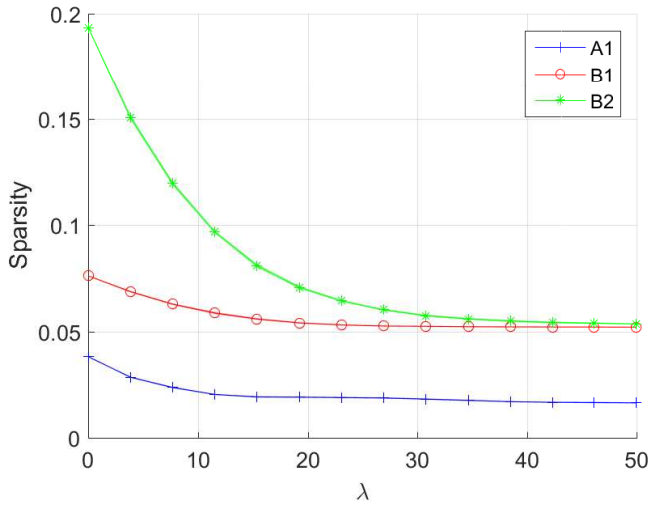


Figure 4-4: The influence of ℓ_1-regularization on the fraction of non-zero elements in the VARX coefficient matrices.

VARX model order has to increase to maintain a sufficient accuracy. The implementation of the method in more realistic circumstances forms an interesting subject for future research.

The sparsity induced by the assumed graph topology could be improved further by introducing ℓ_1-regularization and solving the LASSO problem with ADMM as presented in Section 4-2. The influence of the regularization parameter λ on the sparsity is shown in Figure 4-6. Clearly, the sparsity can be decreased significantly without decreasing the accuracy too much. The most improvement is seen in A_1 and B_2, which can be explained by their dependence on the wind direction, i.e. many elements of the chosen circle that are not approximately in the same direction as the wind are actually zero. The accuracy decreases very gradually when the sparsity increases. The highest accuracy is reached for $\lambda = 0$ with a VAF of 97.7% and for $\lambda = 50$ it is 96.8%.

To see how the identification performs as the turbulence gets heavier, the Fried parameter is decreased, resulting in an increase of the Greenwood frequency. In Figure 4-5, r_0 is varied between 0.03m and 0.3m. Both the accuracy of the VARX model and the innovation model decrease when the Fried parameter decreases. It can be concluded that within this range, the VARX model does not have problems modelling heavy turbulence as it follows the same trend as the Kalman filter.

Finally, the performance of the VARX model is tested for different signal to noise ratios. A clear decrease in performance of the VARX model compared to the innovation model is visible as the noise gets heavier. The trend is explained by the fact that Kalman gain of the innovation model takes the measurement noise into account, while the VARX identification algorithm is overfitting the noise. Possible solutions to this problem include adding an extra regularization term to decrease the sensitivity to the noise or increasing the VARX model order. This discussion will be continued in Section 4-5 where the influence of the SNR on the control performance will be investigated.

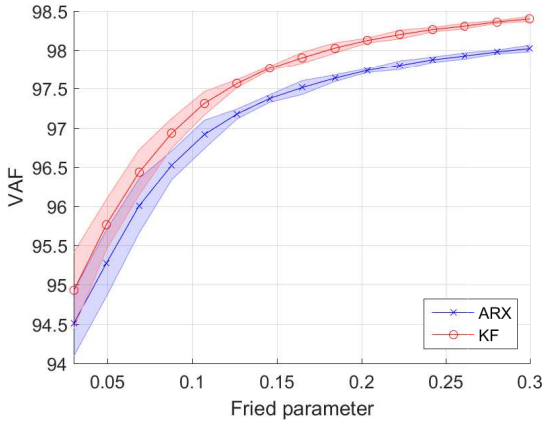


Figure 4-5: Influence of the Fried parameter, displaying the mean VAF and corresponding standard deviation over the different realizations.

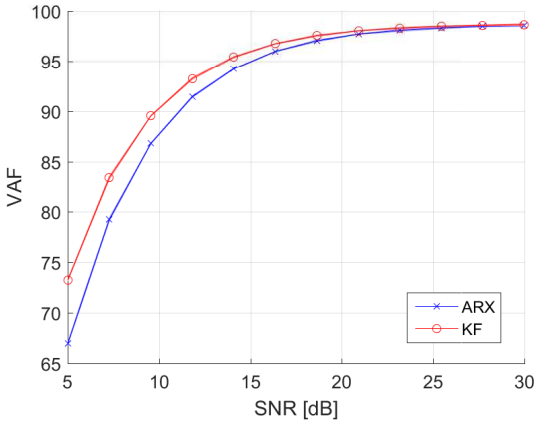


Figure 4-6: VAF versus the SNR for the VARX- and KF-based methods.

The sparse VARX model identification is only the first step of the new algorithm. Next, the new efficient optimal control strategy will be discussed and be tested in a similar validation study in Section 4-5.

4-4 Sparse optimal control for AO

Suppose that a predictor for the VARX model has been successfully identified in the sparse format (4-1). According to (3-42), a control law has to be derived such that the mean squared error of the residual wavefront predictor $\hat{\epsilon}(k+1)$ is minimized. Since the model gives a predictor of $s(k)$ rather than $\epsilon(k)$, the wavefront reconstruction step of (3-33) would be necessary to obtain our predictor, i.e. $\hat{\epsilon}(k+1) = F\hat{s}(k+1)$. However, the matrix F is not sparse, nor can it be enforced sparse without loosing accuracy since the covariance matrix C_ϕ is generally a dense matrix. Moreover, the linear relation between the wavefront residual and the WFS signal highlights the fact that minimizing $\hat{s}(k+1)$ could also serve as objective function. The main difference between both is the fact that the phase contains two unobservable modes (piston and waffle) as discussed in Section 3-4-1 that are not visible in the slopes. Since the piston mode does not influence the image quality and the waffle mode only has very little energy in a Kolmogorov spectrum due to its high frequency, it is chosen to neglect both. Another difference is that minimizing $\hat{s}(k+1)$ flattens the wavefront slopes, resulting in a larger compensation for the higher frequencies. In conclusion, since efficiency and sparsity are very important to this method, it is chosen to replace $\hat{\epsilon}(k+1)$ by $\hat{s}(k+1)$ in the objective function (3-42).

4-4-1 The control objective function

Let the following compact notation be used for the model (4-1):

$$s(k) = y(k-1) + B_1 u(k-1) + e(k) \tag{4-18}$$

with $y(k-1) := \sum_{i=1}^{n_a} A_i s(k-i) + \sum_{i=2}^{n_b} B_i u(k-i)$. A predictor of the slopes of the wavefront according to the model (4-18), neglecting the stochastic signal $e(k)$, could be written as

$$\hat{s}(k+1) = y(k) + B_1 u(k) \tag{4-19}$$

The optimal control input is defined as $u(k)$ that minimizes the mean squared error of the predicted wavefront slopes, i.e.

$$\begin{aligned} \min_{u(k)} \quad & \|\hat{s}(k+1)\|_2^2 \\ \text{s.t.} \quad & \hat{s}(k+1) = y(k) + B_1 u(k) \end{aligned} \tag{4-20}$$

which by substitution evolves into the sparse least-squares problem

$$\min_{u(k)} \quad \|y(k) + B_1 u(k)\|_2^2 \tag{4-21}$$

This minimization problem has to be solved at each time instance. An important property of the identified VARX model the fact that all matrices in this problem are very sparse. The sparsity of the matrices within $y(k)$ and B_1 can be exploited to solve the problem more efficiently (see e.g. Paige and Saunders, 1982). In addition, B_1 is also highly structured and has Kronecker rank 2. All of these properties create many opportunities to a scalable control law that are not possible in the existing control frameworks.

For the minimum variance least squares estimate, one should compensate for the stochastic nature of $e(k)$ in the identified model of (4-1), resulting in a weighted least-squares problem (see e.g. Verhaegen and Verdult, 2007, Sec. 2.7) given by:

$$\|\hat{s}(k+1)\|_{C_e^{-1}}^2 = \|y(k) - B_1 u(k)\|_{C_e^{-1}}^2 \tag{4-22}$$

such that the control law results in the following minimization problem

$$\min_{u(k)} \quad \|y(k) + B_1 u(k)\|_{C_e^{-1}}^2 \tag{4-23}$$

where $\|x\|_W = x^T W x$. This weighted least-squares problem can be solved efficiently if the matrix C_e^{-1} is sparse. With the knowledge of the underlying graphical model of the AO system and according to the influence of conditional independence relations on the inverse of the covariance (see the definition in Section 2-2-2), the inverse covariance is indeed expected to be highly sparse.

Moreover, the weighted least-squares problem can be interpreted as a standard least-squares problem by considering the model of the form (4-2) rather than (4-1). If we define a matrix $A_0 = C_e^{-1/2}$, the estimate in (4-22) is equivalent to

$$\|A_0 \hat{s}(k+1)\|_2^2 = \|\bar{y}(k) + \bar{B}_1 u(k)\|_2^2$$

with $\bar{y}(k-1) = \sum_{i=1}^{n_a} \bar{A}_i s(k-i) + \sum_{i=2}^{n_b} \bar{B}_i u(k-i)$ and $\bar{A}_i = A_0 A_i$ and $\bar{B}_i = A_0 B_i$ the matrices from (4-2). The control objective is to minimize this estimate, i.e.

$$\min_{u(k)} \quad \|\bar{y}(k) + \bar{B}_1 u(k)\|_2^2 \tag{4-24}$$

To put it in other words, we can rewrite (4-1) as (4-2), such that the noise term has an identity covariance matrix. The least-squares for the right hand side of this equation will lead to the minimum variance control law.

Even though there are many conditional independence relations describing the Gaussian signal $e(t)$, inverting its sample covariance matrix generally does not result in an exact sparse matrix. Also, truncating the very small values might drive the resulting VARX model unstable and is not a reliable option. Therefore, the task that still remains is to find a scalable routine to estimate the inverse covariance C_e^{-1}, or its square root represented by A_0, as sparse as possible. This will be of interest in the next paragraph.

4-4-2 Estimating a sparse covariance matrix

Chapter 2 introduced the problem of estimating a sparse inverse covariance matrix from a dataset consisting of samples from a zero mean Gaussian distribution. With the model of the previous section and real measurements, we can find the error signal $e(k)$ which was assumed to have a Gaussian distribution $\mathcal{N}(0, C_e)$. From a large number of samples of $e(k)$, a sample covariance matrix S_e is constructed via (2-7). The topology of graphical models is usually determined using the so-called covariance selection problem:

$$\hat{C}_e^{-1} = \arg\min_X \; -\log\det(X) + \text{trace}(S_e X) + \gamma\|X\|_1 \tag{4-25}$$

which results in the maximum likelihood estimate for $\gamma = 0$. Increasing γ trades off maximality of the likelihood for sparsity. This problem can be solved using ADMM (see Appendix A), but various other efficient algorithms have been proposed lately that solve the covariance selection problem (see an overview in Section 2-2-4).

When the stochastic nature of $e(k)$ is neglected, a direct sparse estimate of S_e^{-1} can be computed to approximate C_e^{-1}. By defining the new variable $X = S_e^{-1}$, the estimation amounts to solving a regularized least squares problem:

$$\hat{S}_e^{-1} = \arg\min_X \|I - XS_e\|_F^2 + \gamma\|X\|_1$$

such that, with a similar reasoning to Section 4-2-1, we can solve this row-by-row only for the non-zero elements that we can define from the spatial 2D graph topology that is expected for $e(k)$. Listing the measurement locations that connect with location i as $V_{e,i}$, and using the notation of Section 4-2-1, the reduced optimization problem becomes:

$$\min_{X_{(i,V_{e,i})}} \|I_{(i,\star)} - X_{(i,V_{e,i})}(S_e)_{(V_{e,i},\star)}\|_F^2 + \gamma\|X_{(i,V_{e,i})}\|_1$$

The optimization problem above is much more efficient than the covariance selection problem because of its distributed nature; however, it does not find the maximum likelihood solution. Furthermore, it should be noted that this optimization problem, in contrast to the covariance selection problem, does not specifically enforce positive definiteness of the variable. Since in practice S_e is diagonally dominant, it does not cause any problems in the considered simulations and is very accurate in practice. It is still a question for future research how to efficiently make this derivation more robust.

4-4-3 Advantages of the sparse VARX model based control method

There are numerous advantages of the presented VARX based control method and most of them already have been shortly mentioned. The main advantage over the MVM method (Section 3-6-2) is the fact that an accurate model of the AO loop and turbulence is used to accurately predict the wavefront slopes $s(k)$.

Moreover, the identification is completely data-driven, where many other methods are based on having accurate statistical knowledge of the atmospheric turbulence (such as C_ϕ). This information is normally necessary for reconstructing the wavefront $\epsilon(k)$ from measurements $s(k)$. In the data-driven approach we assume that we have no knowledge of this turbulence except for the information that is available in the measurements $s(k)$.

Compared to the optimal approaches in the literature, the performance of the new method is always inferior, since the VARX model is an approximation of the innovation model that forms the basis of those methods. However, as was already stated in Section 4-2-3, the linear complexity of the identification routine makes the method superior in terms of scalability. The large-scale Riccati equation that scales cubically with the number of outputs is replaced by a separable least-squares problem that can be solved in parallel.

The sparse VARX model also has the advantage of creating a more efficient control problem. The control action requires a number of sparse matrix-vector multiplications and a solution to a sparse least-squares problem, where all other methods do something similar but with dense matrices. Exploiting both the sparsity and other structures within the least-squares problem, it is has the potential of scaling up to very large-dimensions without explicitly computing a pseudo-inverse.

4-5 Numerical validation of the new control method

In a similar fashion as Section 4-3, the new control method is tested in a validation study. Under the same conditions as before, the performance of the controller is compared to its Kalman filter based equivalent (see Section 3-6-3) and the classical MVM approach (Section 3-6-2). First, the simulation procedure and performance criteria are presented, followed by the results.

4-5-1 Simulation procedure and performance metrics

Before commencing to the validation study, the simulation procedure and a number of performance criteria need to be defined in order to quantify the improvements for all different control methods.

The simulation of the turbulence dynamics is changed compared to the case discussed in Section (4-3). In the identification validations, the wind speed was assumed to be such that it always moves exactly one inter-lenslet spacing δ per sampling time. However, when the wind speed is varied, choosing only integer multiples of δ per sampling time is way too restrictive to represent a useful range. On the other hand, problems might occur when the turbulence does not move in multiples of δ per sampling time. When the movement is (approximately) an

integer number, the estimations are much more accurate than when it falls right in between, since interpolation will cause errors. Therefore, the turbulence is simulated on a fictive fine grid and the WFS centroid locations are centred between four non-adjacent phase grid points. The wind velocity is chosen such that it moves an integer number of this finer grid sample spacings each sampling time. This method makes it possible to select a wider range of wind velocities without having to cope with interpolation issues.

After the model has been identified, M new realization of the same turbulence spectrum are created over a timespan of N samples. The corresponding sensor data of these turbulence realizations, without any action of the DM, will serve as the base line for quantifying the performance. Also the MVM, Kalman filter based and new sparse VARX control method will be applied on each of these realizations. The performance criteria, which will be discussed below, are stored for each method and realization and compared to the situation with no control and each other. From the values of the criteria over all M realization, the mean and standard deviation are extracted and will be presented in this study. Moreover, it should be expected that the MVM method is an improvement compared to the base line and that the VARX- and KF-based optimal methods have a similar performance, much better than MVM. Due to the fact that the VARX model is an approximation of the Kalman filter, a slightly lower performance is expected for the new method.

The goal of the controller is to minimize the mean squared error of the phase residual $\epsilon(k)$ at each iteration. The most straightforward performance measure would hence be the value of this objective function. Equivalently, this means that the variance of $\epsilon(k)$ is our performance measure, i.e. the lower the value of

$$\sigma_\epsilon^2 = \frac{1}{n}\text{trace}(E[\epsilon(k)\epsilon^T(k)])$$

the better the performance of the controller, where n is the number of elements in $\epsilon(k)$. However, since we only have a limited number of sample data, a sample estimate of the mean squared residual wavefront error is computed via

$$\hat{\sigma}_\epsilon^2 = \frac{1}{N}\sum_{k=1}^{N}\frac{1}{n}\epsilon(k)^T\epsilon(k) \tag{4-26}$$

This value will serve as the main performance metric in this validation study.

The most common performance metrics in adaptive optics use intensity information, in particular the point spread function (PSF). The *Strehl ratio*, defined as the ratio between the measured peak intensity of the corrected PSF and the perfect peak intensity of the unaberrated PSF,

$$S = \frac{I_m}{I_p}$$

is one of the most used measures to define the effects of distortions in an optical system. A perfectly compensated wavefront will result in a Strehl ratio of 1 and the lower the ratio gets, the worse the controller performance is. However, in the simulations there is no intensity information available such that the definition above is not very useful. As was shown in (Maréchal, 1947), the Strehl ratio is directly related to the residual phase variance via

$$S \approx \exp(-\sigma_\epsilon^2)$$

Since the Strehl ratio can be written in terms of σ_ϵ^2, the mean square phase error contains the same information as the Strehl ratio in this case.

In theory, the mean squared wavefront error σ_ϵ^2 is composed of a sum of various error sources, such as the mean squared fitting error, temporal error and angular anisoplanatism. In the case of Kolmogorov turbulence, the fitting error is caused by the fact that the DM cannot take any arbitrary shape and it is defined as

$$\sigma_f^2 = a_f \left(\frac{\delta_{act}}{r_0} \right)^{5/3} \tag{4-27}$$

where δ_{act} is the inter-actuator spacing, r_0 the Fried parameter and a_f is a coefficient depending on the influence functions (Hinnen, 2007). However, since an "ideal" mirror is assumed in the simulation with H full row-rank, this fitting error should become approximately zero in our observations.

The temporal error is caused by the fact that there are delays and other temporal limitations within the system inhibiting the the temporal compensation at larger wind speeds. No mirror and sensor dynamics are assumed in the simulations except from a single time delay. Hence this delay will be the only error source in the simulated closed-loop AO system. The effectiveness of the control strategy is directly related to the temporal error. For a classical controller such as the MVM method, it can be found that in case of Kolmogorov turbulence, the mean-square temporal error is approximately

$$\sigma_t^2 = a_t \left(\frac{f_G}{f_s} \right)^{5/3} \tag{4-28}$$

with f_s the sampling frequency, f_G the Greenwood frequency and a_t a constant depending on the controller and bandwidth (Hinnen, 2007). Optimal control strategies, such as the method proposed in this chapter, should be able to compensate better for this time delay as it tries to take it into account by predicting the distortions one step ahead. However, since we consider a simplified version of the turbulence, it is not guaranteed that this curve will hold. Therefore, the exact steepness of the curve is not considered, but the curve of σ_ϵ^2 against the Greenwood frequency in general is still of interest in our validation.

The controller performance is tested in a validation experiment that investigates the influence of f_G/f_s on the mean squared error by changing the wind velocity. Furthermore, since it was shown to influence the accuracy of the identified model, the SNR of the measurement noise is varied. The atmospheric turbulence conditions remain the same with $r_0 = 0.16$m and $L_0 = 10$m. The only parameters that will be varied are the wind speed and the signal to noise ratio.

In the experiment, the WFS is an 11-by-11 grid with a spacing of $\delta = 4$cm between each location. The turbulence is propagated on finer a 34-by-34 grid, such that each sampling location is separated by $\delta^* = 4/3$cm. The sample frequency is kept constant at 250 Hz. The wind speed is varied such that it moves between 1 to 10 times δ^* per sampling time along the grid, corresponding to a wind speed ranging between 3.33m/s and 33.3m/s and a Greenwood frequency $f_G = 0.427v/r_0$ between 8.9 Hz and 89 Hz. The experiments use one identified model for each ratio, and the control performance is measured over 20 different realizations of turbulence, simulating over 1000 sampling periods. The average of the mean squared phase error (4-26) over these realizations is used as a measure of the control performance.

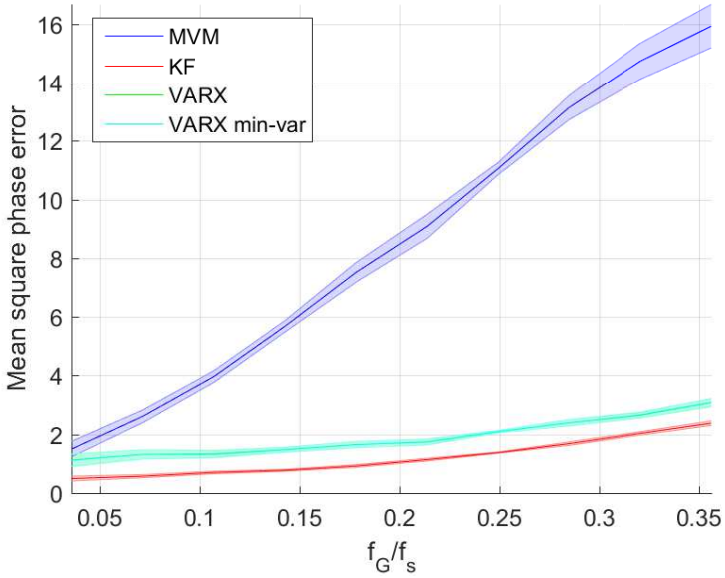


Figure 4-7: Performance of the different controllers with respect to different Greenwood frequency to sample frequency ratios. The line represents the mean over all realizations and the surface the corresponding standard deviation.

4-5-2 Numerical validation

The numerical validation of the control is split into two parts. First, the wind speed is changed to demonstrate the evolution of the temporal error. Afterwards, the SNR is increased to see how the method performs under a range of noise variances.

Figure 4-7 shows the influence of the Greenwood frequency to sample frequency on the mean squared error of the phase residual. Clearly, the line corresponding to the MVM method increases much faster than the new control law. This result supports the effectiveness of the new method compared to the classical approach. There is a small decrease in performance visible in relation to the Kalman filter based optimal controller. This difference remains approximately constant for all wind speeds and is explained by the more accurate predictor of the Kalman filter as shown in Section 4-3. Moreover, the Kalman filter based method in this simulation has the additional advantage that the grid spacing of the estimated phase residual $\epsilon(k)$ is three times as small as that of $s(k)$. Therefore, the VARX model method indirectly still feels the consequences of the interpolation, since it is solely based on the wider spaced WFS signal. It is expected that in real experiments, the Kalman filter and VARX methods will get closer to each other, as was also visible in terms of the accuracy in Section 4-3. The essential conclusion that can be derived from this experiment is that the VARX method significantly outperforms MVM and stays close to the performance of the Kalman filter based method for a wide range of Greenwood to sample frequency ratios. It also should be noted that the minimum-variance estimator does not improve the results significantly as it is indistinguishable form the non-minimum-variance control law in Figure 4-7.

Next, the SNR is varied between 5dB and 30dB with a wind velocity corresponding to a movement of one δ per sampling period. The mean squared phase error is plotted in Figure

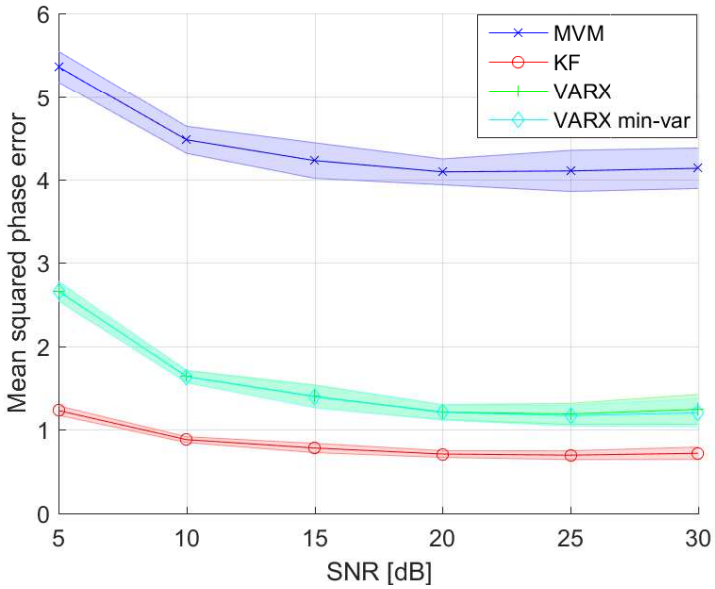


Figure 4-8: Performance of the different controllers with respect to different values for the SNR. The line represents the mean over all realizations and the surface the corresponding standard deviation.

4-8. The results resemble the trend that was visible during the validation of the model identification. When the noise variance gets larger, the performance of the VARX model based method seems to decrease in relation to the Kalman filter based controller. This can be explained by the fact that the least-squares problem identifying the VARX model is overfitting the noise without any form of compensation for it. Because of the lack of compensation for the measurement noise, the controller is not very efficient in situations where the light source is very dim, e.g. when observing a faint stars. It was discussed in Section 4-3 that possible solutions to this problem include adding an extra regularization term to decrease the sensitivity to the noise or increasing the VARX model order. This would mean that this problem could be solved very quickly, but this still has to be confirmed in future research. Furthermore, it is also noticeable that the minimum-variance estimator does not improve the performance at smaller signal to noise ratios.

In conclusion, the combination of accuracy and scalability shows the potential of the new method. The linear complexity and separable nature of the identification algorithm makes the method very suitable for large-scale applications. In addition, it has shown significant improvements over the classical control approach and a similar performance as its Kalman filter based equivalent, especially under moderate measurement noise levels. All major findings of this report will be summarized in the following chapter and a number of recommendations for future research are presented.

Chapter 5

Conclusions and Recommendations

This final chapter is divided into two sections. The first section will summarize the main conclusions presented in this thesis. Afterwards, a number of recommendations for further research are discussed in the second section.

5-1 Concluding Remarks

This thesis has presented a new scalable data-driven control method for large-scale adaptive optics systems. The fundamental idea that created the method was the realization that adaptive optics systems could be represented as graphical models, describing the conditional (in)dependence relations between each output with all other in- and outputs. The spatially separated nature of the WFS measurement grid combined with Taylor's frozen flow hypothesis sparked the idea that the evolution of one output element (i.e. one slope measurement of the WFS) is described only by the measurements and actuator inputs in a close neighbourhood. In a graphical modelling framework, VARX models are very effective as the graph topology can directly be mapped as a sparsity pattern on the coefficient matrices and the inverse of the stochastic input's covariance.

The aforementioned reasoning finds a rough estimation of the sparsity pattern before solving the identification problem. By only identifying the non-zero values, the number of elements per row of the VARX coefficient matrices is decreased significantly. This number is dependent on the nature of the sensor and actuator geometry, turbulence conditions and desired precision; however, it is independent of the size of the system. That is, if the size of the measurement grid is increased, it does not mean that one location is now influenced by a farther neighbour than it was before. Because of this property, it was shown that the complexity of the VARX model identification algorithm scales linearly with the total number of outputs of the system. This computational complexity forms a large contrast with the identification of a state-space innovation model, which requires solving a Riccati equation with a cubic complexity.

The identification routine requires one set of open-loop sensor data of the wavefront phase distortion due to turbulence plus a persistently exciting actuation of the DM. All non-zero elements of the VARX coefficient matrices are identified in a separable least-squares optimization framework. The sparsity of the model can be improved even more by adding an ℓ_1-regularization term and solving the regularized least-squares problem in parallel with the Alternating Direction Method of Multipliers (ADMM). The increase in sparsity is due to the overestimation of the region of influence in the graphical model and the fact that the wind speed is only in one constant direction, while the selected neighbourhood stretches in the same length along the direction of the wind speed as opposite or orthogonal to it. All wrongly assumed conditionally dependence relations causing unnecessary non-zero values in the coefficient matrices will in this step be forced to zero.

Given the identified model of the complete system, the AO control problem can be expressed in an optimal control framework. However, the identified model gives the slopes of the wavefront rather than the phase values. Where classical control methods first reconstruct the phase from the measured slopes, it is chosen not to do so in the new method. The main reason is that the reconstruction step would destroy the obtained sparsity in the model. Moreover, the most important difference between the real phase and the measured slopes are the unobservable piston and waffle modes. Since the piston mode does not influence the image quality and the high frequency waffle mode has only a very low energy, they are neglected in this thesis. Therefore, the new control law computes at each time instance the control input that minimizes the one-step-ahead prediction of the mean squared error of the phase slopes. The achieved sparsity of the control problem should be exploited to solve the resulting large least-squares problem efficiently.

A minimum-variance control law requires the least-squares control problem to be weighted with the inverse of the spatial covariance of the VARX model's stochastic input. If the same degree of sparsity has to be maintained, the estimated covariance matrix should also be approximately as sparse as the coefficient matrices. A certain trade-off between maximum likelihood and sparsity of the estimation can be found by solving the so-called covariance selection problem. Solving this problem for large dimensions is very challenging and has drawn a lot of attention in the literature over the last decade. Another estimate can be obtained by directly computing a sparse estimate of the inverse sample covariance matrix in a separable least-squares framework similar to the one used in the VARX model identification. An advantage of this approach is that similar to the identification algorithm, all conditional independence relations known beforehand can be used to decrease the dimensionality of the system, resulting in a linearly scaling computational complexity.

The proposed data-driven sparse VARX model identification has been validated in simulation. The turbulence is generated as a single layer of Kolmogorov turbulence following Taylor's frozen flow hypothesis. The experiments have shown that the sparse VARX approximation has almost equal accuracy to an identified dense state-space innovation model. Furthermore, ℓ_1-regularization will increase the sparsity caused by the unidirectional wind speed in some of the matrices significantly, only decreasing the VAF value with less than one percent. Moreover, any severity of turbulence can be modelled approximately with the accuracy of the innovation model. The only difference is visible when the measurement noise variance is increased. The accuracy of the VARX model deteriorates faster compared to the accuracy of the innovation model when the SNR is decreased. This can be explained by the fact that the least-squares problem identifying the VARX model is overfitting the noise.

Also the controller performance is tested in a validation study, comparing the new method with a classical control law without any predictive capability and an optimal controller based on a Kalman filter. The simulations have demonstrated that the new control method outperforms the conventional controller, especially for larger Greenwood to sampling frequency ratios, and that it obtains a similar performance to the Kalman filter based technique. However, when the SNR is decreased, the temporal error of the new method starts to increase in comparison with the optimal controller. This decrease in performance for high measurement noise variances makes the method less suitable for correcting faint light sources. The overfitting problem might already be solved by introducing an extra regularization term to compensate for the noise, but this still has to be confirmed in future research. Also, the minimum-variance control law obtains the same performance as the non-minimum-variance controller under all circumstances.

In conclusion, the combination of accuracy and scalability demonstrates the potential of the new method. The linear complexity and parallelizable algorithm make the method very suitable for large-scale applications. It has shown significant improvements over the classical control approach, and a similar performance to the much more complex optimal control law.

5-2 Recommendations

Despite the fact that the new procedure has been shown to have great potential as an efficient control method for large-scale AO systems, the research in this field is far from finished. A number of suggestions for future research are presented below.

First of all, the decrease in performance for lower signal to noise ratios should be addressed. As discussed before, the overfitting of the noise in the identification algorithm might be decreased by simply increasing the VARX model order or by adding an extra regularization term to compensate for the measurement noise. These possibilities should be investigated in further simulations and the improvements should be compared to the performance of the Kalman filter based controller.

The results were obtained under ideal circumstances in simulation. Turbulence is very difficult to simulate reliably and system dynamics might influence the accuracy of model identification, changing the behaviour of the controller. Therefore, it is necessary to test the method in more realistic circumstances in order to draw final conclusions. This means that either the simulations need to include the system dynamics and a less simplistic turbulence generation, or the algorithm has to be tested in a laboratory experiment. Since a realistic well-tested simulation environment is not directly available, testing the method on a practical AO system seems to be the best option of the two. The turbulence can be simulated by a circular plastic plate which is sprayed or machined in such a way that the resulting wavefront distortions have a Kolmogorov distribution. By adjusting the rotational speed, it is possible to simulate different wind speeds to validate the influence of the Greenwood frequency on the controller performance. A similar simulation has been realized in the study of Hinnen (2007).

It should also be noted that even this type of experiments might still use a too restrictive representation of the turbulence. Therefore, it would be of great interest to have a clear, easily accessible and well validated simulator that can model both the turbulence and complete AO

system very accurately. With such a toolbox, it would be possible to have a clear comparison between the performance of different control algorithms.

The new method has so far only been tested against the classical MVM method and a simple Kalman filter based optimal control method. Further validation of the performance should also include a comparison to the current state of the art AO control algorithms. It would be very interesting to see if or under which conditions the new method can compete with state of the art insights.

In practical applications, the turbulence conditions will slowly vary over time. To ensure an accurate prediction, the identified model has to be updated as this happens. It would be a challenge to construct a routine that detects big changes in the turbulence conditions and updates the model during operation based on closed-loop WFS data, without breaking off the observations. This method should then be compared to adaptive control strategies such as the method proposed in Ellerbroek and Rhoadarmer (2001).

It was shown that by adding the regularization term, the sparsity of the VARX coefficient matrices could be increased significantly. One cause is the fact that the causality relations in the graphical model describing the turbulence dynamics in reality do not go in both directions but are dependent on the direction of the wind. So when the wind direction is (approximately) known, at least half of the assumed causality relations can be removed. In other words, the assumed rough topology that reduces the dimension of the identification problem can be fine-tuned if the weather conditions and other parameters are known more precisely. Of course, this also means that the model is less data-driven and more depending on certain input parameters, but it also might get the ℓ_1-regularization to become superfluous.

The sparse inverse estimation problem still requires an efficient and robust solution. The covariance selection problem has drawn a lot of attention in the literature and should be able to solve the problem for very large systems. For example the block-coordinate descent method of Hsieh et al. (2011, 2013) shows great potential. However, since the algorithm is very complex and difficult to implement, it is not an ideal solution. Furthermore, the separable least-squares problem estimating a sparse inverse sample covariance matrix requires additional constraints to guarantee positive definiteness and stability. Implementing these constraints is far from trivial and might increase the complexity significantly.

In this thesis, the piston mode has been ignored and removed from the results afterwards. The second unobservable mode, the waffle mode, which does have a deteriorating effect on the image quality has not been considered since it was assumed to have only a very limited influence. However, due to the fact that these modes do not influence the control objective function (the mean squared error of the wavefront slopes), they can appear in the control action where they will do more harm than good. In many AO control strategies, these modes are removed from the model, which was also shortly discussed in this thesis. If the modes are removed from the model, the computed input will never cause the piston and waffle mode in the DM shape. This reduced model changes the physical meaning and transforms the model into a modal basis. Therefore, the graphical modelling approach used to reduce the dimensionality does not hold any more and the sparsity might even be completely erased. However, another approach is not to remove the piston and waffle modes from the output, but rather during the computation of the optimal control input. Adding constraints on the control objective functions can also avoid the DM from creating these modes. Since both modes will not influence the wavefront slopes, regularization on the input might already take

care of this. It is an interesting question for future research to remove the unobservable modes from the input as efficient as possible.

Finally, the control problem requires the solution of a very large-dimensional but highly sparse least-squares problem. There have been several tools proposed to avoid explicitly computing the pseudo-inverse by exploiting the sparsity pattern. A possible routine to solve large-scale sparse least squares problems is the one of Paige and Saunders (1982). Besides the sparsity pattern in the matrix B_1, it is has several structural properties. The Kronecker product structure can be exploited by using it as a pre-conditioner for the sparse least-squares problem (Bardsley et al., 2011) or using the parallel method of Fausett et al. (1997). Another possibility is to use pre-conditioned conjugate gradient methods similar to the one used to solve the wavefront reconstruction problem in Gilles et al. (2002). Undoubtedly, there are many other possibilities to obtain an efficient solution to the large-scale least squares problem and it forms a very interesting topic for future research.

Appendix A

Alternating direction method of multipliers (ADMM)

ADMM is an algorithm belonging to the class of proximal algorithms which are a tool for solving non-smooth, constrained, large-scale, or distributed versions of convex optimization problems. A detailed overview of proximal algorithms, their interpretations and applications can be found in the monograph of Parikh and Boyd (2013).

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{-\infty\}$ be a closed proper convex function. The proximal operator $\mathbf{prox}_{\lambda f} : \mathbb{R}^n \to \mathbb{R}^n$ of a scaled function λf is defined by

$$\mathbf{prox}_{\lambda f}(v) = \arg\min_x \left(f(x) + \frac{1}{2\lambda}\|x - v\|_2^2 \right) \tag{A-1}$$

This definition has many interpretations and it would be too extensive to explain them all. Multiple algorithms exist that use proximal operators. Examples are *proximal minimization*, the *proximal gradient method*, the *accelerated proximal gradient method* and the *alternating direction method of multipliers (ADMM)*. The ADMM has proven to be particularly efficient for large scale problems and is easily integrated in a distributed framework (Boyd et al., 2011).

The alternating direction method of multipliers (ADMM) can be used to solve problems of the form

$$\min_x f(x) + g(x)$$

where f, $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ are closed proper convex functions (but not necessarily smooth). The method consist of three steps.

$$x^{k+1} := \mathbf{prox}_{\lambda f}(z^k - u^k) \tag{A-2}$$

$$z^{k+1} := \mathbf{prox}_{\lambda g}(x^{k+1} + u^k) \tag{A-3}$$

$$u^{k+1} := u^k + x^{k+1} - z^{k+1} \tag{A-4}$$

In this method, x^k and z^k will converge to each other and to optimality. The advantage of this algorithm is that the objective functions are solved separately. Also, compared to the other proximal algorithms, it is more easily parallelized.

ADMM combines the convergence properties of the method of multipliers with the decomposability of the dual ascent. Consider an equality constrained optimization problem and split the variable in two parts, here denoted as x and z, i.e.

$$\begin{aligned} \min_{x,z} \quad & f(x) + g(z) \\ \text{s.t.} \quad & Ax + Bz = c \end{aligned} \tag{A-5}$$

with $x \in \mathbb{R}^n$, $z \in \mathbb{R}^m$, $A \in \mathbb{R}^{p\times n}$, $B \in \mathbb{R}^{p\times m}$ and $c \in \mathbb{R}^p$. Furthermore, f and g are considered convex. Besides the formulation of proximal operators (A-2), ADMM can be formulated using the augmented Lagrangian

$$x^{k+1} := \arg\min_x L_\rho(x, z^k, y^k) \tag{A-6}$$

$$z^{k+1} := \arg\min_z L_\rho(x^{k+1}, z, y^k) \tag{A-7}$$

$$y^{k+1} := y^k + \rho\left(Ax^{k+1} + Bz^{k+1} - c\right) \tag{A-8}$$

with $\rho > 0$ and L_ρ, the augmented Lagrangian of (A-5), is defined as

$$L_\rho(x, z, y) := f(x) + g(y)y^T(Ax + Bz - c) + \rho/2\|Ax + Bz - c\|_2^2$$

(B-6) is the x-minimization step, (A-7) the z-minimization step and (A-8) the dual variable update using ρ as step size. Often the ADMM algorithm is written in the so called 'scaled form', by combining the linear and quadratic terms in the Lagrangian and scaling the dual variable

$$x^{k+1} := \arg\min_x \left(f(x) + \rho/2\|Ax + Bz^k - c + u^k\|_2^2\right) \tag{A-9}$$

$$z^{k+1} := \arg\min_z \left(g(z) + \rho/2\|Ax^{k+1} + Bz - c + u^k\|_2^2\right) \tag{A-10}$$

$$u^{k+1} := u^k + Ax^{k+1} + Bz^{k+1} - c \tag{A-11}$$

where $u = y/\rho$ is the scaled dual variable. Note that if we consider the x-update step (A-9) we get

$$x^+ = \arg\min_x \left(f(x) + \rho/2\|Ax - v\|_2^2\right)$$

where $v = -Bz + c - y$ is a constant. The right hand side is clearly equivalent to the proximal operator $\mathbf{prox}_{\lambda f}(v)$ for $\lambda = 1/\rho$. By symmetry, the same can be derived for the z-update (A-10).

A-1 ADMM for ℓ_1 regularized minimization

ADMM can be very efficiently used for ℓ_1 regularized problems. Consider the generic problem

$$\min \; l(x) + \lambda\|x\|_1 \tag{A-12}$$

where l is a convex function. In ADMM form, by taking $g(z) = \lambda\|z\|_1$, this is

$$\begin{array}{ll} \min & l(x) + g(x) \\ \text{s.t.} & x - z = 0 \end{array}$$

Whether or not there consists a closed-form solution to the x-update depends on $l(x)$. For the z-update, the closed-form solution can be found using subdifferential calculus (Boyd et al., 2011)

$$z^{k+1} := S_{\lambda/\rho}(x^{k+1} + u^k) \tag{A-13}$$

where S is called the *soft thresholding operator* defined as

$$S_\kappa(a) := \max\left(0, 1 - \frac{\kappa}{|a|}\right) a \tag{A-14}$$

with $S(0) = 0$. An important special case of the ℓ_1 regularized problem is the *lasso*, where $l(x)$ is the least squares $\frac{1}{2}\|Ax - b\|_2^2$ in (A-12). Note that the x-update (A-9) is essentially a *ridge regression* problem which has a closed-form solution

$$x^{k+1} := (A^T A + \rho I)^{-1}(A^T b + \rho(z^k - u^k))$$

Another important example is the *group lasso*, where the regularizer $\|x\|_1$ is replaced by a sum of ℓ_2 norms: $\sum_{i=1}^N \|x_i\|_2$. The x-update remains the same as for the lasso and the z-updates becomes

$$z_i^{k+1} = \mathcal{S}_{\lambda/\rho}(x_i^{k+1} + u^k), \quad i = 1, \ldots, N$$

where the *vector soft thresholding*, $\mathcal{S}_\kappa(a)$ is defined as

$$\mathcal{S}_\kappa(a) = \max\left(0, 1 - \frac{\kappa}{\|a\|_2}\right) a \tag{A-15}$$

with $\mathcal{S}_\kappa(0) = 0$, which reduces to (A-14) when a is a scalar.

A-2 ADMM for sparse inverse covariance selection

For the covariance selection problem, ADMM has already been succesfully applied (for example, see Scheinberg et al. (2010)). Recall that the problem at hand can be written as

$$\min_X \quad \operatorname{trace}(CX) - \log\det X + \lambda\|X\|_1 \tag{A-16}$$

with C the sample covariance matrix. To put it in the form of (A-12), $l(X) = \operatorname{trace}(CX) - \log\det X$. The ADMM algorithm for this problem is

$$\begin{aligned} X^{k+1} &= \arg\min_X \left(\operatorname{tr}(CX) - \log\det X + \frac{\rho}{2}\|X - Z^k + U^k\|_F^2\right) \\ Z^{k+1} &= \arg\min_Z \left(\lambda\|Z\|_1 + \frac{\rho}{2}\|X^{k+1} - Z + U^k\|_F^2\right) \\ U^{k+1} &= U^k + X^{k+1} - Z^{k+1} \end{aligned}$$

The algorithm can be simplified by noting that the Z-minimization step is elementwise soft thresholding

$$Z_{ij}^{k+1} := S_{\lambda/\rho}\left(X_{ij}^{k+1} + U_{ij}^{k+1}\right)$$

and the X-minimization can also be solved analytically. This can be derived by noting that the gradient should be zero at the optimum, i.e.

$$\begin{aligned} C - X^{-1} + \rho(X - Z^k + U^k) &= 0 \\ \rho X - X^{-1} &= \rho(Z^k - U^k) - S \end{aligned} \tag{A-17}$$

By taking the orthogonal eigenvalue decomposition of the right hand side (i.e., $\rho(Z^k - U^k) - S = Q\Lambda Q^T$, with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ and $Q^T Q = QQ^T = I$) and multiplying (A-17) by Q^T on the left and Q on the right this gives

$$\rho\tilde{X} - \tilde{X}^{-1} = \Lambda$$

where $\tilde{X} = Q^T X Q$, or equivalently we have that

$$\rho\tilde{X}_{ii} - \frac{1}{\tilde{X}_{ii}} = \lambda_i \tag{A-18}$$

$X = Q\tilde{X}Q^T$ satisfies the optimality condition (A-17), hence (A-18) is the solution of the X-minimization step. Note that the computational effort of the X-minimization step is reduced to just an eigenvalue decomposition of a symmetric matrix.

Appendix B

Sparse Nuclear Norm Subspace Identification

This appendix contains a study of finding a sparse Kalman filter from data to obtain an optimal one-step-ahead predictor of the turbulence. It should be noted that this algorithm itself requires a large computational effort and therefore is not directly applicable for the large-scale adaptive optics application. However, for systems with only a few inputs and outputs, it has shown to have more potential.

B-1 Problem description

As an alternative to the prediction error method presented in Chapter 4, a sparse state-space model can be derived using subspace identification. When we consider an model in innovation form (3-37), this amounts in identifying sparse system matrices $\bar{A}$, B, C, and K. Following a subspace identification approach, the following data equation is constructed:

$$Y_{i,s,N} = \mathcal{O}_s X_{i,1,N} + \mathcal{T}_s U_{i,s,N} + \mathcal{S}_s S_{i,s,N}$$

Where $Y_{i,s,N}$, $X_{i,1,N}$, $U_{i,s,N}$ and $S_{i,s,N}$ are Hankel matrices as defined in Section 2-1. Furthermore, $\mathcal{O}_s$ represents the extended observability matrix and $\mathcal{T}_s$ and $\mathcal{S}_s$ are two block lower-triangular Toeplitz matrices of the form

$$\mathcal{O}_s = \begin{bmatrix} C \\ C\bar{A} \\ C\bar{A}^2 \\ \vdots \\ C\bar{A}^{s-1} \end{bmatrix} \quad \mathcal{S}_s = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ CK & 0 & \cdots & 0 & 0 \\ C\bar{A}K & CK & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ C\bar{A}^{s-1}K & C\bar{A}^{s-2}K & \cdots & CK & 0 \end{bmatrix} \quad \mathcal{T}_s = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ CB & 0 & \cdots & 0 & 0 \\ C\bar{A}B & CB & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ C\bar{A}^{s-1}B & C\bar{A}^{s-2}B & \cdots & CB & 0 \end{bmatrix}$$

Moreover, we know that the the matrix $\mathcal{O}_s X_{i,1,N}$ is of rank smaller than s, the matrix $\mathcal{O}_s X_{i,1,N} = Y_{i,s,N} - \mathcal{T}_s U_{i,s,N} - \mathcal{S}_s S_{i,s,N}$ is of low rank. Hence nuclear norm minimization could be used to enforce this low-rank condition.

Besides enforcing sparse system matrices ($\bar{A}$,B,C,K), enforcing sparse Toeplitz matrices $\mathcal{T}_s$ and $\mathcal{S}_s$ might be sufficient. Note that a predictor in the form of a VARX model follows from the last block-row of the data equation:

$$\hat{s}(k+1|k) = \begin{bmatrix} C\bar{A}^{s-1}B & C\bar{A}^{s-2}B & \cdots & C\bar{A}B & CB \end{bmatrix} \begin{bmatrix} u(k-s-1) \\ u(k-s-2) \\ \vdots \\ u(k-1) \\ u(k) \end{bmatrix}$$
$$+ \begin{bmatrix} C\bar{A}^{s-1}K & C\bar{A}^{s-2}K & \cdots & C\bar{A}K & CK \end{bmatrix} \begin{bmatrix} s(k-s-1) \\ s(k-s-2) \\ \vdots \\ s(k-1) \\ s(k) \end{bmatrix}$$

This predictor supports the primary focus of VARX models in this thesis. Specifically, if $\bar{A}$ will be estimated as a nilpotent matrix, the model can be described by a low order VARX model without the loss of accuracy.

B-2 The algorithm

It was proposed to solve the identification problem using nuclear norm minimization on $\mathcal{O}_s X$ and fit the data equation. To ease the notation, the subscripts are dropped from the Hankel matrices. Moreover, the block-Toeplitz matrices are parametrized in their blocks, denoted by $t_{s,i}$ and $t_{u,i}$ for $\mathcal{S}_s$ and $\mathcal{T}_s$ respectively and the estimate of Y, $\tilde{Y}$, is parametrized in its vectors $\tilde{y}(k)$, and $\tilde{y}$ is a concatenation of these vectors over all time instances. To take into account that there is noise on the output, we should minimize its variance. One possible problem formulation could be

$$\min_{\tilde{y}(k),t_{u,i},t_{s,i},\mathcal{O}_s,X} \quad \|\tilde{Y} - \mathcal{T}_s(t_{u,i})U - \mathcal{S}_s(t_{s,i})S - \mathcal{O}_s X\|_F^2 + \tfrac{1}{2}(y-\tilde{y})^T(y-\tilde{y}) + \|\mathcal{O}_s X\|_\star$$

where $\|\cdot\|_\star$ denotes the nuclear norm. If we now introduce the operator

$$A(a) = \tilde{Y} - \mathcal{T}_s(t_{u,i})U + \mathcal{S}_s(t_{s,i})S$$
$$a = \mathrm{vec}(\tilde{y}, t_{u,i}, t_{s,i})$$

we get a formulation that is also used in the N2SID method of Verhaegen and Hansson (2014). Using the approximation proposed in Haeffele et al. (2014)

$$\|UV\|_\star = \arg\min_{U,V} \frac{1}{2}\left(\|U\|_F^2 + \|V\|_F^2\right)$$

we obtain a problem that is very similar to that described in Signoretto et al. (2013). Namely,

$$\begin{aligned} \min_{y(\tilde{k}),t_{u,i},t_{s,i},\mathcal{O}_s,X} \quad & \tfrac{1}{2}(\tilde{y}-y)^T(\tilde{y}-y) + \tfrac{1}{2}\left(\|\mathcal{O}_s\|_F^2 + \|X\|_F^2\right) \\ \text{s.t.} \quad & \mathcal{A}(a) = \mathcal{O}_s X \end{aligned} \tag{B-1}$$

The main difference in our problem formulation compared to (B-1) is that we are looking for a solution that is as sparse as possible. One possible way to achieve sparsity is by adding an ℓ_1-norm regularization term on $\mathcal{O}_s$, $\mathcal{S}_s(t_{s,i})$ and $\mathcal{T}_s(t_{u,i})$. For this purpose, we introduce a second operator $\mathcal{O}(o) = \mathcal{O}_s$, where o is simply the vectorization of $\mathcal{O}_s$ and a new variable α for the ℓ_1-norm regularization. Below, all previously introduced and some new notations are listed

Notation		Size
a	$= \text{vec}(\tilde{y}(k), t_{u,i}, t_{s,i})$	$Np+(s-1)mp+(s-1)p^2$
o	$= \text{vec}(\mathcal{O}_{s,i})$	nps
α	$= \text{vec}(\mathcal{O}_{s,i}, t_{u,i}, t_{s,i})$	$nps+(s-1)mp+(s-1)p^2$
$\mathcal{A}(a)$	$= \tilde{Y} - T_u(t_{u,i})U + T_s(t_{s,i})S$	$sp \times N-s+1$
$\mathcal{O}_s(o)$	$= \mathcal{O}_s$	$sp \times n$
$P_t a$	$= \text{vec}(t_{u,i}, t_{s,i})$	$(s-1)mp+(s-1)p^2$
y	$= \text{vec}(y(k))$	Np
a_y	$= \text{vec}(y(k), 0)$	$Np+(s-1)mp+(s-1)p^2$
$(a-a_y)^T P_y (a-a_y)$	$= (\tilde{y}-y)^T(\tilde{y}-y)$	1

where p, m and N are the dimensions of the output vector, input vector and number of identification time samples respectively. The final optimization problem can now be written in the following form

$$\begin{aligned} \min_{a,o,X,\alpha} \quad & \tfrac{1}{2}\left(\|\mathcal{O}(o)\|_F^2 + \|X\|_F^2\right) + \tfrac{1}{2}\lambda_y (a-a_y)^T P_y (a-a_y) + \gamma\|\alpha\|_1 \\ \text{s.t.} \quad & \mathcal{A}(a) = \mathcal{O}(o)X \\ & \alpha = \begin{bmatrix} o \\ P_t a \end{bmatrix} \end{aligned} \tag{B-2}$$

In the following, (B-2) is solved using ADMM and the algorithm is tested in simulations.

B-2-1 ADMM formulation

In Appendix A, the basics of ADMM have already been discussed. The optimization problem (B-2) is already in ADMM form and its augmented Lagrangian is

$$\begin{aligned} L(a,o,X,\alpha,Z,Z_\alpha) = {} & \frac{1}{2}\|\mathcal{O}(o)\|_F^2 + \frac{1}{2}\|X\|_F^2 + \frac{1}{2}\lambda_y (a-a_y)^T P_y (a-a_y) + \\ & \langle Z, \mathcal{A}(a) - \mathcal{O}(o)X\rangle + \frac{\rho}{2}\|\mathcal{A}(a) - \mathcal{O}(o)X\|_F^2 + \\ & \langle Z_\alpha, \alpha - \begin{bmatrix} o \\ P_t a \end{bmatrix}\rangle + \frac{\mu}{2}\|\alpha - \begin{bmatrix} o \\ P_t a \end{bmatrix}\|_F^2 + \gamma\|\alpha\|_1 \end{aligned} \tag{B-3}$$

with dual variables Z and Z_α and $\langle \cdot, \cdot \rangle$ denoting the inner product. By taking partial derivatives of (B-3), we find the 6 updating steps

$$\begin{aligned} a^{k+1} = \arg\min_a \ & \frac{1}{2}\lambda_y (a-a_y)^T P_y (a-a_y) + \frac{\rho}{2}\|\mathcal{A}(a) - \mathcal{O}(o^k)X^k + \frac{1}{\rho}Z^k\|_F^2 \\ & + \frac{\mu}{2}\|\alpha^k - \begin{bmatrix} o^k \\ P_t a \end{bmatrix} + \frac{1}{\mu}Z_\alpha^k\|_2^2 \end{aligned} \tag{B-4}$$

$$\begin{aligned} o^{k+1} = \arg\min_o \ & \frac{1}{2}\|\mathcal{O}(o)\|_F^2 + \frac{\rho}{2}\|\mathcal{A}(a^{k+1}) - \mathcal{O}(o)X^k + \frac{1}{\rho}Z^k\|_F^2 \\ & + \frac{\mu}{2}\|\alpha^k - \begin{bmatrix} o \\ P_t a^{k+1} \end{bmatrix} + \frac{1}{\mu}Z_\alpha^k\|_2^2 \end{aligned} \tag{B-5}$$

$$X^{k+1} = \arg\min_x \ \frac{1}{2}\|X\|_F^2 + \frac{\rho}{2}\|\mathcal{A}(a^{k+1}) - \mathcal{O}(o^{k+1})X + \frac{1}{\rho}Z^k\|_F^2 \tag{B-6}$$

$$\alpha^{k+1} = \arg\min_\alpha \ \gamma\|\alpha\|_1 + \frac{\mu}{2}\|\alpha - \begin{bmatrix} o^{k+1} \\ P_t a^{k+1} \end{bmatrix} + \frac{1}{\mu}Z_\alpha^k\|_2^2 \tag{B-7}$$

$$Z^{k+1} = Z^k + \rho\left(\mathcal{A}(a^{k+1}) - \mathcal{O}(o^{k+1})X^{k+1}\right) \tag{B-8}$$

$$Z_\alpha^{k+1} = Z_\alpha^k + \mu\left(\alpha^{k+1} - \begin{bmatrix} o^{k+1} \\ P_t a^{k+1} \end{bmatrix}\right) \tag{B-9}$$

There are many convergence results for ADMM that apply to the convex problem of Liu et al. (2013), but do not immediately apply for our non-convex problem. The convergence can be showed from simulations, but is not likely to be found for the algorithm described above, as was also the case for the related problem of Signoretto et al. (2013).

B-2-2 Analytical solution to the update step

For all updating steps it is possible to find an analytic solution. Here we can use that for any operator $\mathcal{G}(g)$ we can define its adjoint $\mathcal{G}^*(g)$ satisfying

$$\langle \mathcal{G}(g), \Gamma \rangle = \langle g, \mathcal{G}^*(\Gamma) \rangle \tag{B-10}$$

The derivative of this expression with respect to g is

$$\frac{\partial \langle \mathcal{G}(g), \Gamma \rangle}{\partial g} = \mathcal{G}^*(\Gamma)$$

Furthermore, since $\langle A, B \rangle = \mathrm{tr}(AB^T)$ we can use the fact that it is invariant to cyclic permutations, i.e.

$$\langle A, B \rangle = \left\langle AB^T, I \right\rangle = \left\langle B^T, A^T \right\rangle = \left\langle B^T A, I \right\rangle$$

Using these definitions we can derive the analytic expressions for the update steps. The a-update can be found by setting the partial derivative $\frac{\partial L}{\partial a}$ (i.e. the right hand side of (B-4)) to zero, i.e.

$$\left(\lambda_y P_y + \rho M_a + \mu P_t^T P_t\right) a = \lambda_y P_y a_y + \mathcal{A}^*(\rho \mathcal{O}_s(o) X^T - Z) + \begin{bmatrix} 0 & P_t^T \end{bmatrix}(\mu\alpha + Z_\alpha) \tag{B-11}$$

where

$$M_a a = \mathcal{A}^*(A(a)) \tag{B-12}$$

Besides straightforward calculation, the matrix M_a can be calculated much more efficiently by exploiting the Hankel structure and using MATLAB's fast Fourier transform routine as described in Liu et al. (2013). Also, M_a is block diagonal with equal blocks and the number of blocks equal to the number of outputs of the system. Hence only one block needs to be constructed and inverted to define the inverse of the term on the left-hand side of (B-11).

The o-update can be calculated similar to a, by setting $\frac{\partial L}{\partial o}$ to zero, i.e.

$$(\lambda_O I_p + \mu I_p + \rho M_o)\, o = \left(\rho \mathcal{O}_s^*(\mathcal{A}(a)X^T) + \mathcal{O}^*(ZX^T)\right) + \begin{bmatrix} I_p & 0 \end{bmatrix}(\mu\alpha + Z_\alpha) \tag{B-13}$$

where

$$M_o o = \mathcal{O}^*(\mathcal{O}(o)X^T X)$$

Since the operation $\mathcal{O}(o)$ is just reshaping a vector, $\mathcal{O}^*(\mathcal{O}(o)) = I$ and the matrix M_o ends up to be a large sparse matrix $M_o = XX^T \otimes I_{sp}$. This Kronecker structure can furthermore be exploited in the efficient computation of the inverse.

Likewise, the X-update follows from straightforward calculations by setting $\frac{\partial L}{\partial o}$ to zero

$$\left(\lambda_X I_n + \rho \mathcal{O}(o)^T \mathcal{O}(o)\right) X = \left(\rho \mathcal{O}(o)^T \mathcal{A}(a) + \mathcal{O}(o)^T Z\right) \tag{B-14}$$

The α-update can be found using subdifferential calculus

$$\alpha = S_{\gamma/\mu}\left(\begin{bmatrix} o \\ P_t a \end{bmatrix} - \frac{1}{\mu} Z_\alpha\right) \tag{B-15}$$

where $S_\kappa(c) = \max\left(0, 1 - \frac{\kappa}{|c|}\right) c$ is the soft thresholding operator and should be interpreted elementwise.

B-2-3 Stopping criteria and parameter selection

To test for optimality, one possibility is to follow the criteria proposed by Boyd et al. (2011). For our specific example, they would read as:

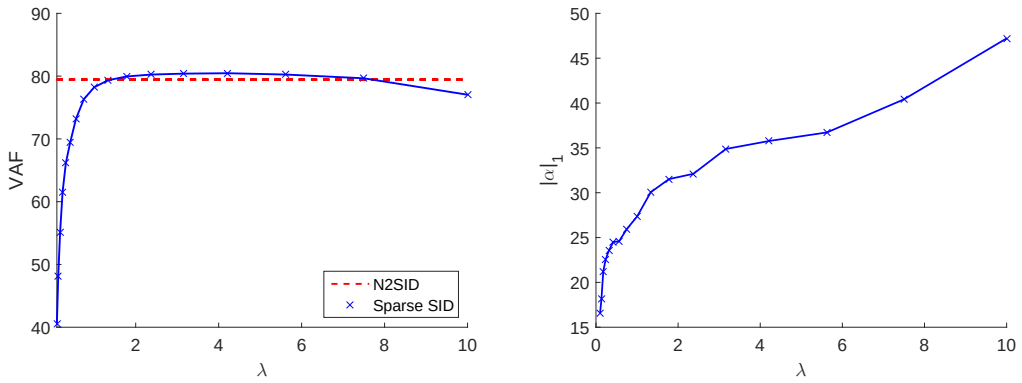
$$\begin{aligned}
r_{p1} &= \mathcal{A}(a^k) - \mathcal{O}(o^k)X^k \\
r_{p2} &= \alpha^k - \begin{bmatrix} o^k \\ P_t a^k \end{bmatrix} \\
r_{d1} &= \rho \mathcal{A}^*(\mathcal{O}(o^{k-1})X^{k-1} - \mathcal{O}(o^k)X^k) \\
r_{d2} &= \mu\left(\begin{bmatrix} o^{k-1} \\ P_t a^{k-1} \end{bmatrix} - \begin{bmatrix} o^k \\ P_t a^k \end{bmatrix}\right) \\
\epsilon_{p1} &= \sqrt{pq}\epsilon_{abs} + \epsilon_{rel} \max\left\{\|\mathcal{A}(a^k)\|_F,\ \|\mathcal{O}(o^k)X^k\|_F\right\} \\
\epsilon_{p2} &= \sqrt{n_\alpha}\epsilon_{abs} + \epsilon_{rel} \max\left\{\|\alpha^k\|_F,\ \left\|\begin{bmatrix} o^k \\ P_t a^k \end{bmatrix}\right\|_F\right\} \\
\epsilon_{d1} &= \sqrt{n_x}\epsilon_{abs} + \epsilon_{rel}\|\mathcal{A}^*(Z)\|_2 \\
\epsilon_{d2} &= \sqrt{n_\alpha}\epsilon_{abs} + \epsilon_{rel}\|Z_\alpha\|_2
\end{aligned}$$

There is however no general rule for selecting and/or updating the regularization parameters. Since a higher value for γ will result in a higher sparsity, and a higher value of λ_y in a better fit to the data, it is expected that the sparsity of the model is depending on both values. A higher value for γ and a lower value for λ_y will increase the sparsity.

B-2-4 Simulation results

The algorithm described in this section has been tested in simple simulations. The simulations are done using small-scale innovation models with randomly generated matrices A, B, C and K, such that the poles of eigenvalues of A and $A-KC$ lie within the unit circle. Furthermore the parameters $\rho = \mu = 1$ and $\gamma = 0.1$ are kept constant, while λ will be varied.

Figure B-1 show the results that are obtained. Clearly, it seems to work as well as N2SID for small dimensions. When λ decreases in comparison to γ it becomes clear that the sparsity increases while the accuracy of the model starts to decrease. The drop in accuracy for too large values of λ can be explained by overfitting.



(a) VAF of the model compared to a model obtained using N2SID.

(b) ℓ_1-norm of α

Figure B-1: An example of the identification results of sparse SID in comparison with N2SID.

B-3 Discussion

The main goal of this research was to find a scalable alternative to the standard Kalman filter in modelling an adaptive optics system. The method that is described here does find a sparse model, but the process of deriving this model is not efficient at all. The large matrices (such as M_a and M_o) that grow exponentially when the size of the system increases are either complex to derive or used in operations such as inversions. Since there is no clear solution to this problem, this method was not further investigated in this thesis. For small-scale problems it still might have some contributions. However, there are a number of problems that have to be solved first.

The algorithm can obtain either a sparse ARX or state-space model. For the first case, obtaining a sparse $\mathcal{O}_s$, $\mathcal{S}_s$ and $\mathcal{T}_s$ is sufficient, but a very complicated and inefficient approach

compared to the method of Chapter 4. For the state-space model, we need one extra step of extracting the matrices A, B, C and K as sparse as possible. Normally (e.g. in Verhaegen and Hansson, 2014), one would extract C and $A-KC$ from the matrix $\mathcal{O}_s$, then K using the matrix $\mathcal{S}_s$ and then B from $\mathcal{T}_s$. The problem is however that apart from C, the matrices will in general be dense, since they are calculated as a simple least squares problem. A new method has to be developed for this purpose that enforces sparsity in this problem, for example by using the fact that we have sparse representations of the matrices C, CA, CB, CK, etc.

Another fundamental problem is that the matrices $\mathcal{O}_s$, $\mathcal{S}_s$ and $\mathcal{T}_s$ that have been estimated so far are approximately sparse, but not exactly since the residual r_{p2} will never be exactly zero. This means that all very small values in a, o (and α) should be truncated during the iterations. It is however another parameter that would decide at what value it will be truncated. Also, truncation might lead to a small shift in the poles of the system. Turbulence models tend to have poles close to the unit circle such that truncation often results in an instability.

Bibliography

F. Assémat, R. Wilson, and E. Gendron. Method for simulating infinitely long and non stationary phase screens with optimized memory storage. *Optics express*, 14(3):988–999, 2006.

O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516, 2008.

J. M. Bardsley, S. Knepper, and J. Nagy. Structured linear algebra problems in adaptive optics imaging. *Advances in Computational Mathematics*, 35(2-4):103–117, 2011.

A. Beghi, A. Cenedese, and A. Masiero. Stochastic realization approach to the efficient simulation of phase screens. *JOSA A*, 25(2):515–525, 2008.

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *Information Theory, IEEE Transactions on*, 52(2):489–509, 2006.

R. Conan. Mean-square residual error of a wavefront after propagation through atmospheric turbulence and after correction with zernike polynomials. *JOSA A*, 25(2):526–536, 2008.

R. Dahlhaus. Graphical interaction models for multivariate time series. *Metrika*, 51(2):157–172, 2000.

R. Dahlhaus and M. Eichler. Causality and graphical models in time series analysis. *Highly Structured Stochastic Systems*, pages 115–137, 2003. ISSN 0952-9942.

C. C. de Visser, E. Brunner, and M. Verhaegen. On distributed wavefront reconstruction for large-scale adaptive optics systems. *JOSA A*, 33(5):817–831, 2016.

A. P. Dempster. Covariance selection. *Biometrics*, pages 157–175, 1972.

Q. T. Dinh, A. Kyrillidis, and V. Cevher. A proximal newton framework for composite minimization: Graph learning without cholesky decompositions and matrix inversions. *arXiv preprint arXiv:1301.1459*, 2013.

D. L. Donoho. De-noising by soft-thresholding. *Information Theory, IEEE Transactions on*, 41(3):613–627, 1995.

D. L. Donoho. For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution. *Communications on pure and applied mathematics*, 59(6):797–829, 2006.

J. Duchi, S. Gould, and D. Koller. Projected subgradient methods for learning sparse gaussians. In *Proceedings of the Conference on Uncertainty in AI*, 2008.

B. Ellerbroek and T. Rhoadarmer. Adaptive wavefront control algorithms for closed loop adaptive optics. *Mathematical and Computer modelling*, 33(1):145–158, 2001.

D. W. Fausett, C. T. Fulton, and H. Hashish. Improved parallel qr method for large least squares problems involving kronecker products. *Journal of computational and applied mathematics*, 78(1):63–78, 1997.

D. L. Fried. Statistics of a geometric representation of wavefront distortion. *JOSA*, 55(11): 1427–1435, 1965.

D. L. Fried. Least-square fitting a wave-front distortion estimate to an array of phase-difference measurements. *JOSA*, 67(3):370–375, 1977.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

L. Gilles, C. R. Vogel, and B. L. Ellerbroek. Multigrid preconditioned conjugate-gradient method for large-scale wave-front reconstruction. *JOSA A*, 19(9):1817–1822, 2002.

R. Gilmozzi and J. Spyromilio. The european extremely large telescope (e-elt). *The Messenger*, 127(11):3, 2007.

C. W. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

C. W. Granger. Testing for causality: a personal viewpoint. *Journal of Economic Dynamics and control*, 2:329–352, 1980.

D. P. Greenwood. Bandwidth specification for adaptive optics systems. *JOSA*, 67(3):390–393, 1977.

B. Haeffele, E. Young, and R. Vidal. Structured low-rank matrix factorization: Optimality, algorithm, and applications to image processing. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 2007–2015, 2014.

K. Hinnen, M. Verhaegen, and N. Doelman. A data-driven-optimal control approach for adaptive optics. *Control Systems Technology, IEEE Transactions on*, 16(3):381–395, 2008.

K. J. G. Hinnen. *Data-driven optimal control for adaptive optics*. TU Delft, Delft University of Technology, 2007.

C.-J. Hsieh, I. S. Dhillon, P. K. Ravikumar, and M. A. Sustik. Sparse inverse covariance matrix estimation using quadratic approximation. In *Advances in Neural Information Processing Systems*, pages 2330–2338, 2011.

C.-J. Hsieh, M. A. Sustik, I. S. Dhillon, P. K. Ravikumar, and R. Poldrack. Big & quic: Sparse inverse covariance estimation for a million variables. In *Advances in Neural Information Processing Systems*, pages 3165–3173, 2013.

R. H. Hudgin. Wave-front reconstruction for compensated imaging. *JOSA*, 67(3):375–378, 1977.

A. N. Kolmogorov. The local structure of turbulence in incompressible viscous fluid for very large reynold's numbers. *C.R. (Dokl.) Acad. Sci. URSS*, 30:301–305, 1941a.

A. N. Kolmogorov. Dissipation of energy in the locally isotropic turbulence. *C.R. (Dokl.) Acad. Sci. URSS*, 32:15–18, 1941b.

C. Kulcsár, H.-F. Raynaud, C. Petit, J.-M. Conan, and P. V. de Lesegno. Optimal control, observers and integrators in adaptive optics. *Optics express*, 14(17):7464–7476, 2006.

S. L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

L. Li and K.-C. Toh. An inexact interior point method for l1-regularized sparse covariance selection. *Mathematical Programming Computation*, 2(3-4):291–315, 2010.

A. Lindquist and G. Picci. *Linear stochastic systems: A geometric approach to modeling, estimation and identification*, volume 1. Springer, 2015.

Z. Liu, A. Hansson, and L. Vandenberghe. Nuclear norm system identification with missing inputs and outputs. *Systems & Control Letters*, 62(8):605–612, 2013.

D. P. Looze. Minimum variance control structure for adaptive optics systems. *JOSA A*, 23 (3):603–612, 2006.

Z. Lu. Adaptive first-order methods for general sparse inverse covariance selection. *SIAM Journal on Matrix Analysis and Applications*, 31(4):2000–2016, 2010.

A. Maréchal. Étude des effets combinés de la diffraction et des aberrations géométriques sur lâĂŹimage dâĂŹun point lumineux. *Rev. Opt*, 26:257–277, 1947.

P. Massioni, H.-F. Raynaud, C. Kulcsar, and J.-M. Conan. An approximation of the riccati equation in large-scale systems with application to adaptive optics. *Control Systems Technology, IEEE Transactions on*, 23(2):479–487, 2015.

R. Mazumder and T. Hastie. The graphical lasso: New insights and alternatives. *Electronic journal of statistics*, 6:2125, 2012.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, pages 1436–1462, 2006.

C. C. Paige and M. A. Saunders. Lsqr: An algorithm for sparse linear equations and sparse least squares. *ACM transactions on mathematical software*, 8(1):43–71, 1982.

N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3): 123–231, 2013.

A. Reeves. *Laser Guide Star Only Adaptive Optics: The Development of Tools and Algorithms for the Determination of Laser Guide Star Tip-Tilt.* PhD thesis, Durham University, 2015.

F. Roddier. The effects of atmospheric turbulence in optical astronomy. *In: Progress in optics. Volume 19. Amsterdam, North-Holland Publishing Co., 1981, p. 281-376.*, 19:281–376, 1981.

F. Roddier. *Adaptive optics in astronomy.* Cambridge university press, 1999.

M. Rosensteiner. Wavefront reconstruction for extremely large telescopes via cure with domain decomposition. *JOSA A*, 29(11):2328–2336, 2012.

K. Scheinberg, S. Ma, and D. Goldfarb. Sparse inverse covariance selection via alternating linearization methods. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23, pages 2101–2109. Curran Associates, Inc., 2010.

J. D. Schmidt. Numerical simulation of optical wave propagation with examples in matlab. SPIE Washington, 2010.

M. Signoretto, V. Cevher, and J. A. Suykens. An svd-free approach to a class of structured low rank matrix optimization problems with application to system identification. In *IEEE Conference on Decision and Control*, number EPFL-CONF-184990, 2013.

J. Songsiri and L. Vandenberghe. Topology selection in graphical models of autoregressive processes. *The Journal of Machine Learning Research*, 11:2671–2705, 2010.

J. Songsiri, J. Dahl, and L. Vandenberghe. Graphical models of autoregressive processes. *Convex Optimization in Signal Processing and Communications*, pages 89–116, 2009.

W. H. Southwell. Wave-front estimation from wave-front slope measurements. *JOSA*, 70(8): 998–1006, 1980.

V. Sriram and D. Kearney. An ultra fast kolmogorov phase screen generator suitable for parallel implementation. *Optics express*, 15(21):13709–13714, 2007.

V. Tatarski, R. Silverman, and N. Chako. Wave Propagation in a Turbulent Medium. *Physics Today*, 14:46, 1961.

G. I. Taylor. The spectrum of turbulence. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 164, pages 476–490. The Royal Society, 1938.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

E. Treister and J. S. Turek. A block-coordinate descent approach for large-scale sparse inverse covariance estimation. In *Advances in Neural Information Processing Systems*, pages 927–935, 2014.

M. Verhaegen and A. Hansson. Nuclear norm subspace identification (n2sid) for short data batches. *arXiv preprint arXiv:1401.4273*, 2014.

M. Verhaegen and V. Verdult. *Filtering and system identification: a least squares approach*. Cambridge university press, 2007.

M. Verhaegen, G. Vdovin, and O. Soloviev. Lecture notes on control for high resolution imaging. Delft Center for Systems and Control, Delft University of Technology, April 2015.

E. P. Wallner. Optimal wave-front correction using slope measurements. *JOSA*, 73(12): 1771–1776, 1983.

C. Wang, D. Sun, and K.-C. Toh. Solving log-determinant optimization problems by a newton-cg primal proximal point algorithm. *SIAM Journal on Optimization*, 20(6):2994–3013, 2010.

H. Wang, A. Banerjee, C.-J. Hsieh, P. K. Ravikumar, and I. S. Dhillon. Large scale distributed sparse precision estimation. In *Advances in Neural Information Processing Systems*, pages 584–592, 2013.

A. Wiesel and A. O. Hero III. Distributed covariance estimation in gaussian graphical models. *Signal Processing, IEEE Transactions on*, 60(1):211–220, 2012.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

X. Yuan. Alternating direction method for covariance selection models. *Journal of Scientific Computing*, 51(2):261–273, 2012.