OPTIMISING ALIGNMENT OF A MULTI-ELEMENT TELESCOPE

Morgan M. KAMGA

A thesis submitted to the Faculty of Science
in fulfillment of the requirements of the degree of
Doctor of Philosophy

School of Computational and Applied Mathematics
University of the Witwatersrand

September 20, 2012

Declaration

I declare that this thesis is my own, unaided work. It is being submitted for the Degree of Doctor of Philosophy to the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination to any other University.

(Signature)

(Date)

Abstract

In this thesis, we analyse reasons for poor image quality on the Southern African Large Telescope (SALT) and we analyse control methods of the segmented primary mirror. Errors in the control algorithm of SALT (circa 2007) are discovered. More powerful numerical procedures are developed and in particular, we show that singular value decomposition method is preferred over normal equations method as used on SALT. In addition, this method does not require physical constraints to some mirror parameters. Sufficiently accurate numerical procedures impose constraints on the precision of segment actuator displacements and edge sensors. We analyse the data filtering method on SALT and find that it is inadequate for control. We give a filtering method that achieves improved control. Finally, we give a new method (gradient flow) that gives acceptable control from arbitrary, imprecise initial alignment.

Acknowledgements

Without the help, support and contribution of many people, this work would not have been possible.

I would first of all like to thank the *CAM Staff* for offering me a place in the School of Computational and Applied Mathematics.

The School of Computational and Applied Mathematics for providing financial support to start and to finish this work.

SALT staff for offering me the Stobie-Salt Scholarship, and allowing me to experiment on the telescope.

I am also grateful to *Prof. Bao-Zhu Guo* for accepting to supervise my work, for leading me through some areas of mathematics, and for financial support amongst many things.

I would also like to thank *Mr. Colin Myburgh* for his availability, for the majority of scientific discussions and for co-supervising the current work.

My gratitude also goes to *Dr. John Menzies* as a co-supervisor, *Mr. Hitesh Gajjar*, *Mr. Deon Bester* and the SALT technical staff for assisting me with my work and experiments on SALT.

A particular thanks to *Prof. David Sherwell* for bringing this topic which took me in new areas of mathematics, and for his support, advice and encouragements.

I don't forget to thank my friends *Naval Andrianjafinandrasana*, *Charles Lebon Mberi Kimpolo*, *Virginie Konlack*, *Olivier Menoukeu* and *Hugues Tchantcho* for their support, collaboration and availability for discussions.

I am very grateful to my family for their unconditional support and gracious assistance despite the distance between us.

I finally express my gratitude to all the people who in one way or another contributed to the achievement of this work.

Contents

List of Figures

List of Tables

Chapter 1

Review of Multi-element Reflecting Telescopes

This thesis concerns computational and numerical aspects of SALT, the Southern African Large Telescope. SALT is a reflecting telescope with a 9.4 metres segmented primary mirror. The alignment of segments is maintained by automatic control. Image quality, on commission in 2005, was unsatisfactory, owing to flaws in the optical path (spherical aberration corrector) and, it is believed, for errors of measurements owing to the effect of humidity on capacitive edge sensors. This thesis, in Chapter 2, objectively decides on significant environmental variables that corrupt segment alignment. In Chapters 3 and 4, we analyse the control of mirror segments. From this, it is clear that we approach the problem of SALT image quality as applied mathematicians, using mathematical statistics and control theory. We do not concern ourselves with physical optics (individual mirror segments meet specifications, as does, now, the spherical aberration corrector) or instrumentation.

Telescopes with a primary mirror built from many segments are multi-element telescopes. Note that mirrors of telescopes can be solid, segmented, meniscus, honeycomb, liquid. Examples of telescopes are given in Table 1.1, with basic information.

Note in Table 1.1 that:

- Aperture is measured in metres.
- The term *Date* refers to the year when the telescope was commissioned, or is expected to be operational.
- IRTF is the Infrared Telescope Facility, located at the National Aeronautics and Space Administration (NASA), Mauna Kea, Hawaii.

Table 1.1: A short list of telescopes

Name	**Date**	**Type**	**Aperture (m)**
IRTF	1979	Solid	3
INT	1984	Solid	2.5
KECK	1992/96	Segmented	10
ARC	1994	Honeycomb	3.48
HET	1997	Segmented	9.5
SALT	2005	Segmented	9.5
LSST	2013	Honeycomb	8.4
E-ELT	2018	Segmented	42
TMT	2018	Segmented	30
GMT	2019	Honeycomb	21.4

- INT is the Isaac Newton Telescope, located at the Observatory Roque de los Muchachos, La Palma, Canary Islands.
- KECK is composed of two telescopes, the first operational since 1992 and the second since 1996; the two telescopes are located at the W. M. KECK Observatory, Mauna Kea, Hawaii.
- ARC is the Astrophysical Research Consortium, located at the Apache Point Observatory, Sacramento Peak, New Mexico.
- HET is the Hobby-Eberly Telescope, located at the Mc-Donald Observatory, Mt. Fow Ikes, Texas.
- SALT is the Southern African Large Telescope, located at the South African Astronomical Observatory, Sutherland, South Africa.
- LSST is the Large Synoptic Survey Telescope, under construction, expected to be operational in 2013 and located at the Cerro Tololo Inter-American Observatory, Cerro Pachon, Chile.
- E-ELT is the European Extremely Large Telescope, under construction, expected to be operational in 2018 and located in Cerro Armazones, Chile.
- TMT is the Thirty Meter Telescope, previously known as CELT (California Extremely Large Telescope), under construction, expected to

be operational in 2018 and located in Mauna Kea, Hawaii.

- GMT is the Giant Magellan Telescope, under construction, expected to be operational in 2019 and located at Las Campanas Observatory, Cerro Las Campanas, Chile.

Until 2018, SALT is, with KECK and HET, among the largest multi-element telescopes in operation in the world. SALT is important because it is the largest telescope in the Southern Hemisphere. However, SALT is not yet in good working order. Of the telescopes listed in Table 1.1, SALT is the reflector on which our studies and experiments are conducted. We are only interested in telescopes with a segmented primary mirror, that is, multi-element telescopes.

1.1 Background

We give basic specifications of a few examples of multi-element telescopes.

1.1.1 SALT

SALT is the Southern African Large Telescope and is an example of a multi-element telescope, that is, a telescope with a segmented primary mirror. It is located in Sutherland near Cape Town in South Africa. It is based on the design of the Hobby-Eberly Telescope (HET is located in Texas).

SALT primary mirror has a spherical shape and each segment has a spherical top surface. The radius of curvature is GRoC = 26.165 metres. GRoC stands for Global Radius of Curvature, that is, the radius of the best-fit spherical surface to the mirror surface after a possible change in alignment. It (SALT primary mirror) is composed of 91 hexagonal (regular hexagon) interchangeable numbered mirror segments each with an inscribed diameter $H = 1$ metre, three mounting points on an equilateral triangle at a distance $R = 0.313125$ metre from the center of the segment, sensors with one emitting plate and one receiving plate on each edge between any two adjacent segments, at a distance $d = 13.5$cm from the nearest corners. This gives 273 mounting points and 480 sensors. The primary mirror viewed from the top is illustrated in Figure 1.1, and a few segments with mounting points and sensor positions in Figure 1.2. Note that SALT uses capacitive sensors and these are sensitive to humidity. SALT primary mirror rotates in

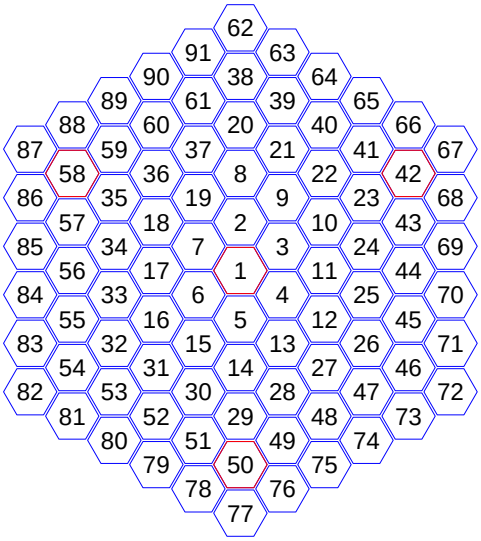

Figure 1.1: SALT primary mirror viewed from the top

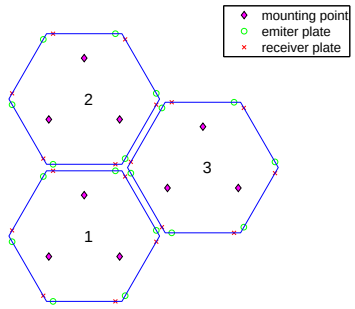

Figure 1.2: SALT segments with sensors and mounting points

azimuth only [42]. SALT was commissioned in 2005 and by 2007 it was clear that image quality was not satisfactory. The effect of humidity on capacitive edge sensors were considered important and presently (2012), SALT is tendering for inductive sensors. Errors in the spherical aberration corrector have now been corrected. In this thesis, we re-examine available data from 2007 in order to give an objective assessment of causes of poor image quality. Time, truss temperature and humidity have been found significant from our study. Changes to SALT's control algorithm were indicated. These were implemented in 2011 and tested. SALT operates with 91 segments in place. Alignment is made at nightfall in order for the control to take place through the night.

1.1.2 HET

As mentioned before (Section 1.1.1), HET (Hobby-Eberly Telescope) is the source of inspiration for the design of SALT. HET is located at the McDonald Observatory in Texas. Its construction started in 1994 and ended in 1996. It has a primary mirror composed of 91 hexagonal segments with a 1 metre inscribed diameter, a thickness of 52mm and about 115kg weight each, a total area of 78m^2, an aperture of 9.2 metres. The telescope ro-

tates in azimuth to access 85% of the sky. The primary mirror is spherically shaped with a radius of curvature (RoC) of 26 metres. All the specifications mentioned above are similar to those of SALT primary mirror. Hence the layout of HET primary mirror is the same as that of SALT primary mirror. There are (on HET) three spectrographs of low, medium and high resolution. The low resolution spectrograph (LRS) is at prime focus on the tracker (13 metres above the primary mirror). Medium resolution spectrograph (MRS) and high resolution spectrograph (HRS) are beneath the telescope in a climate controlled basement, and fed by fiber optic cable. The spectrographs are used as optical corrector facilities. The primary mirror weights about 13 tons and the telescope weights about 80 tons.

The control is performed using Singular Value Decomposition (SVD) with no constraints, which is different from the method used on SALT (details are given below). GRoC corrections needed due to thermal expansion of the truss are estimated based on gap measurements at the SAMS sensors, and the corrections are added open loop according to the solutions of the control equations. In contrast to SALT, HET uses inductive sensors. GRoC is similarly used on SALT.

1.1.3 KECK

There are two KECK telescopes and both have the same design. They are located in Hawaii. The first one was completed in 1992 and the second one in 1996. Each of the telescopes has a primary mirror with a 10 metres diameter, made up of 36 hexagonal segments. The shape of the primary mirror is a hyperboloid of revolution. The 36 segments of the primary mirror are disposed over three rings and no central segment. Each segment has a 1.8 metre inscribed diameter. The primary mirror has 168 sensors and 108 motorised adjusting devices (actuators). KECK uses capacitive sensors. A simple illustration of a KECK primary mirror is given in Figure 1.3. Sensors measure relative heights between segments. Actuator adjustments are done twice a second, which is different from SALT where adjustments are done once every four minutes. The two telescopes (the KECK telescopes) combined together are used for *interferometry*. Other specifications follow: Focal length is 17.5 metres, segment weight is about 400kg, segment thickness is about 75mm, light collection area is 76m^2, total weight of the primary mirror is about 16 tons, the total moving weight of the telescope is about 700 tons. Note

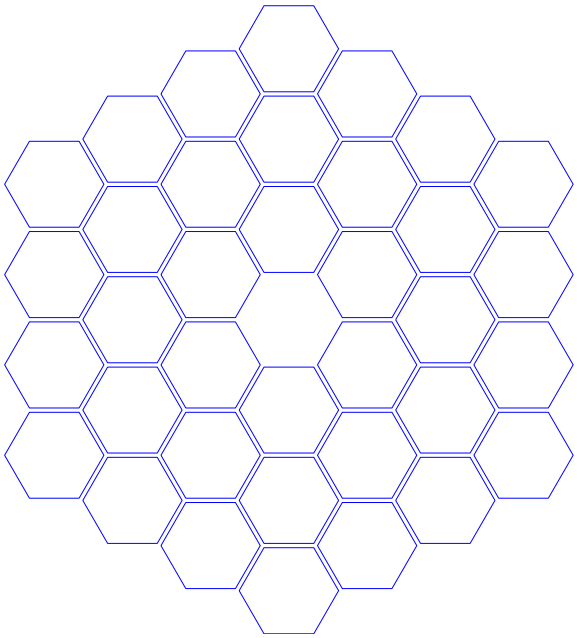

Figure 1.3: Layout of one KECK primary mirror

that the air on the site of the KECK telescopes is almost always clear, dry and not turbulent. The KECK primary mirror has a concave hyperbolic curvature. Each of the KECK telescopes also has a secondary mirror with convex hyperbolic curvature, and a flat tertiary mirror.

1.1.4 TMT (Previously known as CELT)

The Thirty Meter Telescope (TMT), previously known as the California Extremely Large Telescope (CELT) is still under design and is expected to be completed in 2018. It is supposed to be located in Hawaii, on the same site as the KECK telescopes, and is inspired by the success of the KECK telescopes. This success is confirmed, for example, by the collaboration between KECK and NASA, that resulted in generating a substantial number of scientific papers. In the initial project as CELT, the primary mirror was designed to have a 30 metres diameter, a shape of a hyperboloid of revolution, 1080 hexagonal segments with a circumscribed radius of 0.5 metre, where the out-of-plane degrees of freedom resulting from segment displacements will be actively controlled by 3240 actuators receiving feedback from 6204 edge sensors. The radius of curvature is 90 metres, the segment thickness is 0.045

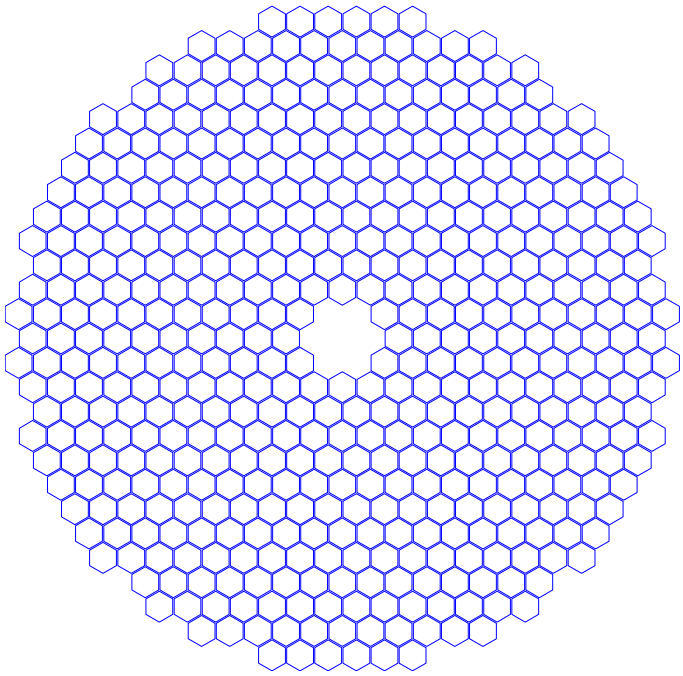

Figure 1.4: Layout of the TMT primary mirror

metre, the focal length is 45 metres and the F-ratio is 1.5. The CELT project has been revised to TMT. The primary mirror of TMT (view from the top is illustrated in Figure 1.4) is in the shape of a hyperboloid of revolution, has a 30 metres diameter and is composed of 492 hexagonal segments with a 1.4 metre inscribed diameter each, 2772 sensors, 1476 actuators. TMT also has a 3 metres diameter secondary mirror and a flat rectangular (3.6 metres by 2.5 metres) tertiary mirror. The moving mass of TMT is almost 2000 tons. Decision has not been made yet about the type of edge sensors to be used on TMT, but they will likely be capacitive sensors [29]. TMT control system is similar to that of KECK and is described in [6].

1.1.5 E-ELT

The European Extremely Large Telescope (E-ELT) is still under design and is expected to be completed in 2018. It is supposed to be located in Cerro Armazones, Chile. Its primary mirror, illustrated in Figure 1.5, has 984 hexagonal segments with circumscribed diameter of 1.4 metre, three whiffle trees and three actuators per segment, which gives 2952 actuators. Edge

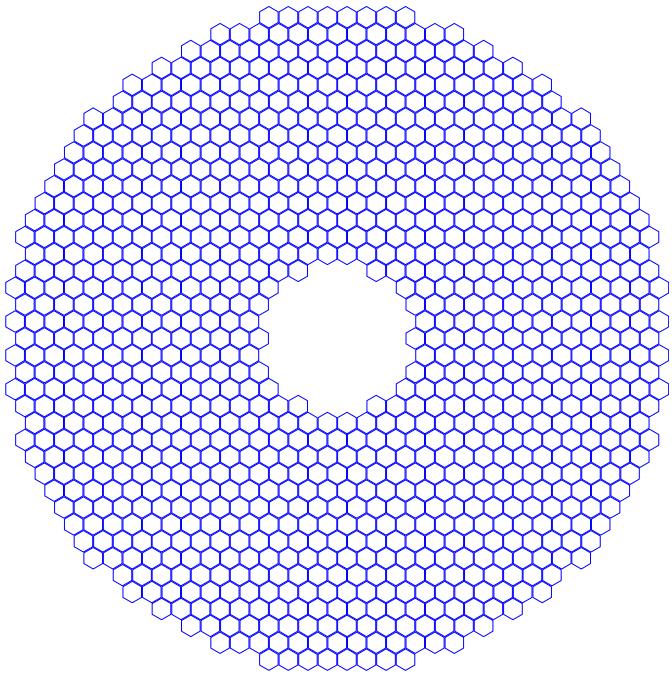

Figure 1.5: Layout of the E-ELT primary mirror

sensors are used to measure the relative heights between neighboring segments. Two edge sensors are used for each edge between any two neighboring segments. There will be approximately 6000 edge sensors. E-ELT intends to use inductive sensors. The control system is designed and explored in [13].

One common thing about segmented mirrors is that the control is performed on actuators using information about relative heights as measured by edge sensors.

1.2 Overview of Optimal Control Problems

The following methods are well known. This thesis focuses on determining the cause of image deterioration on SALT and thus requires us to critique the existing control. For later use, we outline some relevant background results in control theory.

1.2.1 Continuous time Formulation

We consider problems formulated as follows:

$$\begin{aligned} &\min_{u} J = h\left(z(T)\right) + \int_0^T f_0(t, z(t), u(t))dt \\ &\text{subject to } \begin{cases} \dot{z}(t) = f\left(t, z(t), u(t)\right), \quad z(0) = z_0 \\ s(t) = g\left(t, z(t), u(t)\right) \end{cases} \end{aligned} \tag{1.1}$$

The main goal is to find an *optimal control* u^* that minimises J where

- z is the state variable and at each time t, $z(t) \in \mathbb{R}^n$ where n is a positive integer. In this thesis, $z(t)$ will refer to the vector of actuator positions at time t.
- u is the control variable and at each time t, $u(t) \in \mathbb{R}^p$ where p is a positive integer. In this thesis, $u(t)$ will refer to the vector of corrections to be performed on actuator positions at time t.
- s is the output variable and at each time t, $s(t) \in \mathbb{R}^m$ where m is a positive integer. In this thesis, $s(t)$ will refer to the vector of relative heights at time t.
- $J : \mathbb{R}^{n+p+1} \longrightarrow \mathbb{R}$ is the *objective function* ($J = J\left(t, z(t), u(t)\right)$). In this thesis, J will refer to the overall relative heights (combined with the effort to perform the control if the integrand contains a term of the form $u^T(t)Ru(t)$ where R is a symmetric positive definite matrix).
- $h : \mathbb{R}^n \longrightarrow \mathbb{R}$, $g : \mathbb{R}^{n+p+1} \longrightarrow \mathbb{R}$, $f : \mathbb{R}^{n+p+1} \longrightarrow \mathbb{R}$ and $f_0 : \mathbb{R}^{n+p+1} \longrightarrow \mathbb{R}$ are functions with *nice* properties, that is, h, g and f are continuous and f_0 is at least piecewise continuous.
- T is the final time and can be a positive real number or infinite.

Finding an optimal control u^* that minimises the objective function J in Problem (1.1) is possible only when the system in Problem (1.1) is *controllable*, that is, when there exists a control u that can bring the system from the initial state z_0 to any given final state z_f in a finite time T. The optimal control u^* can therefore be given in the *state feedback form* using well known methods such as the Pontryagin Maximum Principle or Dynamic Programming [3, 22, 40]. Another option, recommended in practical problems, is to determine the optimal control in the *output feedback form*. This

is possible only when the system (in Problem (1.1)) is *output controllable*, that is, when there exists a control u that can bring the output of the system from the initial value $s_0 = s(0)$ to any given final value s_f at time T. The optimal control u^* can therefore be given in the output feedback form using techniques that involve problem transcriptions and numerical approaches.

1.2.2 Discrete time Formulation

In computational control, when a problem is given in continuous form, we must discretise in order to be able to implement it using a computer. For discrete time problems, we consider formulations given as follows:

$$\begin{aligned} &\min_u J = h_N\left(z_N\right) + \sum_{k=0}^{N-1} h_k(z_k, u_k, w_k) \\ &\text{subject to } \begin{cases} z_{k+1} = f_k\left(z_k, u_k, w_k\right), \quad z_0 \text{ given} \\ s_k = g_k\left(z_k, u_k, w_k\right) \end{cases} \end{aligned} \tag{1.2}$$

The main goal is to find an *optimal control* u^* that minimises J where

- k is a step in the process.
- z_k is the value of the state variable at step k, and $z_k \in \mathbb{R}^n$.
- u_k is the value of the control variable at step k, and $u_k \in \mathbb{R}^p$.
- s_k is the value of the output variable at step k, and $s_k \in \mathbb{R}^m$.
- w_k is the value of a random disturbance at step k, and $w_k \in \mathbb{R}^n$.
- J is the *objective function* and $J: \mathbb{R}^{n+p+1} \longrightarrow \mathbb{R}$.
- N is the number of steps, that is, the number of times the control is performed, and is a positive integer but can be infinite.
- h_N, h_k, g_k, and f_k are functions with *nice* properties, the same as in the continuous case, that is, $h_N : \mathbb{R}^n \longrightarrow \mathbb{R}$, $g_k : \mathbb{R}^{2n+p} \longrightarrow \mathbb{R}$, $f_k : \mathbb{R}^{2n+p} \longrightarrow \mathbb{R}$ are continuous and $h_k : \mathbb{R}^{2n+p} \longrightarrow \mathbb{R}$ is at least piecewise continuous. Note that if disturbance is not considered, then the functions f_k, g_k and h_k are defined from $\mathbb{R}^{n+p}$ to $\mathbb{R}$, and not from $\mathbb{R}^{2n+p}$ to $\mathbb{R}$.

The meaning of the variables in this section is as in the previous section on continuous time formulation, except variables being evaluated at step k instead of time t.

Finding an optimal control u^* that minimises the objective function J in Problem (1.2) is possible only when the system in Problem (1.2) is *controllable*, that is, when there exists a control u (or a discrete set of controls $(u_k)_{0\leq k\leq N-1}$ also known as *policy*) that can bring the system from the initial state z_0 to any given final state z_f in a finite number of steps N. The optimal control u^* can be given in the *state feedback form* using for example Dynamic Programming [3, 40]. Another option, recommended in practical problems, is to determine the optimal control in the *output feedback form*. This is possible only when the system in Problem (1.2) is *output controllable*, that is, when there exists a control u (or, as above in this paragraph, a *policy* $(u_k)_{0\leq k\leq N-1}$) that can bring the output of the system from the initial value $s(0)$ to any given final value s_f in a finite number of steps N. The optimal control u^* can eventually be given in the output feedback form using techniques that involve problem transcriptions and numerical approaches [5].

1.2.3 From Continuous to Discrete Time and Vice Versa

This will be explained on the particular case of a linear system (that is, a system governed by an equation of type (1.3) where x is the state variable and u is the control variable), as that will be the case of interest in our work. This is explored in [32] pages 703-704, but we recall the description. From the state equation of a system, if we need to use a computer to determine the state $x(t)$, we have to take a continuous-time state equation and convert it into a discrete-time state equation. In the lines to follow, we describe this procedure. The assumption is that the input vector $u(t)$ changes exclusively at equally spaced sampling instants. Here the discrete-time state equation which yields the exact values at $t = kT$, $k = 0, 1, 2, \ldots$ is derived. Note that T is the time interval between two consecutive steps (steps k and $k + 1$).

Consider the continuous-time state equation

$$\dot{x} = Ax + Bu \tag{1.3}$$

where at each time t, $x \in \mathbb{R}^n$, $u \in \mathbb{R}^p$, A is an $n \times n$ matrix and B is an

$n \times p$ matrix, all with real elements. This is a special case of

$$\dot{x}(t) = f\left(t, x(t), u(t)\right)$$

where f is actually linear in $x(t)$ and $u(t)$ at all times t.

In the following, in order to clarify the analysis, we use the notation kT and $(k+1)T$ instead of k and $k+1$. The discrete-time representation of Equation (1.3) will take the form

$$x\left((k+1)T\right) = G(T)x(kT) + H(T)u(kT) \tag{1.4}$$

Note that the matrices G and H depend on the sampling period T.

In order to determine $G(T)$ and $H(T)$, we use the solution of Equation (1.3), that is,

$$x(t) = e^{At}x(0) + e^{At}\int_0^t e^{-A\tau}Bu(\tau)d\tau \tag{1.5}$$

We assume that all the components of $u(t)$ are constant over the interval between any two consecutive sampling instants, or $u(t) = u(kT)$ for the k^{th} sampling period, that is, for $t \in [kT, (k+1)T)$. Since

$$x\left((k+1)T\right) = e^{A(k+1)T}x(0) + e^{A(k+1)T}\int_0^{(k+1)T} e^{-A\tau}Bu(\tau)d\tau \tag{1.6}$$

and

$$x\left(kT\right) = e^{AkT}x(0) + e^{AkT}\int_0^{kT} e^{-A\tau}Bu(\tau)d\tau \tag{1.7}$$

it follows that multiplication of Equation (1.7) by e^{AT} followed by subtraction of the obtained result from Equation (1.6) leads to

$$\begin{aligned} x\left((k+1)T\right) &= e^{AT}x\left(kT\right) + e^{A(k+1)T}\int_{kT}^{(k+1)T} e^{-A\tau}Bu(\tau)d\tau \\ &= e^{AT}x\left(kT\right) + e^{AT}\int_0^{T} e^{-At}Bu(kT)dt \\ &= e^{AT}x\left(kT\right) + \int_0^{T} e^{A\lambda}Bu(kT)d\lambda \end{aligned} \tag{1.8}$$

where $\lambda = T - t$. If we define

$$G(T) = e^{AT} \tag{1.9}$$

$$H(T) = \left(\int_0^T e^{At}dt\right)B \tag{1.10}$$

then Equation (1.8) becomes

$$x\left((k+1)T\right) = G(T)x(kT) + H(T)u(kT) \tag{1.11}$$

which is indeed Equation (1.4). Thus Equations (1.9) and (1.10) give the desired matrices $G(T)$ and $H(T)$.

Remark 1.1. The following apply:

1. From the above study, it is possible to move back from discrete-time dynamics (Equation (1.4)) to continuous-time dynamics (Equation (1.3)) and determine A and B in terms of G and H, knowing the sampling period T. More precisely, we obtain:

$$A = \frac{1}{T}\log(G) \tag{1.12}$$

$$B = \left(\int_0^T e^{At}dt\right)^{-1} H \tag{1.13}$$

   where log is the matrix logarithm.

2. The method described above in an example also applies on any state equation (linear or nonlinear ODE) that can be solved analytically and has to undergo a discretisation process or vice versa.

From Equation (1.4), it can be established that

$$x_n = G^n x_0 + \sum_{i=0}^{n-1} G^{n-1-i} H u_i \qquad \forall n \in \mathbb{N},\ n \neq 0 \tag{1.14}$$

where $x_n = x(nT)$, $u_n = u(nT)$, $G = G(T)$, $H = H(T)$ and $\mathbb{N}$ is the set of natural numbers (non-negative integers).

1.3 Statement of the Problem for SALT

From the approach described in the SAMS (Segment Alignment Measurement System) control algorithm, the main goal is to keep the relative heights as close to zero as possible. These relative heights are given from sensors on the edges of the primary mirror. Indeed, when the segments are in their ideal position, that is, they approximate a single large spherical mirror, all together, the relative heights are all zero. When the segments move, the

relative heights are given for each sensor as the distance between the emitting plate and the corresponding receiving plate in the direction towards the center of curvature, considered as the z direction in the three dimensional (x, y, z) space. The control is performed on the actuator displacements. The actuator displacements are unknown but can be estimated from the relative heights which are obtained from the sensors. There is a linear relationship between the relative heights S and the actuator displacements Z and it is given by $S = AZ$ where A is the *actuator-to-heights* matrix (known in the language of segmented mirrors as the *interaction matrix*) and is obtained from the geometry of the primary mirror. The optimal control problem for SALT can be formulated in discrete time as follows:

$$\begin{aligned} &\min_u J = \|s_N\|^2 + \sum_{k=0}^{N-1} \|s_k\|^2 \\ &\text{subject to } \begin{cases} z_{k+1} = z_k + u_k, \quad z_0 \text{ given} \\ s_k = Az_k \end{cases} \end{aligned} \tag{1.15}$$

and can be translated in continuous time as follows:

$$\begin{aligned} &\min_u J = \|s(T)\|^2 + \int_0^T \|s(t)\|^2 dt \\ &\text{subject to } \begin{cases} \dot{z}(t) = Mz(t) + Nu(t), \quad z(0) = z_0 \\ s(t) = Az(t) \end{cases} \end{aligned} \tag{1.16}$$

where M and N are 273×273 matrices that can be determined using the conversion from discrete to continuous time as described above (See item 1 of Remark 1.1), and A (the actuator-to-heights matrix) is a 480×273 matrix known from the geometry of SALT primary mirror. It is to be noted that here, $\|\cdot\|$ is the Euclidean norm (or the $\mathcal{L}^2$ norm). Note that the random disturbance is not considered. It can be established from the above formulation (in discrete and continuous time, given by Equations (1.15) and (1.16)) that the SALT control problem is an output feedback control problem.

With regard to SALT image quality, numerical techniques, rather than correctness of formulation, will be of importance.

1.4 Conclusion

SALT is not yet in acceptable configuration as the images are distorted. Capacitive edge sensors are suspected to be sensitive to humidity, and as we will show later in this thesis, the SALT mirror control algorithm gives rise for concern. This feature is a mathematical problem of robust formulation of the control algorithm and, as applied mathematicians, this will be our main interest.

Table 1.2: Dimension of the control vector

Number of Rings	1	2	3	4	5	6	7
Number of Segments	7	19	37	61	91	127	169
Size of the Control Vector	21	57	111	183	273	381	507

Table 1.2 shows the rise of dimension of the control vector. There is a linear relationship between the size of the control vector and the number of segments (since each segment has three actuators) and also a quadratic relationship between the size of the control vector and the number of rings. These sizes of vectors and matrices involved in the computation constitute a major source of computational errors. It is also of concern that the single real number J (value of the objective function which in this case is the overall measure of relative heights) might not be adequate for the control of segmented mirrors. It is hoped that this work will help in the control of next large telescopes with segmented primary mirror, such as TMT and E-ELT.

Chapter 2

Statistical Analysis of SALT Historical Data

Optimising alignment of a multi-element telescope involves theoretical and experimental work. It is essential to know what the problem is, in order to attempt providing a solution. Diagnosing the problem requires analysis of existing data. This includes finding which of the data are the most likely to explain the problem under consideration, and also finding if there is multicollinearity in the data. Note that SALT was built under the strong belief that the deformation (of the truss) is essentially temperature based. Moreover, there is no geometric information available from measurements. The data of interest in the SALT case are as in Table 2.1 and respectively represent: measurement time (time), temperature of truss (ttr), temperatures of igloos 1 and 2 (ti1 and ti2), wind speed inside the operating room (windin), wind speed outside at about 30 metres above ground level (wind30), humidity inside the operating room (humin), humidity outside at about 30 metres above ground level (hum30), change in radius of curvature (dgroc), reference temperature in the room (tref). Other data of interest are relative heights, and also figure of merit which is computed and tells us how far the measured relative heights are from the range of the linear transformation that maps the actuator positions to relative heights via the actuator-to-heights matrix known from the geometry of the primary mirror. In other words, it tells us mostly about errors in the measurements. Note that these data are collected to find out which explain the figure of merit. Also, change in radius of curvature is computed and is a linear combination of change in temperature of truss. This means information we have from temperature of truss can be obtained from change in radius of curvature and vice versa. Change in radius of curvature is computed as follows, as given in [42]:

$$\Delta\text{GRoC} = \Delta\text{T}_{\text{truss}} \times \text{GRoC} \times \text{CTE}_{\text{steel}}$$

where GRoC is the global radius of curvature of the primary mirror and is known (26.165 metres), $\Delta \mathrm{T}_{\mathrm{truss}}$ is the change in temperature of truss and is given in ^oC/hour, and $\mathrm{CTE}_{\mathrm{steel}}$ is the coefficient of thermal expansion of the steel which is the material used to build the truss, and is also known (11.7×10^{-6}). Change in radius of curvature also gives rise to adjustments in tip/tilts and pistons as follows [42], and respectively denoted by α and c, where for notation convenience, GRoC and ΔGRoC are respectively replaced by R_0 and ΔR:

$$c = \sqrt{\Delta R^2 + (R_0 + \Delta R)^2 - 2\Delta R\,(R_0 + \Delta R)\cos(\omega)} - R_0$$

$$\text{tip/tilt} = \alpha = \arccos\left[\frac{(R_0 + c)^2 + (R_0 + \Delta R)^2 - \Delta R^2}{2\,(R_0 + c)\,(R_0 + \Delta R)}\right]$$

Here ω is the *dihedral angle* between the segment under consideration and the central segment. It is to be noted that the above equations for tip/tilts and piston adjustments can be derived from basic knowledge of Euclidean geometry in the plane.

Table 2.1: Data of interest for SALT in diagnosing imperfection

Variable	Units
time	hours
ttr	^{o}C
ti1	^{o}C
ti2	^{o}C
windin	m/s
wind30	m/s
humin	%
hum30	%
dgroc	metres
tref	^{o}C

Figure 2.1 shows the figure of merit, that is, the Root Mean Square (RMS) of $s - Az$ where at each time, $s(t)$ is the vector representing the relative heights as measured by sensors and $z(t)$ is the estimation of the corresponding actuator displacements using the least squares method.

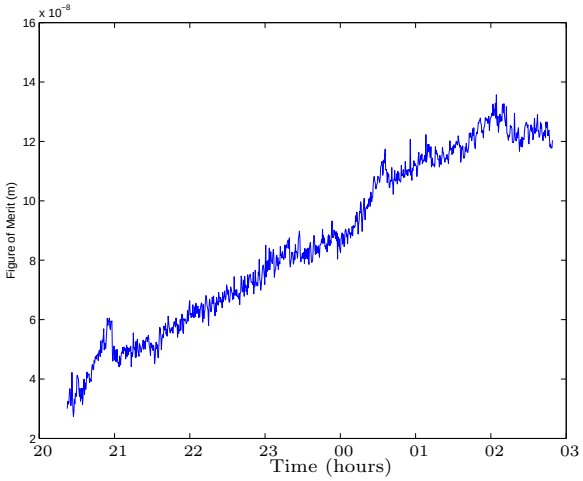

Figure 2.1: Figure of merit in one night (08 March 2007). Figure of merit increasing with time and going beyond 60nm $= 6 \times 10^{-8}$m indicates failure of primary mirror control with time

Recall that the RMS of a vector $v = (v_1, v_2, \ldots, v_n)$ is given by

$$\mathrm{RMS}\,(v) = \sqrt{\frac{v_1^2 + v_2^2 + \cdots + v_n^2}{n}} \tag{2.1}$$

This figure of merit was evaluated during the night of 08 March 2007. Note that there was a failure of primary mirror control with time.

2.1 Method

In this section, we lay out some statistical notions used in the data analysis investigated in this chapter. The methods are well known. Suppose we have a list of variables (data sets, such as time series) Y, X_1, X_2, ..., X_p from a sample of size n (the *sample size* is the length of each of the data sets) where Y is our *response variable*. The other variables are the *explanatory variables*, as in Table 2.1 for the SALT case.

2.1.1 Tests for Multicollinearity

If we have two or more explanatory variables, then there is *multicollinearity* [35, 36] in the model when one of these variables can be approximated as a linear combination of the other variables. Thus, dgroc (change in radius

of curvature) is linearly related to change in temperature. In this case an important optical measure is possibly explained by temperature, which in turn suggests control solution. Three methods for testing multicollinearity are explored below.

Correlation between Explanatory Variables

We consider two random vectors X and Y, where $X = (X_1, X_2, \ldots, X_n)$ and $Y = (Y_1, Y_2, \ldots, Y_n)$. The covariance [2] between X and Y is defined as follows:

$$\sigma_{XY} = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X}\right) \left(Y_i - \bar{Y}\right)$$

and the correlation coefficient [2] between X and Y is defined as follows:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right) \left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2} \sqrt{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}}$$

where σ_X and σ_Y respectively stand for the *standard deviations* of X and Y ($\sigma_X = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}$), and $\bar{X}$ and $\bar{Y}$ stand for the *means* of X and Y respectively ($\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$). Moreover, ρ_{XY} is always such that $-1 \leq \rho_{XY} \leq 1$. Note that the covariance and correlation coefficient also apply for samples of random variables.

Multicollinearity is of concern when for example, two or more explanatory variables are highly correlated, that is, if the correlation coefficient of two of them, X and Y, say, is such that $\rho_{XY} \simeq \pm 1$. This method gives a hint on how to divide explanatory variables into small groups, depending on their correlation coefficients.

Condition Number

As we have seen in Chapter 1, matrix computation is required for the primary mirror control. We consider an $n \times n$ matrix A. The $\| \cdot \|_2$ *condition number* of A [19, 28] is defined as follows:

$$\kappa(A) = \begin{cases} \|A\|_2 \|A^{-1}\|_2 & \text{if } A \text{ is nonsingular;} \\ \infty & \text{if } A \text{ is singular.} \end{cases}$$

If $A = U\Sigma V^T$ is a *singular value decomposition* of A (SVD will be explored later in this thesis), and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$ are the singular values of A (note that $\sigma_i \geq 0 \ \forall i$), then

$$\kappa(A) = \frac{\sigma_1}{\sigma_n}. \tag{2.2}$$

Also note that, considering A as the matrix of a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$ [28],

$$\|A\|_2 = \sigma_1 = \max_{\|x\|_2=1} \max_{\|y\|_2=1} \left|y^T A x\right| \tag{2.3}$$

where x varies in $\mathbb{R}^n$, y varies $\mathbb{R}^m$ and $\|\cdot\|_2$ is the standard Euclidean norm.

Another way to investigate multicollinearity (and that can be of interest to SALT) is to check the condition number of the matrix $X^T X$ where X is the *augmented* matrix of explanatory variables, that is, the matrix of explanatory variables with a column of 1s added at the beginning, more precisely:

$$X = \begin{bmatrix} \mathbf{1}_n & X_1 & X_2 & \cdots & X_p \end{bmatrix}$$

where $X_1, X_2, \ldots, X_p$ are the explanatory variables in column vectors of length n each, and $\mathbf{1}_n$ in a column vector of length n with all entries equal to one. This (exploration of the condition number of $X^T X$) is due to the fact that the ordinary least squares determination of the regression coefficients associated to the explanatory variables involves the inversion of the matrix $X^T X$ mentioned above in this section. If the condition number of X is high (an informal approach is to use $\kappa(X) > 30$), it means that X is ill-conditioned. In that case, multicollinearity is of concern. This method just gives global information about multicollinearity with no specific detail on any of the explanatory variables.

Variance Inflation Factor

The variance inflation factor [9, 17, 34, 37, 49] is an indicator of multicollinearity. The study of multicollinearity using the variance inflation factor involves multiple regression analysis (as laid out later in section 2.1.2) between explanatory variables. The variance inflation factor (for each explanatory variable) can be determined in three steps:

1. Choose one variable X_k amongst all p variables and run the multiple

regression analysis against the other variables

$$X_k = \beta_{k,0}\mathbf{1}_n + \sum_{\substack{i=1 \\ i\neq k}}^{p} \beta_{k,i}X_i + \varepsilon_k$$

and do the same for each of the explanatory variables

2. Compute the R^2 (coefficient of determination) of the regression model (denoted by R_k^2 for the variable X_k)

3. The corresponding variance inflation factor is $\text{VIF}_k = \frac{1}{1-R_k^2}$

Note that the variance inflation factor is always such that VIF ≥ 1. It can be established from the definition of the R^2 that

$$\text{VIF}_k = \frac{\sum_{i=1}^{n}(X_{k,i} - \bar{X}_k)^2}{\sum_{i=1}^{n}(X_{k,i} - \hat{X}_{k,i})^2} \qquad 1 \leq k \leq p$$

where $\hat{X}_k$ is the estimate of X_k in the multiple regression against the other $(X_l)_{1\leq l\leq p,\ l\neq k}$.

A variance inflation factor VIF > 10, meaning $R^2 > 0.9$ [9, 17, 34, 37, 49] (or VIF > 5, meaning $R^2 > 0.8$) for the k^{th} explanatory variable is an indicator that this k^{th} variable is involved in multicollinearity.

How to Handle Multicollinearity

One interpretation of multicollinearity is that the corresponding variables (those involved in multicollinearity) are very likely to convey similar information. One way to deal with multicollinearity is *sequential variable selection* [35, 36]. In this case, one or more of the many collinear variables may be dropped. Another option is *principal component analysis (PCA)* [9] provided the new variables have a meaningful interpretation. These new variables are called *principal components*. However, an efficient way to find a model with the minimum possible number of explanatory variables is *stepwise regression*, explained in the next section.

2.1.2 Regression Analysis

Linear regression analysis is concerned with the estimation of the response variable as a linear combination of the explanatory variables. In this thesis,

linear regression will simply be referred to as *regression*. If we have one response variable Y and p explanatory variables $X_1, X_2, \ldots, X_p$, all column vectors of length n where n is the number of observations, the regression problem (approximating Y as a linear combination of $X_1, X_2, \ldots, X_p$) can be formulated as follows [9]:

$$Y = X\beta + \varepsilon \tag{2.4}$$

where $\beta = \begin{bmatrix}\beta_0 & \beta_1 & \cdots & \beta_p\end{bmatrix}^T$, $X = \begin{bmatrix}\mathbf{1}_n & X_1 & \cdots & X_p\end{bmatrix}$, $\mathbf{1}_n$ is a column vector of length n with all components equal to one, and $\varepsilon = \begin{bmatrix}\varepsilon_1 & \cdots & \varepsilon_n\end{bmatrix}^T$ is the *error term*. If we are only using one explanatory variable, in other words $p = 1$, then we are performing a *simple regression*. If we are using all the (more than one) explanatory variables, in other words $p > 1$, then we are performing a *multiple regression*. Since the regression coefficients are determined in order to minimise the error $S(\beta) = (Y - X\beta)^T (Y - X\beta)$, we obtain [9]:

$$\left(X^T X\right) \hat{\beta} = X^T Y \tag{2.5}$$

which is the set of *normal equations* corresponding to Problem (2.4). If $\left(X^T X\right)$ is nonsingular, then we obtain the estimated regression coefficients in a vector form as follows [9]:

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y \tag{2.6}$$

and consequently, the fitted response is given by [9]:

$$\hat{Y} = X\hat{\beta} = PY \quad \text{where} \quad P = X\left(X^T X\right)^{-1} X^T. \tag{2.7}$$

Note that $\hat{\beta}$ is known as the *best linear unbiased estimator (BLUE)* of β, and that P is symmetric and idempotent.

The significance of an explanatory variable in a regression model is measured by its *p-value*. This is obtained from inference on regression coefficients done as follows: For a variable X_j, $(1 \leq j \leq p)$, we test the *null hypothesis* H_0 against the *alternate hypothesis* H_1 where:

$$\begin{cases} H_0: & \beta_j = 0; \\ H_1: & \beta_j \neq 0. \end{cases}$$

The test is done by computing the t-statistic

$$t_j = \frac{\hat{\beta}_j}{s.e.\left(\hat{\beta}_j\right)} \tag{2.8}$$

and this t-statistic has a student's t-distribution with $n-p-1$ degrees of freedom. Note that *s.e.* stands for *standard error*, which is an estimate of the standard deviation, for each regression coefficient. Moreover,

$$s.e.\left(\hat{\beta}_j\right) = \hat{\sigma}\sqrt{c_{jj}} \tag{2.9}$$

where $C = \left(X^T X\right)^{-1}$ and

$$\hat{\sigma}^2 = \frac{\varepsilon^T \varepsilon}{n-p-1} = \frac{Y^T\left(I_n - P\right)Y}{n-p-1} \tag{2.10}$$

with I_n being the $n \times n$ identity matrix and P as given in (2.7). We compare t_j with $t_{(n-p-1,\alpha/2)}$ obtained from the t-table, where α is the significance level. H_0 is rejected at significance level α if [9]

$$|t_j| \geq t_{(n-p-1,\alpha/2)}, \quad \text{or equivalently,} \quad p\left(|t_j|\right) \leq \alpha.$$

Here, $p\left(|t_j|\right)$ is the p-value of the test (for β_j) and is the probability that a random variable having a student's t-distribution with $n-p-1$ degrees of freedom, is greater than $|t_j|$ in magnitude, that is, the area above the x-axis and under the curve of the probability density function of the student's t-distribution with $n-p-1$ degrees of freedom, outside the range $[-|t_j|, |t_j|]$. This is given by:

$$p\left(|t_j|\right) = 1 - 2\int_0^{|t_j|} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} dt \tag{2.11}$$

where the integrand in (2.11) is the probability density function of the student's t-distribution with ν degrees of freedom [23] and is an even function defined for all $t \in \mathbb{R}$. Here $\nu = n-p-1$ and Γ is the *Gamma* function defined as follows:

$$\Gamma\left(z\right) = \int_0^{\infty} x^{z-1}e^{-x}dx \quad z > 0. \tag{2.12}$$

An explanatory variable is considered significant if its p-value is less than a specified value α, and insignificant otherwise. The goodness of fit of a regression model is explained by the R^2 also called *coefficient of determination* [14, 27] and defined as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \tag{2.13}$$

or, equivalently [9] by the *multiple correlation coefficient* $R = \sqrt{R^2}$ of the response variable Y on the explanatory variables $X_1, \ldots, X_p$, where Y_i is the i^{th} component of Y; $\bar{Y}$ is the mean of Y (that is $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$); $\hat{Y}_i$ is the estimated value for Y_i from the regression (as given in expression (2.7) or equivalently in (2.21) for a general case). Since $0 \leq R^2 \leq 1$, the more R^2 is close to 1, the better the data fits the model. A modification of the R^2 that adjusts for the number of explanatory variables is called the *adjusted* R^2 and is defined as follows [9]:

$$\bar{R}^2 = 1 - \left(1 - R^2\right) \frac{n-1}{n-p-1}. \tag{2.14}$$

Note that $\bar{R}^2 \leq R^2$ and it is possible that $\bar{R}^2 < 0$. In some documents from the literature (see for example [9]), the adjusted R^2 is denoted by R_a^2.

The global significance of all the variables appearing in a model can be measured by the *general p-value* of the model. This is obtained by testing the adequacy of a model against another model as follows [9]: We consider two models M_0 and M_1 with q_0 and q_1 explanatory variables respectively. We assume that M_0 is a *sub-model* of M_1, that is, all the variables appearing in M_0 also appear in M_1. We also assume that $0 \leq q_0 < q_1 \leq p$. We test the null hypothesis H_0 against the alternate hypothesis H_1 defined as follows:

$$\begin{cases} H_0: & M_0 \text{ is adequate;} \\ H_1: & M_1 \text{ is adequate.} \end{cases}$$

If the goodness of fit (the R^2) of M_0 is greater or equal to the goodness of fit of M_1, then H_0 is considered and H_1 is rejected. The respective degrees of freedom for models M_0 and M_1 are $df_0 = n - q_0 - 1$ and $df_1 = n - q_1 - 1$. The *F-test* or *F-statistic* to see whether model M_0 is adequate is given by:

$$\begin{aligned} F &= \frac{\left[\text{SSE}\left(M_0\right) - \text{SSE}\left(M_1\right)\right] / (df_0 - df_1)}{\text{SSE}\left(M_1\right) / df_1} \\ &= \frac{\left[\text{SSE}\left(M_0\right) - \text{SSE}\left(M_1\right)\right] / (q_1 - q_0)}{\text{SSE}\left(M_1\right) / \left(n - q_1 - 1\right)} \end{aligned} \tag{2.15}$$

where

$$\text{SSE}\left(M_0\right) = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i^0\right)^2 \quad \text{and} \quad \text{SSE}\left(M_1\right) = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i^1\right)^2 \tag{2.16}$$

are the sums of squared residuals due to respectively fitting models M_0 and M_1 to the data, with $\hat{Y}^0$ and $\hat{Y}^1$ being the estimates (predictions) of Y using

models M_0 and M_1 respectively. Note that the F-statistic from Equation (2.15) has F distribution with $q_1 - q_0$ and $n - q_1 - 1$ degrees of freedom. We compare F, the observed value of the F-test as given in (2.15), with $F_{(q_1-q_0,n-q_1-1;\alpha)}$ which is the corresponding critical value obtained from the F table, where α is the significance level. H_0 is rejected at significance level α if [9]

$$F \geq F_{(q_1-q_0,n-q_1-1;\alpha)} \quad \text{or equivalently} \quad p(F) \leq \alpha$$

where $p(F)$ is the p-value for the F-test, that is, the probability that a random variable having F distribution with $q_1 - q_0$ and $n - q_1 - 1$ degrees of freedom, is greater than the observed F-test as given in (2.15). The p-value is the area above the x axis and under the curve of the probability density function of the F distribution with $q_1 - q_0$ and $n - q_1 - 1$ degrees of freedom, in the range $[F, \infty)$ where F is given in (2.15), that is:

$$p(F) = 1 - \int_0^F \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}}}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)\left(\frac{1+\nu_1}{\nu_2}t\right)^{\frac{\nu_1+\nu_2}{2}}} t^{\frac{\nu_1-2}{2}} dt \tag{2.17}$$

where the integrand in (2.17) is the probability density function of the F distribution with ν_1 and ν_2 degrees of freedom [23] and is defined for all $t > 0$, $\nu_1 = q_1 - q_0$, $\nu_2 = n - q_1 - 1$ and Γ is the Gamma function as given in (2.12).

Simple Regression

Simple regression [14, 34, 49] is used to check if each explanatory variable alone is significant to explain the response variable, and how good the data used fits the corresponding regression model. If we are dealing with only one explanatory variable, say X_1, which is a vector of length n, then from Equation (2.4), the *simple regression* problem can be formulated as follows:

$$Y = \beta_0 \mathbf{1}_n + \beta_1 X_1 + \varepsilon \tag{2.18}$$

where $\mathbf{1}_n$ and ε are the same as in (2.4).

In this case, $\hat{Y} = X\hat{\beta}$ can be rewritten as

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i,1}, \quad 1 \leq i \leq n \tag{2.19}$$

with $\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$.

Note that the p-value of the simple regression model is obtained from (2.17) by taking $q_0 = 0$ and $q_1 = 1$.

Multiple Regression

Multiple regression [11, 14, 15, 27, 35, 36, 45] is used to check how many explanatory variables together explain the response variable, and how well the data fits the regression model. If we want to perform a multiple regression analysis on the variables given at the beginning of this section (Section 2.1.2), of course under the assumption that $p > 1$, then Equation (2.4) can be reformulated as follows:

$$Y = \beta_0 \mathbf{1}_n + \sum_{j=1}^{p} \beta_j X_j + \varepsilon \tag{2.20}$$

where $\mathbf{1}_n$ and ε are the same as in (2.4). A detailed expression using vector matrix formulation is as follows:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{1,1} & X_{1,2} & \cdots & X_{1,p} \\ 1 & X_{2,1} & X_{2,2} & \cdots & X_{2,p} \\ 1 & X_{3,1} & X_{3,2} & \cdots & X_{3,p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & X_{n,1} & X_{n,2} & \cdots & X_{n,p} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

In this case, $\hat{Y} = X\hat{\beta}$ can be rewritten as

$$\hat{Y}_i = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j X_{i,j}, \quad 1 \leq i \leq n \tag{2.21}$$

with $\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$.

Note that the p-value of the multiple regression model is obtained from (2.17) by taking $q_0 = 0$ and $q_1 = p$.

Stepwise Regression

Stepwise regression [14, 27] is an improvement on multiple regression. It is used to determine which explanatory variables are significant in explaining the response variable. This is extremely important in the process of building a regression model with the least possible explanatory variables. It is a statistical procedure that considers all the variables as in the multiple regression procedure, and step by step adds significant explanatory variables and excludes insignificant explanatory variables from the regression model. This tells us which explanatory variables are the main causes for the response variable. Stepwise regression is a combination of *forward selection*

(FS) and *backward elimination* (BE) [9], depending on the criteria for FS or BE. The procedure for stepwise regression is as follows:

1. On a data set X with p explanatory variables $X_1, X_2, \ldots, X_p$, start with an initial multiple regression model. The default is usually no term (no explanatory variable), in which case the response variable Y is approximated by a constant.

2. Add to the regression model the term with the smallest p-value if this p-value is less than a given *entrance tolerance*, and repeat the process until there is no term left to add. The default entrance tolerance is $p_{in} = 0.05$.

3. Remove from the model the term with the highest p-value if this p-value is greater than a given *exit tolerance* and go back to step 2. The default exit tolerance is $p_{out} = \max(p_{in}, 0.1)$.

4. If there is no term left to remove from the multiple regression model, then stop.

Note that the p-value mentioned in the stepwise regression procedure given above is the p-value of an F-statistic used to compare two models (the current model and the previous model, or the larger model and the reduced model). The explanatory variables that appear in the final model are the significant variables, and those not appearing in the final model are insignificant.

The p-value of the stepwise regression model is obtained after the last F-test. At each step, since we are adding or removing one variable at a time, and M_0 represents the reduced model whilst M_1 represents the larger model, from (2.15) we take $q_1 = q_0 + 1$ and therefore

$$F = \frac{[\mathrm{SSE}\,(M_0) - \mathrm{SSE}\,(M_1)]}{\mathrm{SSE}\,(M_1)\,/\,(n - q_1 - 1)} \tag{2.22}$$

has F distribution with 1 and $n - q_1 - 1$ degrees of freedom. For the forward selection process (rejecting H_0), the default value for significance level α is 0.05 and for the backward elimination process (accepting H_0), the default value for α is 0.1 [20, 49].

2.1.3 Autocorrelation and Spectral Analysis

These concepts are explored in the literature, see for example [4, 7, 8, 23]. These (autocorrelation and spectral analysis) will be performed on open-loop[1] data sets, especially on relative heights from the portion of the primary mirror made of the central segment and the first ring, due to the unavailability of enough man power to maintain the whole primary mirror for our experiments at the time. This makes 7 segments, 21 actuators and 24 sensors. This can also be performed on the whole primary mirror. These measured relative heights are considered as a random process $X(t, n)$ where at each time t, the relative heights as given by the sensors constitute a vector denoted by $X_{t,*}$. This $X_{t,*}$ is considered as a realization of the random process at time t. Similarly, for each n, the time series of measurements as given by a specific sensor (the sensor corresponding to index n) is denoted by $X_{*,n}$. When there is no confusion, $X_{t,*}$ will be denoted by X_t and $X_{*,n}$ will be denoted by X_n. One of the objectives is to test for the stationarity of the process.

Autocorrelation

For one random process $X(t, n)$ with realizations X_s and X_t at times s and t, the corresponding *autocorrelation* is defined as follows:

$$R_{XX}(s,t) = \mathbb{E}\left(X_s X_t\right) \tag{2.23}$$

and the *autocovariance* is defined as follows:

$$C_{XX}(s,t) = \mathbb{E}\left[\left(X_s - \mu_X(s)\right)\left(X_t - \mu_X(t)\right)\right] = R_{XX}(s,t) - \mu_X(s)\mu_X(t)$$

where $\mu_X(s) = \mathbb{E}\left(X_s\right)$ and $\mu_X(t) = \mathbb{E}\left(X_t\right)$ are the respective means of the (realization of the) random process at times s and t.

A random process $X(t, n)$ is said to be *(wide sense) stationary* or *weakly stationary* [8, 17, 39] if the following conditions are satisfied:

- the mean $\mu_X(t)$ does not depend on t, that is, there exists a constant μ such that $\mu_X(t) = \mu$ for all $t \geq 0$

[1] It consists, in this context, of taking measurements without performing any control on the system

- the variance $\sigma^2_{X_t}$, or equivalently the standard deviation σ_{X_t} does not depend on t, that is, there exists a constant σ such that $\sigma_{X_t} = \sigma$ for all $t \geq 0$

- the autocovariance $C_{XX}(s,t)$ between X_s and X_t only depends on the difference between time units $t - s$, not X_s and X_t. In this case, if we have $\tau = t - s$ then $C_{XX}(s,t) = C_{XX}(s, s+\tau) = C_{XX}(\tau)$ is called the *autocovariance coefficient* at lag τ, and is sometimes denoted by γ_τ or $C_X(\tau)$.

If $X(t,n)$ is wide sense stationary, then the autocorrelation $R_{XX}(s,t)$ only depends on $t - s$, not X_s and X_t. So if $\tau = t - s$, then $R_{XX}(s,t) = R_{XX}(s, s+\tau) = R_{XX}(\tau)$ is called the *autocorrelation coefficient* at lag τ and is sometimes denoted by $R_X(\tau)$. However, if $X(t,n)$ is not wide sense stationary, then we can approximate the autocorrelation coefficient at lag τ as follows: $\hat{R}_X(\tau) = \mathbb{E}_t\left[R_{XX}(t, t+\tau)\right]$.

Moreover, in practice, time series are given as measurements in discrete time, most likely equally spaced, and in a finite range $(X_t)_{t=0,1,\ldots,N-1}$ where N is a positive integer. In that case, the (biased) estimate of the autocorrelation is given by

$$\hat{R}_X(\tau) = \frac{1}{N} \sum_{t=0}^{N-\tau-1} X_t X_{t+\tau} \tag{2.24}$$

and the unbiased estimate is given by

$$\bar{R}_X(\tau) = \frac{1}{N-\tau} \sum_{t=0}^{N-\tau-1} X_t X_{t+\tau}. \tag{2.25}$$

Note that in this case $\bar{R}_X(\tau) = \frac{N}{N-\tau}\hat{R}_X(\tau)$.

From now on, unless otherwise specified, the term *stationary* will refer to *weakly stationary*.

Spectral Analysis

Spectral analysis is analysis of the *spectrum* of a time series. The spectrum is defined for continuous as well as discrete time series, and involves the notion of *Fourier Transform* [12, 39], which is also defined in the continuous as well as the discrete sense. We will focus on the discrete Fourier Transform and its inverse. Again, we recall that the method is well known.

Consider a time series, that is, a discrete-time signal, most generally a complex-valued sequence $(x_n)_{0\leq n\leq N-1}$ where N is a positive integer. The *discrete Fourier transform (DFT)* maps the sequence $(x_n)_{0\leq n\leq N-1}$ into a sequence $(X_k)_{0\leq k\leq N-1}$ defined as follows:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N}kn} \qquad 0 \leq k \leq N-1$$

where $e^{\frac{2\pi i}{N}}$ is a primitive N^{th} root of unity in the set $\mathbb{C}$ of complex numbers. A short notation is $X = \mathcal{F}(x)$, or $\mathcal{F}\{x\}$, or $\mathcal{F}_x$. The *inverse discrete Fourier transform (IDFT)* maps the sequence $(X_k)_{0\leq k\leq N-1}$ back to the sequence $(x_n)_{0\leq n\leq N-1}$ and is defined as follows:

$$x_n = \frac{1}{N}\sum_{k=0}^{N-1} X_k e^{\frac{2\pi i}{N}kn} \qquad 0 \leq n \leq N-1$$

and can be denoted as $x = \mathcal{F}^{-1}(X)$ or $\mathcal{F}^{-1}\{X\}$ or $\mathcal{F}_X^{-1}$.

Let (x_n) be defined for all $n \in \mathbb{Z}$ and

$$\omega = \frac{2\pi}{N}k = 2\pi fT$$

The discrete-time Fourier transform (or DTFT) of x_n, is defined as follows:

$$X(\omega) = \sum_{n=-\infty}^{\infty} x_n e^{-i\omega n} \qquad -\pi \leq \omega < \pi. \tag{2.26}$$

Here, $\omega = 2\pi fT$ is the *continuous normalised radian frequency variable*, T is the period of the DTFT and f is the frequency.

The original discrete-time sequence can be recovered by applying to $X(\omega)$ the inverse transforms defined as follows:

$$\begin{aligned} x_n &= \frac{1}{2\pi}\int_{-\pi}^{\pi} X(\omega)\cdot e^{i\omega n}\, d\omega \\ &= T\int_{-\frac{1}{2T}}^{\frac{1}{2T}} X_T(f)\cdot e^{i2\pi fnT}\, df \end{aligned} \tag{2.27}$$

where $X_T(f) = X(\omega) = X(2\pi fT)$.

In practice, for numerical evaluation of the DTFT, a finite-length sequence is needed and recommended. A long sequence can be modified by truncation (that is, by applying for example a rectangular window function), resulting in:

$$X(\omega) = \sum_{n=0}^{L-1} x_n \, e^{-i\omega n} \qquad -\pi \leq \omega < \pi \tag{2.28}$$

where L is the modified sequence length. This is often a useful approximation of the spectrum of the unmodified sequence. In numerical procedures, it is natural, or common, to evaluate $X(\omega)$ at an arbitrary number N of uniformly-spaced frequencies across one period (interval of length 2π):

$$\omega_k = \frac{2\pi}{N} k \quad \text{for } k = 0, 1, \ldots, N-1 \tag{2.29}$$

which gives:

$$X_k = X(\omega_k) = \sum_{n=0}^{L-1} x_n \, e^{-i2\pi \frac{k}{N} n} \tag{2.30}$$

When $N \geq L$, this can also be written:

$$X_k = \sum_{n=0}^{N-1} x_n \, e^{-i2\pi \frac{k}{N} n} \tag{2.31}$$

provided we define $x_n = 0$ for $n \geq L$.

This adjustment makes the X_k sequence now recognizable as a discrete Fourier transform (DFT). Here N is the resolution at which the DTFT is sampled, and L limits the inherent resolution of the DTFT itself. So N and L usually have similar (or equal) values.

If a time series is considered as a random process $X(t, n)$ with realization $X_{t,*} = X_t$ at time t, $R_X(\tau)$ is the autocorrelation as defined in Equation (2.23) when the process is stationary and $\tau = t - s$, and $\hat{R}_X(\tau)$ is the biased estimate of the autocorrelation as defined in Equation (2.24) when the process is not necessarily stationary. The spectrum of the time series (also sometimes called *power spectral density*) is a periodic function with period 2π, and is defined as follows:

$$S_{XX}(\omega) = \mathcal{F}(R_X(\tau)) = \sum_{\tau=-\infty}^{\infty} R_X(\tau) e^{-i\omega\tau} = R_X(0) + 2\sum_{\tau=1}^{\infty} R_X(\tau)\cos(\omega\tau)$$

if the process is stationary [23], and

$$S_{XX}(\omega) = \mathcal{F}\left(\mathbb{E}_t\left[R_{XX}(t, t+\tau)\right]\right) = \mathcal{F}\left(\hat{R}_X(\tau)\right)$$

otherwise. Here $\mathbb{E}_t$ denotes the *expected value with respect to* t, τ denotes the *lag*, and $\mathcal{F}$ denotes the *discrete time Fourier transform*. However, if we are concerned with a discrete time series in a bounded time range $(X_t)_{0\leq t\leq N-1}$ where N is a positive integer, then the alternate form of the spectrum is given as follows:

$$S_{XX}(\omega) = \frac{1}{N}\left|\sum_{k=0}^{N-1} X_k e^{-i\omega k}\right|^2 = \frac{1}{N}\sum_{m=0}^{N-1}\sum_{k=0}^{N-1} X_k X_m e^{-i\omega(k-m)}$$

and it can be established that [23]

$$S_{XX}(\omega) = \sum_{\tau=-(N-1)}^{N-1} \hat{R}_X(\tau) e^{-i\omega\tau}$$

2.2 Application to SALT

Previous studies and assessments before our study resulted in the conclusion that the poor image quality of SALT was due to high humidity conditions. Certainly, the performance was unsatisfactory in extremely high humidity situations, but high humidity situations were not the only causes for unsatisfactory performance. Besides relative heights and figure of merit (FoM), the variables we will study are listed in Table 2.1. In this thesis, we perform analysis of all available data, both from environmental measurements and computational methods. We illustrate statistical analyses for the observations of the night of 08 March 2007, and give a global outcome from 240 data sets. In each of the earlier cases, SALT primary mirror was actively controlled by the original SALT algorithm. In Figure 2.1, we have illustrated the behavior of FoM. It can be seen that FoM increases almost linearly with time, which suggests that errors in measurements are getting bigger with time. This should be a reason for concern, since faulty sensors alone are not sufficient enough to explain the behavior of FoM. Figures 2.2 to 2.10 provide visual comparisons between FoM and each of the explanatory variables from Table 2.1

Figures 2.2 and 2.3 provide comparison between FoM and both inside and outside humidities. Inside and outside humidities are increasing on average, and have similar patterns. Outside humidity is higher than inside humidity and fluctuations are higher in amplitude on inside humidity than on outside humidity. Although the figure of merit is highly correlated to

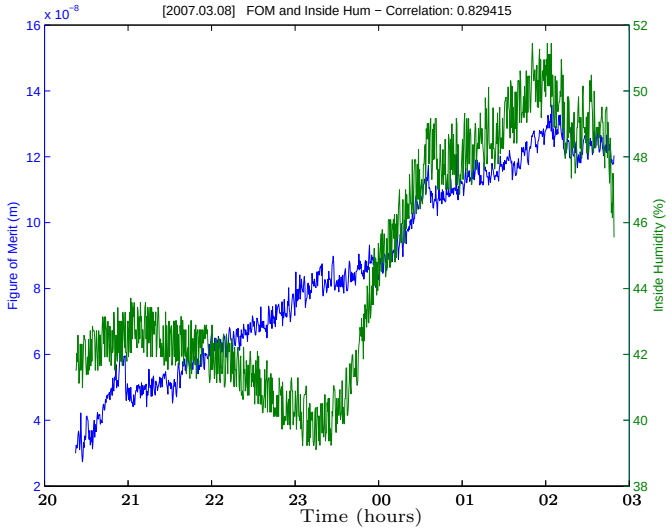

Figure 2.2: FoM and inside humidity vs time (08 March 2007): humidity is relatively low and highly correlated to FoM, but they don't have similar behavior

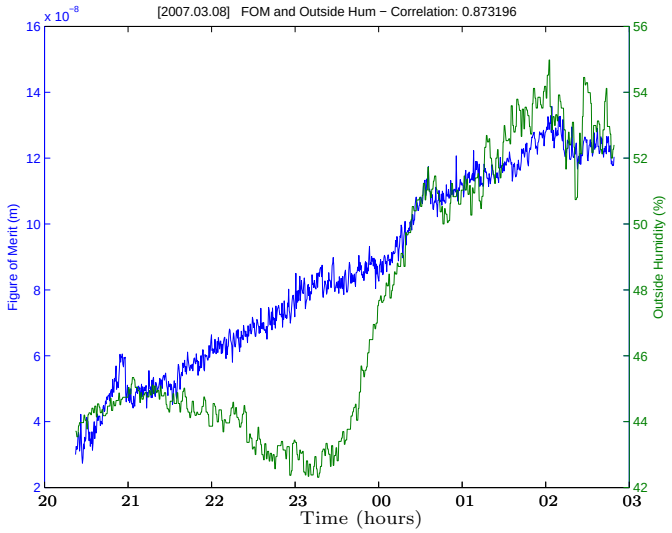

Figure 2.3: FoM and outside humidity vs time (08 March 2007): humidity is relatively low and highly correlated to FoM, but they don't have similar behavior

both humidities (correlation coefficients 0.83 and 0.87), the correlation is higher (in amplitude) with outside humidity (0.87).

Comparison between FoM and both inside and outside wind speeds is provided in Figures 2.4 and 2.5. Both wind speeds are stable on average

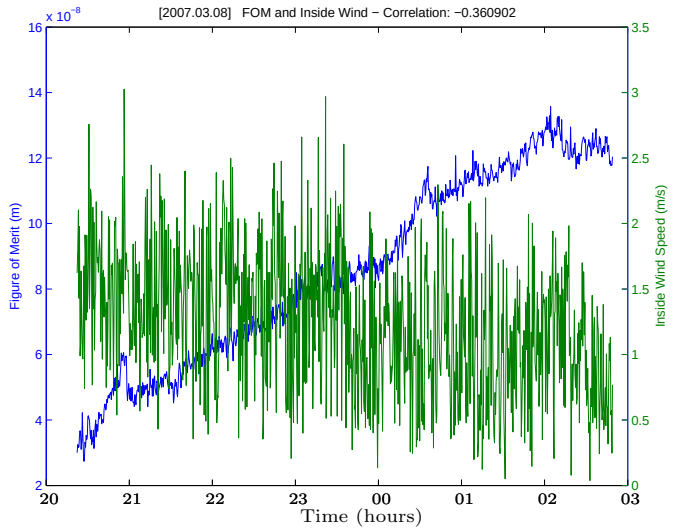

Figure 2.4: FoM and inside wind speed vs time (08 March 2007): wind speed is relatively stable and not highly correlated to FoM, and they don't have similar behavior

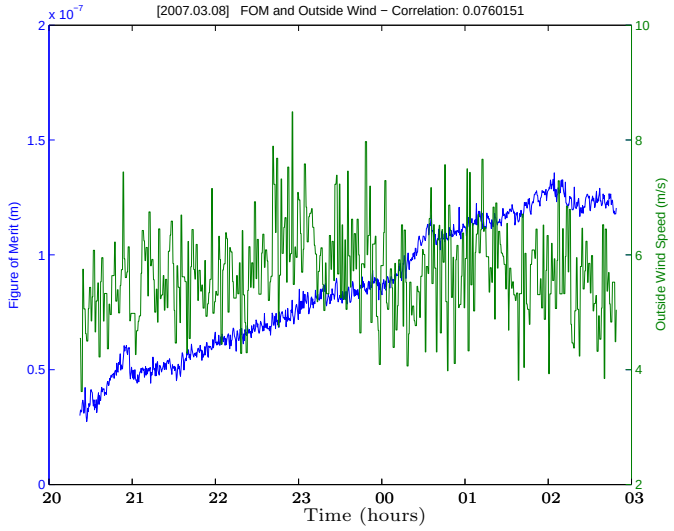

Figure 2.5: FoM and outside wind speed vs time (08 March 2007): wind speed is relatively stable and not highly correlated to FoM, and they don't have similar behavior

(around 1 to 1.5m/s for inside wind speed and around 6m/s for outside wind speed). Outside wind speed is indeed higher than inside wind speed. The

correlation of FoM with both wind speeds is very low (correlation coefficients -0.36 and 0.08) and the correlation is higher in amplitude with inside wind speed (-0.36).

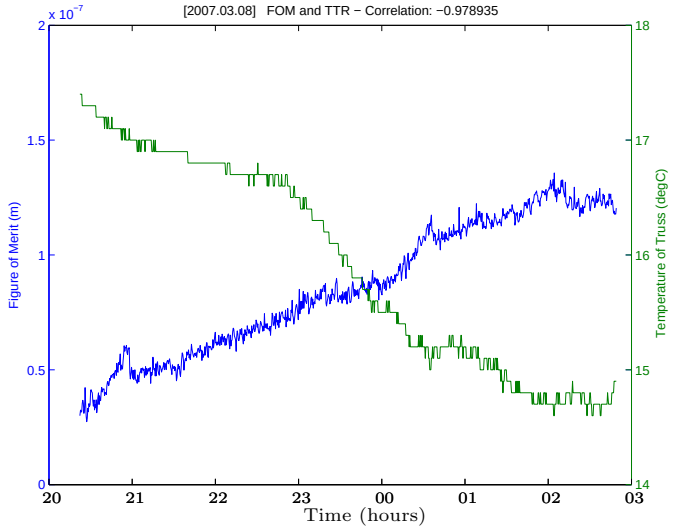

Figure 2.6: FoM and temp of truss vs time (08 March 2007): temperature is relatively low and highly correlated to FoM, and FoM increases as temperature decreases

Figures 2.6 and 2.7 provide visual comparison between FoM and both temperature of truss and reference temperature. Both temperatures are decreasing on average (from about 17.5^oC to about 14.5^oC for both), and have similar patterns. Note that FoM is highly correlated to both temperatures (correlation coefficients -0.9789 and -0.9793) and the correlation is slightly higher with the temperature of truss (-0.9793).

Figures 2.8 and 2.9 illustrate the behavior of FoM and both igloo temperatures. Note that the igloo temperatures have to be kept as steady as possible since they have an impact on the electronics of the telescope. And from both figures, the igloo temperatures are kept as close as possible to 25^oC. Whilst temperature of igloo 1 fluctuates mostly between 24.9^oC and 25.1^oC, temperature of igloo 2 is more stable at 25^oC. Moreover, correlation of FoM with both temperatures is very low (-0.007 and 0.006). This suggests that the igloo temperatures are not likely to contribute in explaining the behavior of FoM.

Figure 2.10 illustrates the behavior of FoM and the change in radius

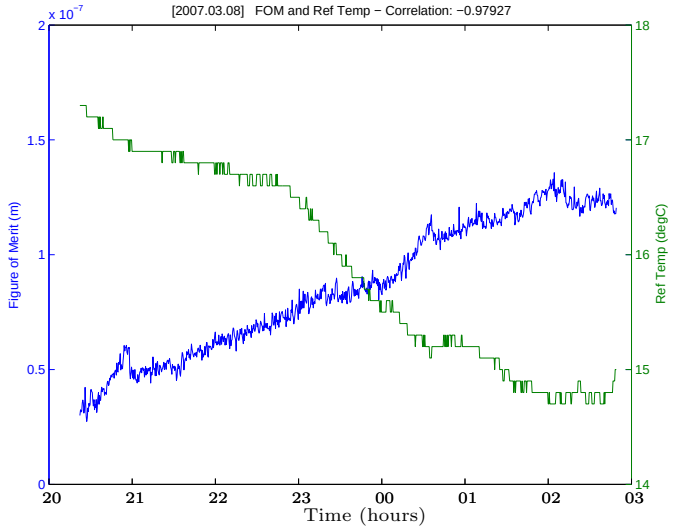

Figure 2.7: FoM and reference temp vs time (08 March 2007): temperature is relatively low and highly correlated to FoM, and FoM increases as temperature decreases

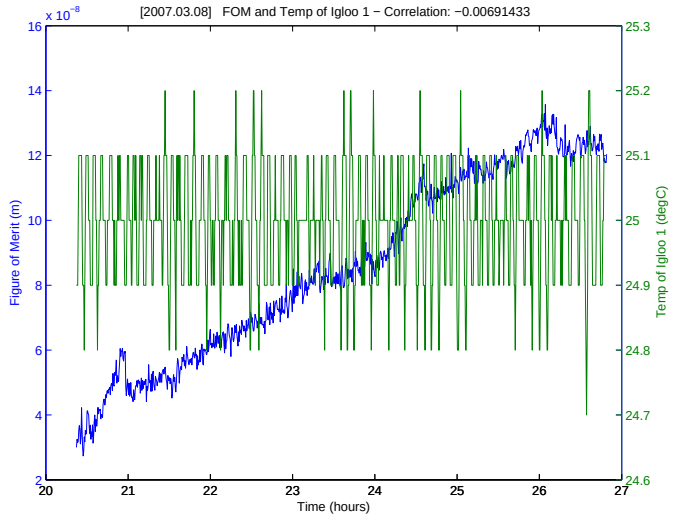

Figure 2.8: FoM and temp of igloo 1 vs time (08 March 2007): temperature is relatively low, stable and not highly correlated to FoM, and they don't have similar behavior

of curvature. The change in radius of curvature decreases in general, and has a pattern similar to those of the temperature of truss and the reference

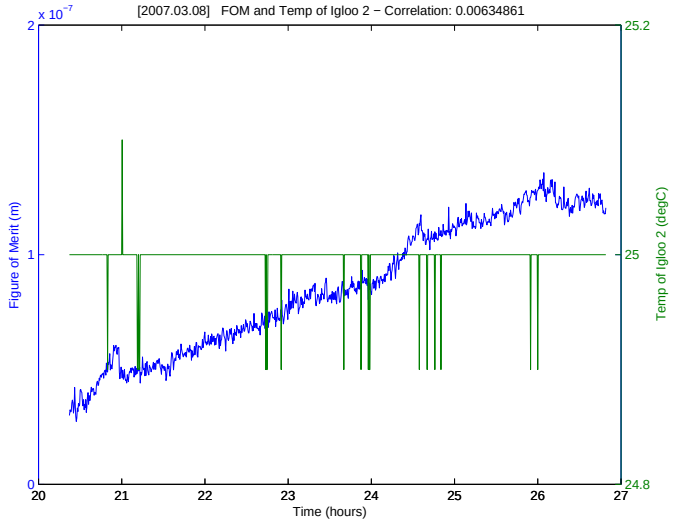

Figure 2.9: FoM and temp of igloo 2 vs time (08 March 2007): temperature is relatively low, stable and not highly correlated to FoM, and they don't have similar behavior

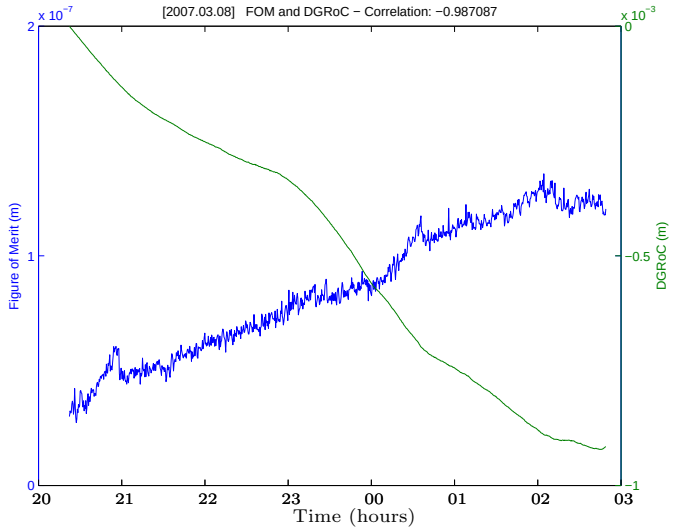

Figure 2.10: FoM and DGRoC vs time (08 March 2007): DGRoC is relatively low, highly correlated to FoM, and decreases as FoM increases

temperature. This is to be expected since the change in radius of curvature is computed and is linearly dependent on the change in temperature. FoM is highly correlated to the change in radius of curvature (correlation coefficient

-0.987) and this is to be expected from inspection of Figures 2.6 and 2.7.

All these conclusions suggest temperature of truss and possibly humidity as the main explanation of the behavior of FoM. Further statistical analysis will give us more information about the main reasons to explain the behavior of the figure of merit.

2.2.1 Tests for Multicollinearity on SALT Data

We perform the tests described in Section 2.1.1 and give results from the night of 08 March 2007 as an illustration, and also global results over 240 data sets.

Correlation Between Explanatory Variables

Table 2.2: Test for multicollinearity using correlation coefficients (08 March 2007)

	time									tref
time	1									
ttr	-0.98	1								
ti1	-0.003	0.006	1							
ti2	0.01	0	0.1	1						
windin	-0.38	0.39	-0.1	-0.08	1					
wind30	0.05	-0.02	0.03	-0.05	0.06	1				
humin	0.81	-0.87	-0.01	0.01	-0.33	-0.1	1			
hum30	0.86	-0.9	0.004	0.01	-0.34	-0.09	0.98	1		
dgroc	-0.99	0.99	0.01	-0.01	0.38	-0.03	-0.86	-0.91	1	
tref	-0.98	0.998	0.01	0.001	0.38	-0.03	-0.86	-0.9	0.99	1

Table 2.2 illustrates the correlation coefficients between explanatory variables for the night of 08 March 2007. Note that this table is a summary of a symmetric matrix, in the sense that the missing values above the main diagonal can be obtained from below the diagonal by transposition. The high correlation coefficients (we chose correlation coefficients greater than 0.75 in magnitude) indicate that there is a strong correlation between 6 variables: time, temperature of truss, inside humidity, outside humidity, change in radius of curvature and reference temperature. This is an indicator of the

fact that there is multicollinearity in the data.

Table 2.3: Overall test for multicollinearity using correlation coefficients (counting high correlations over 240 data sets)

	time									tref
time										
ttr	151									
ti1	16	10								
ti2	42	30	19							
windin	0	4	2	2						
wind30	24	16	5	2	3					
humin	108	115	7	14	4	12				
hum30	117	104	10	21	4	25	110			
dgroc	163	150	15	34	1	16	93	90		
tref	53	69	4	7	1	4	40	34	56	

Table 2.3 (which is also a summary of a symmetric matrix) indicates how many times each pair of explanatory variables has a correlation coefficient higher than 0.75 in magnitude. This suggests that globally, multicollinearity is a concern between the following variables: time, temperature of truss, inside humidity, outside humidity and change in radius of curvature. Note that inside and outside humidities have the lowest rate of high correlation (less than 120 over 240 data sets).

Condition Number

The condition number of X^TX (as given in Section 2.1.1) on the data obtained on 08 March 2007 with SALT is 6.35×10^6, and over 240 data sets, this condition number varies from 1.71×10^4 to ∞. This suggests that globally, multicollinearity should seriously be of concern.

Variance Inflation Factor

The last column of Table 2.6 gives results of the test for multicollinearity for the night of 08 March 2007, and the fact that six of the variance inflation factors are greater than 10 is an indicator of multicollinearity. These variables are: time, temperature of truss, inside humidity, outside humidity,

change in radius of curvature and reference temperature. Note that these are exactly the same variables with high correlation in the previous study.

Table 2.4: Test for multicollinearity on 240 data sets using the variance inflation factor (VIF)

Variable	VIF > 5	VIF > 10
time	226	203
ttr	187	161
ti1	30	25
ti2	51	37
windin	2	1
wind30	68	49
humin	135	105
hum30	155	117
dgroc	185	170
tref	66	61

Table 2.4 gives results of the test for multicollinearity with 240 data sets using the variance inflation factor. It indicates, (out of 240) how many times each explanatory variable has a variance inflation factor VIF > 5, and also VIF > 10. This suggests that globally, the variables involved in multicollinearity are mostly the following: time, temperature of truss, inside humidity, outside humidity, and change in radius of curvature.

2.2.2 Regression Analysis of SALT Data

From the available data, the figure of merit (FoM) is our response variable and is expected to stay as close as possible to zero, with a stable or stationary behavior. The remaining variables are explanatory variables. The ten explanatory variables are those given in Table 2.1. We recall that the response variable in regression analysis is expected to stay as close as possible to zero and the objective function in the optimal control formulation is to be kept as close as possible to zero, but they are two separate concerns. Also note that the observatory is not completely closed and air can flow in and out of the building. Control of humidity, temperature and wind speed in the dome is not practicable since these are environmental and can only be measured for interpretation and analysis.

Simple Regression

Table 2.5: Simple regression with the response variable in terms of each of the explanatory variables (08 March 2007)

Variable	β	**Std Error**	p**-value**	R^2	$\bar{R}^2$
time	1.51×10^{-8}	8.62×10^{-11}	0	0.970	0.970
ttr	-3.11×10^{-8}	2.10×10^{-10}	0	0.958	0.958
ti1	-2.23×10^{-9}	1.05×10^{-8}	0.83	4.78×10^{-5}	-0.001
ti2	1.23×10^{-8}	6.29×10^{-8}	0.84	4.03×10^{-5}	-0.001
windin	-1.98×10^{-8}	1.65×10^{-9}	0	0.130	0.129
wind30	2.73×10^{-9}	1.16×10^{-9}	0.018	0.006	0.005
humin	6.66×10^{-9}	1.45×10^{-10}	0	0.688	0.688
hum30	6.22×10^{-9}	1.12×10^{-10}	0	0.762	0.762
dgroc	-9.88×10^{-5}	5.19×10^{-7}	0	0.974	0.974
tref	-3.25×10^{-8}	2.18×10^{-10}	0	0.959	0.959

Table 2.5 is an illustration of the simple regression analysis on each of the explanatory variables, for the data set of 08 March 2007. It can be established from this table that according to the p-value, each of the explanatory variables, except the temperatures of the igloos, is sufficient to explain the response variable. The case of the igloo temperatures is confirmed by the fact that the standard error is bigger in magnitude than the regression coefficient (β). Moreover, according to the R^2 (and the adjusted R^2), the wind speeds inside and outside the building do not fit the simple regression model very well, although their respective p-values suggest each of them is significant in explaining the figure of merit (response variable). All this together leads to the statement that six explanatory variables are each sufficient to explain the response variable. It means these variables might be of concern regarding multicollinearity. Among these six, the variables time, ttr, dgroc and tref cannot be separated in significance, at all. We note that humin and hum30 are neatly significant. The scenario of more than one explanatory variable being each sufficient to explain the response variable appears to be common in all the data sets provided. This suggests that multicollinearity should be a concern.

Multiple Regression

Table 2.6: Multiple regression with the response variable in terms of all the explanatory variables (08 March 2007)

Variable	β	**Std Error**	t-**stat**	p-**value**	**VIF**
time	2.16×10^{-9}	1.10×10^{-9}	1.9679	0.049	225.99
ttr	-3.79×10^{-10}	2.82×10^{-9}	-0.1343	0.893	349.81
ti1	-1.34×10^{-10}	1.56×10^{-9}	-0.0858	0.932	1.03
ti2	6.90×10^{-9}	9.35×10^{-9}	0.7382	0.461	1.02
windin	8.53×10^{-10}	2.87×10^{-10}	2.9739	0.003	1.21
wind30	1.23×10^{-9}	1.79×10^{-10}	6.8917	1.01×10^{-11}	1.10
humin	1.55×10^{-10}	2.08×10^{-10}	0.7442	0.457	29.72
hum30	-6.63×10^{-10}	2.06×10^{-10}	-3.2241	0.001	36.85
dgroc	-8.29×10^{-5}	1.06×10^{-5}	-7.8087	1.53×10^{-14}	496.34
tref	-2.69×10^{-9}	2.90×10^{-9}	-0.9303	0.352	335.40
$R^2 = 0.97858$; $\bar{R}^2 = 0.978353$; General p-value: 0					

Table 2.6 is an illustration of the multiple regression analysis on all of the explanatory variables for data obtained on 08 March 2007. It can be established from this table that according to the general p-value, the R^2 (and the adjusted R^2), the data fits the regression model very well. Moreover, according to the p-value, the temperature of truss, the temperatures of both igloos, the inside humidity and the reference temperature are not significant, and the remaining variables (time, wind speed in and out of the building, outside humidity and change in radius of curvature) are significant. Once again, the non significance of the explanatory variables ti1, ti2, wind30 and tref is confirmed by their respective regression coefficients being smaller in magnitude than the corresponding standard errors. All this together, with the simple regression outcome, leads to the statement that indeed some variables might be involved in multicollinearity. The scenario of variables being significant in the simple regression case and insignificant in the multiple regression case and vice versa (like for example ttr, humin and tref in the case of data from 08 March 2007) is also common and suggests that multicollinearity should indeed be a concern. It is therefore necessary to deal with the issue of multicollinearity and detect the variables that really

explain the response.

Stepwise Regression

Table 2.7: Stepwise regression (08 March 2007)

Variable	β	**Std Error**	t-**stat**	p-**value**	**VIF**
time	1.83×10^{-9}	1.001×10^{-9}	1.8317	0.067	188.37
ttr	-2.64×10^{-9}	1.26×10^{-9}	-2.0951	0.036	69.85
windin	8.54×10^{-10}	2.83×10^{-10}	3.0125	0.003	1.18
wind30	1.23×10^{-9}	1.78×10^{-10}	6.8996	9.52×10^{-12}	1.09
hum30	-5.50×10^{-10}	1.34×10^{-10}	-4.0928	4.62×10^{-5}	15.75
dgroc	-8.62×10^{-5}	1.001×10^{-5}	-8.6085	0	442.67
$R^2 = 0.978536$;		$\bar{R}^2 = 0.9784$;		General p-value: 0	

Table 2.7 is an illustration of the stepwise regression analysis on all of the explanatory variables for data obtained on 08 March 2007. It can be established from this table that according to the general p-value, the R^2 (and the adjusted R^2), the data fits the model very well. The variables in this table are the only significant variables from the final model of the stepwise regression process. Note that all the p-values are smaller than 0.1 and all the standard errors are smaller than the corresponding regression coefficients in magnitude. The presence of time and change in radius of curvature suggests that errors grow systematically, independently of the environmental conditions (temperature, wind speed and humidity in the dome). Systematic growth of errors in the computation of control of the mirrors is to be considered as an additional explanation. The scenario of variables being significant in the stepwise regression case and insignificant in the multiple regression case and vice versa is also common. However, since stepwise regression is an improvement on multiple regression, the stepwise regression is better in providing explanation to the response. The final model from the stepwise regression process varies with the environmental conditions. We have to determine which of the explanatory variables appear the most in the final model of the stepwise regression process, among all 240 available data sets.

Table 2.8 summarises the results of the stepwise regression analysis for

Table 2.8: SALT overall stepwise regression output from the 240 data sets provided

Data	**All**	**Dis tref**	**No tref**	**With tref**	**TEnv**
time	171	169	115	56	177
ttr	125	131	80	45	149
ti1	64	62	43	21	60
ti2	91	88	56	35	87
windin	66	66	47	19	66
wind30	89	89	61	28	98
humin	119	122	70	49	120
hum30	123	125	77	46	142
dgroc	186	189	128	58	
tref	42			42	
	/240	**/240**	**/157**	**/83**	**/240**

the 240 samples of data sets provided. Note that these data sets are collected, not continuously, meaning not everyday, over a period from March 2005 to April 2007. Also note that the spherical aberration corrector was faulty. Reference temperature in the dome was sometimes (157/240) not provided. It (Table 2.8) indicates how many times each explanatory variable was significant (part of the final model) in the stepwise regression analysis. Note that in this table:

- The first column (All) gives results of the overall stepwise regression on all the 240 data sets, whether the reference temperature was provided or not. The regression analysis takes into account the reference temperature when it is provided.
- The second column (Dis tref) gives results of the overall stepwise regression on all the 240 data sets, whether the reference temperature was provided or not. The regression analysis does not take into account the reference temperature whether it is provided or not.
- The third column (No tref) gives results of the overall stepwise regression only on the 157 data sets where the reference temperature was not provided.

- The fourth column (With tref) gives results of the overall stepwise regression only on the 83 data sets where the reference temperature was provided, and indeed takes into account the reference temperature.
- The fifth column (TEnv) gives results of the overall stepwise regression on all the 240 data sets, whether the reference temperature was provided or not. The regression analysis only takes into account time and environmental data.

Note that considering all the data sets, the main explanations for the figure of merit are time (which suggests computation issues), temperature of truss and humidity, but also change in radius of curvature (which is computed and temperature dependent). Moreover, the rate of significance of reference temperature cannot be fairly assessed since it is provided only 83 times out of 240. We can use alternate approaches: discard the reference temperature, discard reference temperature and change in radius of curvature, split the data sets in two (those with reference temperature and those without). Discarding reference temperature slightly decreases the rate of significance of time, although it still remains a serious concern. It increases the rate of significance of temperature of truss as well as humidity and change in radius of curvature. In brief, the main significant explanatory variables remain the same. Considering only data sets without reference temperature, the main significant explanatory variables remain time, temperature of truss, change in radius of curvature and to some extent, humidity. On the other hand, considering only data sets with reference temperature, the main significant explanatory variables are time, temperature of truss, change in radius of curvature, humidity and reference temperature. However, considering all data sets while discarding reference temperature and change in radius of curvature, which we consider as a better approach, leads to the main significant explanatory variables being time, temperature of truss and humidity. Note that the main concern is time, followed by temperature of truss, and then comes humidity. The occurrence of time as significant suggests improvement in computation. Temperature of truss being significant suggests flaws in engineering design, due to the fact that temperature was a preoccupation in the designing process.

2.2.3 Autocorrelation and Spectral Analysis of SALT Data

Autocorrelation

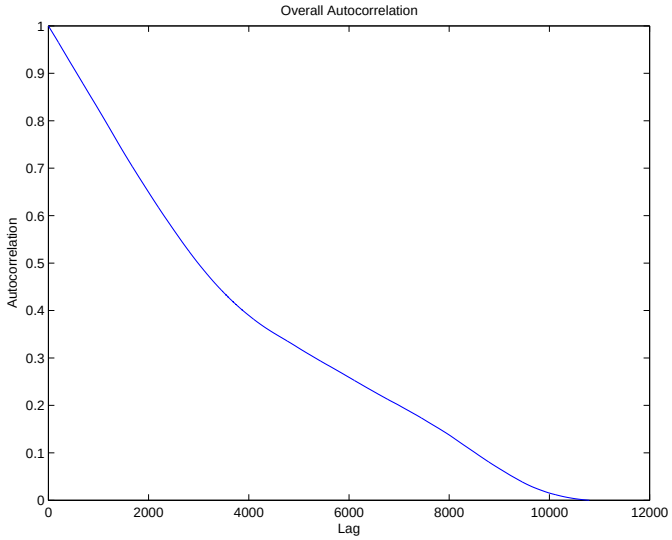

Figure 2.11: Autocorrelation of relative heights (30 July 2010): this indicates a non stationary process

Figure 2.11 illustrates the autocorrelation of the relative heights from an open-loop test conducted on 30 July 2010. It suggests (since it is not close to a straight line) that the stochastic process illustrated by relative heights is not stationary. For the overall assessment, we choose the relative heights at each time to be the RMS (Root Mean Square) of the measurements given by all the sensors. This indicates on average how far the system is from the ideal position (corresponding to the best possible alignment). The lag is measured in seconds and the autocorrelation is a dimensionless output. Note that autocorrelation can also be done for measurements from each sensor, and the results are different for different sensors. The purpose of this on an open-loop test is to extract information, find out if there is a process happening in a given time scale. From inspection of the figure, we can conclude that the autocorrelation decreases almost exponentially with the lag, which doesn't give us much information about stationarity of the process under study. Our next move is to move our inspection from time scale to frequency scale.

Spectral Analysis

Spectral analysis is mostly investigated to tell us in which frequency range most of our information is gathered, whether processes are happening fast or slow, that is, if we have a low pass or high pass process.

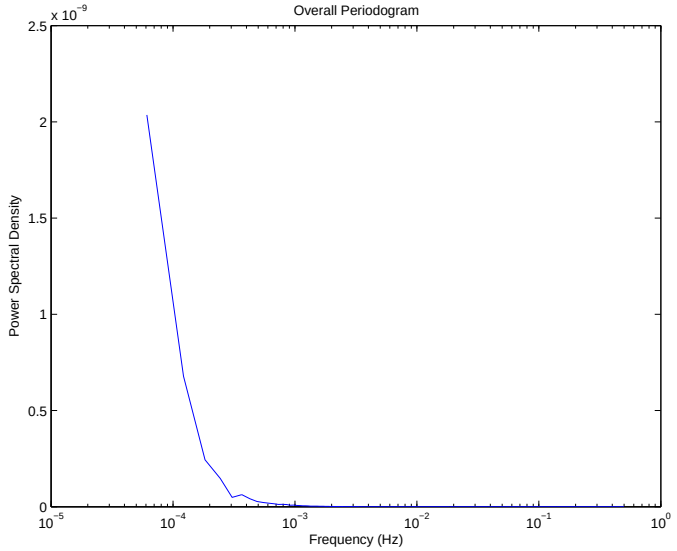

Figure 2.12: Spectrum of relative heights (30 July 2010): this indicates a low pass process

Figure 2.12 illustrates the spectrum of the relative heights from the open-loop test conducted on 30 July 2010. The frequency is given in Hertz (but can also be given in radians per sample) and the spectrum is given in units of power per Hertz (but can also be given in units of power per radian per sample). The frequency $f \in [0, 1]$ and is illustrated in log scale for visibility purposes. The figure shows that the spectrum is small in magnitude and the further we move from 0 in terms of frequency, the smaller the value of the spectrum. This suggests that most (more than 90%) of the power of the spectrum is in the frequency range below 10^{-3}Hz (which corresponds to a time range larger that 20 minutes on average), which indicates a low pass process. Note that the illustration of spectrum can also be done for each sensor, and the results are different for different sensors.

2.3 Conclusion

Available data have been explored for explanation about the behavior of the figure of merit. Note that in the early stages of our data analysis, we were provided data issued from experiments conducted in 11 non consecutive days. Stepwise regression then gave us results as in Table 2.9.

Table 2.9: SALT overall stepwise regression output from the 11 data sets provided

Data	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**Total**
time	1	1		1	1	1	1	1		1	1	**9**
ttr	1	1			1		1			1	1	**6**
ti1												**0**
ti2												**0**
wx	1		1		1					1		**4**
wy	1		1	1		1				1	1	**6**
wz	1	1	1	1	1		1	1	1	1	1	**10**
hum					1			1		1		**3**
dgroc	1	1		1			1			1	1	**6**
tref				1	1	1	1					**4**

In this table, *wx*, *wy* and *wz* stand for wind speeds in the x, y and z directions respectively, and are given in metres per second; *hum* stands for relative humidity, and is given in %. The remaining variables are as given in Table 2.1 (page 17). Note in Table 2.9 that the first column gives the list of explanatory variables in the regression process, the last column gives the rate of significance of each explanatory variable over the 11 experiments, and the columns inbetween are specific to each of the experiments. Each nonempty cell from the columns inbetween indicates that the corresponding explanatory variable was found to be significant in the stepwise regression analysis for that specific experiment. Recall that explanatory variables in this case, and in all types of regression in general, are variables from which we seek explanation to the behavior of the response variable. In this case, the response variable is the figure of merit. Explanatory variables are not to be confused with control variables. In the SALT case, the control variables are the adjustments to be made at the actuators in order to align the mirror

by minimising the relative heights, which is the objective function in the optimal control problem. The optimal control problem and the regression problem are two different problems. In the optimal control problem, the goal is to align the mirror while in the regression problem, the goal is to detect variables that explain the behavior of a specific variable we choose to study – in particular, detect variables that can have some external impact on the control system. Results on explanatory variables are from a process beyond the scope of the control system. Recall that environmental variables cannot be controlled at all. First note from Table 2.9 an unexpected result whereby the most significant explanatory variable (that is, the most significant explanation to the figure of merit) is *wz* – the wind speed in the z direction, which in the first night of the 11 experiments for example varies between 42 and 55 metres per second. This is indeed very unrealistic and requires more analysis and further studies. Overall, the most significant explanatory variables (in order) are *wz*, *time*, *ttr*, *wy* and *dgroc*, as they appear to be significant more than 50% of the time. After a discussion with SALT staff to understand the results of our regression analysis, we were informed that *wx*, *wy*, *wz* and *hum* should actually stand for inside wind speed (*windin* – in m/s), outside wind speed (*wind30* – in m/s), outside humidity (*hum30* – in %), and inside humidity (*humin* – in %), respectively. This clearly shows that data had been wrongly labelled. More data was then provided to us (with new labels) for further analysis. Table 2.8 is the final corrected stepwise regression output.

Tests for multicollinearity have been conducted and they reveal that multicollinearity should be a serious concern. Regression analysis was also conducted and it suggests that multicollinearity is a concern. Autocorrelation and spectral analysis conducted on relative heights from an open-loop test suggests that the process described by the data is not stationary, and is also a low-pass process.

SALT staff are well aware that sensors fail at high humidity. Note that stepwise regression is aimed at measuring a defect in the control algorithm. Stepwise regression applied to a good operating regime shows that at best, humidity has a relatively minor role to play, since its rate of significance comes after that of time and temperature of truss. Also note that the appearance of both temperature and humidity as significant variables in the final model of stepwise regression shows that they are actually independent

significant variables. These two variables (in order, temperature of truss and humidity) are therefore of obvious interest since regression analysis shows failure to control against these variables. But also, it is clear that time itself is an independent variable. Thus, stepwise regression analysis suggests (Table 2.8) that the main reasons for degrading figure of merit are computation (because of time as a significant variable), temperature of truss and to some extent, humidity. The main concern however, is computation, which suggests numerical problems, followed by temperature of truss, which suggests an imperfection in structure design. This result is quite different from the (correct) knowledge of the SALT staff that the capacitive edge sensors work in a particular humidity range.

Temperature is not controlled in the dome, deformation of the truss is a real concern, and therefore it is essential to control the SALT primary mirror continuously. In turn, it is important that the control code be re-examined.

Humidity is not controlled in the dome and is found to be a mildly significant explanation to the degradation of the figure of merit. SALT has independently identified capacitive edge sensors to be very sensitive to humidity and has (as of January 2012) prepared a call for tenders to replace them with inductive devices. Our results indicate that this will contribute to improved figure of merit only if mathematical control is effective.

Considering that time is the main concern in our data analysis, exploration of computation is an important step in resolving the issue of improving the performance of the telescope.

Chapter 3

Control Algorithm

It was established in Chapter 2 that computation is a serious concern in the unsatisfactory performance of SALT. In this chapter we discuss the performance of the control algorithm of SALT and propose a few techniques for improvement of this control algorithm, which can also be helpful for other large segmented telescopes.

3.1 Overview on the Existing Segment Alignment Measurement System (SAMS)

As previously mentioned (Section 1.1.1), SALT has a primary mirror composed of 91 segments disposed on a steel truss with a spherical shape. Between every two neighboring segments, we have two sensors and each segment is controlled by three actuators. This gives 273 actuators and 480 sensors. The sensors measure the relative heights between neighboring segments, and these relative heights are due to the fact that the segments can move independently of one another, and therefore depart from their ideal positions. The main goal is to keep the segments as close as possible to the ideal position all the time. The mirror is equipped with a Primary Mirror Alignment System (PMAS) divided into four (building blocks) subsystems, precisely the Mirror Alignment Control System (MACS), the Center of Curvature Alignment Sensor (CCAS), the Segment Alignment Measurement System (SAMS) and the Segment Positioning System (SPS). We are interested in CCAS and SAMS, due to the fact that relative heights measured by CCAS and computed by SAMS are supposed to match at some level of tolerance, to confirm the efficiency of the control algorithm. The control problem of the SALT mirrors is a large scale problem and therefore involves large matrix manipulations. The most reasonable way to handle this problem is the numerical way.

SALT was built in such a way that among other conditions, the Primary Mirror Alignment System must meet some technical requirements as described in [41]. In brief, the Primary Mirror Alignment System interfaces with the Telescope Control System (TCS), the computer room, the mirror truss, the mirror mounts, the primary mirror array and the CCAS tower.

3.1.1 Geometric Modelisation

It is well known from Geometry [1, 16] that in a 3-dimensional space, the equation of a plane can be written in the form $Lx + My + Nz + K = 0$, where L, M, N and K are constants. In particular, if the plane passes through three known and not aligned points $P_1(x_1, y_1, z_1)$, $P_2(x_2, y_2, z_2)$ and $P_3(x_3, y_3, z_3)$, then it can be established that

$$
\begin{aligned}
L &= y_1(z_3 - z_2) + y_2(z_1 - z_3) + y_3(z_2 - z_1) \\
M &= z_1(x_3 - x_2) + z_2(x_1 - x_3) + z_3(x_2 - x_1) \\
N &= x_1(y_3 - y_2) + x_2(y_1 - y_3) + x_3(y_2 - y_1) \\
K &= x_1(y_2z_3 - y_3z_2) + x_2(y_3z_1 - y_1z_3) + x_3(y_1z_2 - y_2z_1)
\end{aligned}
$$

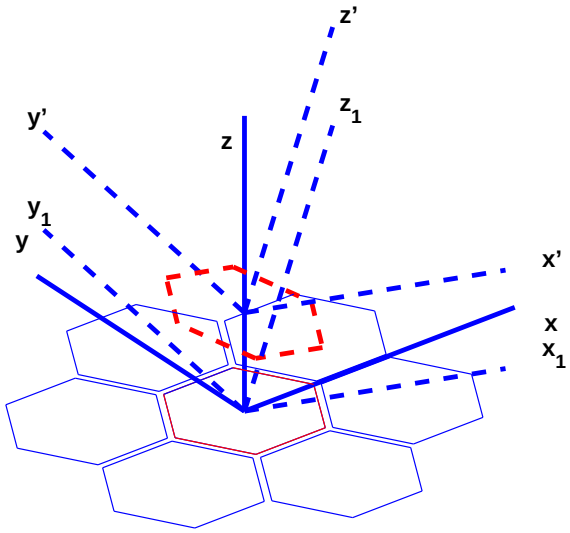

Figure 3.1: Different axes types for the SALT primary mirror

If we consider our x, y and z axes as in Figure 3.1 (in their respective positive directions) where for illustration purposes we only displayed the

central segment and the first ring, then they define the *Telecentric Axes System* as laid out in [38] where we can also find more about the geometry behind the construction of the SALT primary mirror array. If the x, y and z axes are defined with reference to any segment at the ideal position, they define the *Segment Local Axes System (SLAS)* as explained in [30, 31, 41, 42] and illustrated in Figure 3.1 as the x, y and z axes for the central segment. If the axes system is fixed with reference to a segment (and therefore moves as the segment moves) then it defines the *Segment Body Frame (SBF)*, illustrated in Figure 3.1 as the x', y' and z' axes for the central segment. When the segment moves, the angle θ between the x axes of the SLAS and the SBF (the x and x' axes or equivalently the x and x_1 axes in Figure 3.1 for the central segment) is called *tip* and is measured in arcseconds; the angle φ between the y axes of the SLAS and the SBF (the y and y' axes or equivalently the y and y_1 axes in Figure 3.1 for the central segment) is called *tilt* and and is measured in arcseconds; the *piston*, sometimes denoted as p, is the distance covered by the center of the segment in the positive z direction of the SLAS (the distance between the origin of the x, y and z axes and the origin of the x', y' and z' axes in Figure 3.1 for the central segment) and is measured in microns. The *Global Radius of Curvature (GRoC)* is the radius of curvature of the spherical surface forming the closest approximation of the primary mirror array's optical surface.

The segment alignment takes place as follows: during observations, SAMS continuously measures segment movement and MACS calculates corrections and sends commands to SPS to perform the corrections.

Exploring the geometry of the Primary Mirror, the equation of each segment in SLAS at the ideal position is $z = 0$; and after a movement, (assuming this is done in a very small range of actuator displacements) the equation becomes $Lx + My + Nz + K = 0$ and since $N \neq 0$ because of the assumption (leading to the new equation $z = -\frac{L}{N}x - \frac{M}{N}y - \frac{K}{N}$), we can easily find that the tip (θ) and tilt (φ) are given by: $\tan(\varphi) = \frac{L}{N} \simeq \varphi$ and $\tan(\theta) \simeq -\frac{M}{N} \simeq \theta$, as explained in [30, 31]. Indeed, from a simple approximation principle, if we consider the respective angles θ and φ of the rotations around the x and y axes, their matrices are given as follows:

$$R_x = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{pmatrix} \qquad R_y = \begin{pmatrix} \cos\varphi & 0 & \sin\varphi \\ 0 & 1 & 0 \\ -\sin\varphi & 0 & \cos\varphi \end{pmatrix}$$

Their effect on the positive unit vector in the z direction is described as follows:

$$\begin{aligned}\vec{n} = R_y * R_x * \begin{pmatrix}0\\0\\1\end{pmatrix} &= \begin{pmatrix}\cos\varphi & 0 & \sin\varphi\\ 0 & 1 & 0\\ -\sin\varphi & 0 & \cos\varphi\end{pmatrix}\begin{pmatrix}1 & 0 & 0\\ 0 & \cos\theta & -\sin\theta\\ 0 & \sin\theta & \cos\theta\end{pmatrix}\begin{pmatrix}0\\0\\1\end{pmatrix}\\ &= \begin{pmatrix}\cos\varphi & 0 & \sin\varphi\\ 0 & 1 & 0\\ -\sin\varphi & 0 & \cos\varphi\end{pmatrix}\begin{pmatrix}0\\-\sin\theta\\\cos\theta\end{pmatrix}\\ &= \begin{pmatrix}\sin\varphi\cos\theta\\-\sin\theta\\\cos\varphi\cos\theta\end{pmatrix}\end{aligned}$$

This is a unit vector and is normal to the plane with equation

$$Lx + My + Nz + K = 0$$

and is therefore to be identified to the vector

$$\begin{pmatrix}\frac{L}{\sqrt{L^2+M^2+N^2}}\\ \frac{M}{\sqrt{L^2+M^2+N^2}}\\ \frac{N}{\sqrt{L^2+M^2+N^2}}\end{pmatrix}$$

from which we obtain

$$\begin{aligned}\sin\varphi\cos\theta &= \frac{L}{\sqrt{L^2+M^2+N^2}}\\ -\sin\theta &= \frac{M}{\sqrt{L^2+M^2+N^2}}\\ \cos\varphi\cos\theta &= \frac{N}{\sqrt{L^2+M^2+N^2}}\end{aligned}$$

and since θ and φ are very small, the approximation rule gives

$$\begin{aligned}\tan\varphi &= \frac{L}{N} \simeq \varphi\\ \tan\theta &= -\frac{M}{N\cos\varphi} \simeq -\frac{M}{N} \simeq \theta.\end{aligned}$$

Also, the *piston* is given by $p = -\frac{K}{N}$. Therefore

$$z = -x\tan(\varphi) + y\tan(\theta) + p.$$

The assumption of very small range of actuator displacements also gives a one-to-one correspondence between actuator displacements and segment's displacement (piston/tip/tilt) for each segment (see [25, 31, 41]), summarised as follows:

$$\begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & R \\ 1 & -R\frac{\sqrt{3}}{2} & -\frac{R}{2} \\ 1 & R\frac{\sqrt{3}}{2} & -\frac{R}{2} \end{pmatrix} \begin{pmatrix} \text{piston} \\ \text{tip} \\ \text{tilt} \end{pmatrix} \tag{3.1}$$

which is equivalent to:

$$\begin{pmatrix} \text{piston} \\ \text{tip} \\ \text{tilt} \end{pmatrix} = \frac{1}{3R} \begin{pmatrix} R & R & R \\ 0 & -\sqrt{3} & \sqrt{3} \\ 2 & -1 & -1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} \tag{3.2}$$

where R, piston, as well as the actuator displacements a_1, a_2 and a_3 are in metres, tip and tilt in radians (and can be converted to arcseconds, with π rad $= 180^\circ$ and $1^\circ = 3600$ arc-sec). This helps in simplifying the calculation of relative heights between neighboring segments. These relative heights are measured by edge sensors. Each edge sensor has two plates: one active, the emitter plate, and one passive, the receiver plate. The relative heights are defined for each sensor as the difference of heights between the emitter plate and the corresponding receiver plate on the neighboring segment. So for the whole array, this can be written as:

$$S_l = \sum_{m=1}^{273} a_{lm} Z_m \quad \text{or} \quad S = AZ \tag{3.3}$$

where S_l is the height as read from sensor l and Z_m is the displacement of actuator m. So A is a (sparse) 480×273 matrix, S is a column vector of length 480 and Z is a column vector of length 273. Using the heights obtained from the sensor readings, the problem is to find the tip/tilt/piston for each segment. This is done by using the Least Squares method [10, 12, 21, 36] to determine the actuator displacements from relative heights, and the one-to-one correspondence (3.2) above to find the tip/tilt/piston we required.

The rank of A is 269 and we have 273 unknowns. Hence finding z from s using the relation $s = Az$ does not have a unique solution. This is illustrated in introductory examples on least squares problems (page 58). Therefore four constraints are needed to bring the rank of A to 273. This is achieved by locking one segment (the central segment for example) and

locking the piston of another segment (on the outer ring for example) to be zero; another way is to lock the pistons of four segments to be zero. The SALT initial algorithm has the option of choosing between the two alternatives. This changes the dimensions of the A matrix (from 480 to 483 or 484 according to the number of additional constraints). On the other hand, it is also required to deal only with valid sensors (those which are working properly and are not situated around uncontrolled segments) and controlled segments (those which have at least three of twelve sensors (see Figure 1.2 for illustration of a few segments with sensors and actuators) to control their position). This reduces A to a $L \times M$ matrix, where $L \leq 480$ and $M \leq 273$. Note that considering the option to block pistons of some segments is based on the assumption that the truss is spherical and does not deform. Hence a better option is to consider constraints not collinear to the information we have, that is, not collinear to pistons and actuator displacements.

3.2 On the Numerical Methods for SALT

The numerical issues to handle for linear control problems demand accurate solutions of linear algebraic systems. This can be done manually for small size systems. Large scale systems are necessarily handled numerically.

3.2.1 Preliminary Theory

Formulation of the Least Squares Problem

We consider the linear system $Ax = b$ where A is an $m \times n$ matrix and b is a column vector of length m. The problem is to find a column vector x of length n that minimises $\|Ax - b\|$, which is the same as minimising $\|Ax - b\|^2$ (the norm is the $\|\cdot\|_2$ norm or the Euclidean norm). There are many approaches to solve the least squares problem. It can be solved analytically for reasonable sizes of the matrix A. However, for large scale problems, numerical approaches are more indicated for solving least squares problems.

Overview We consider $A \in \mathbb{R}^{m \times n}$, the set of real $m \times n$ matrices, and $\mathcal{R}(A)$ the range of A, that is, $\mathcal{R}(A) = \{Ax : x \in \mathbb{R}^n\}$.

Remark 3.1. The following results are well known from the literature [10]

1. $\text{rank}(A) \leq \min(m, n)$
2. If $m \geq n$, then $\text{rank}(A) = n \iff \text{rank}\left(A^T A\right) = n$, that is, $A^T A$ is nonsingular
3. If $m \geq n$ and $\text{rank}(A) = n$, then the unique vector x^* that minimises $\|Ax - b\|^2$ is $x^* = \left(A^T A\right)^{-1} A^T b$
4. If $m \geq n$ and $\text{rank}(A) = n$, if $h \in \mathcal{R}(A)$ and $h - b$ is orthogonal to $\mathcal{R}(A)$, then $h = Ax^* = A\left(A^T A\right)^{-1} A^T b$
5. If $m \leq n$ and $\text{rank}(A) = m$, then the unique solution to $Ax = b$ that minimises $\|x\|$ is $x^* = A^T \left(AA^T\right)^{-1} b$

The following results can also be found in the literature.

Lemma 3.2 (See [10]). *Given $A \in \mathbb{R}^{m \times n}$ with rank$(A) = r$, there exist two matrices $B \in \mathbb{R}^{m \times r}$ and $C \in \mathbb{R}^{r \times n}$ such that $A = BC$ with rank$(B) =$ rank$(C) = r$.*

Definition 3.3 (See [10]). Given $A \in \mathbb{R}^{m \times n}$, a matrix A^+ is a *pseudo inverse* of A if:

- $AA^+A = A$
- There exist $U \in \mathbb{R}^{n \times n}$ and $V \in \mathbb{R}^{m \times m}$ such that $A^+ = UA^T = A^T V$

Theorem 3.4 (See [10]). *If the pseudo inverse of A exists, it is unique.*

Theorem 3.5 (See [10]). *Let $A \in \mathbb{R}^{m \times n}$, with full rank factorization $A = BC$ where rank$(A) =$ rank$(B) =$ rank$(C) = r$; then $A^+ = C^+ B^+$ where $B^+ = \left(B^T B\right)^{-1} B^T$ and $C^+ = C^T \left(CC^T\right)^{-1}$*

Remark 3.6. The following is valid for full rank factorization:

- If $A = BC$ is a full rank factorization of A, then another (equivalent) expression for A^+ is $A^+ = C^T \left(B^T AC^T\right)^{-1} B^T$
- The equality $A^+ = C^+ B^+$ from Theorem 3.5 does not necessarily hold if $A = BC$ is not a full rank factorization of A

Theorem 3.7 (See [10]). *Let $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$; then $x^* = A^+ b$ minimises $\|Ax - b\|$ on $\mathbb{R}^n$. Furthermore, amongst all vectors in $\mathbb{R}^n$ that minimise $\|Ax - b\|$, $x^* = A^+ b$ is the unique vector with minimum norm.*

Remark 3.8. The following can be verified [10]:

1. $\left(A^T\right)^+ = \left(A^+\right)^T$
2. $\left(A^+\right)^+ = A$
3. A^+ is the pseudo inverse of A if and only if
   - $AA^+A = A$
   - $A^+AA^+ = A^+$
   - $\left(AA^+\right)^T = AA^+$
   - $\left(A^+A\right)^T = A^+A$

The Least Squares Problem as an Optimisation Problem

The problem under consideration is

$$\min_x \|Ax - b\| \tag{3.4}$$

where A is an $m \times n$ matrix and b is a column vector of length m. Problem (3.4) is equivalent to minimising $\|Ax-b\|^2 = (Ax-b)^T(Ax-b)$. This arises in the SALT control process when for example actuator positions have to be determined from relative heights. There are three main methods relevant to solving (3.4). These will be explored after a few introductory examples.

Introductory Examples Consider the following vectors and matrices:

$$A_1 = \begin{pmatrix} 3 & 2 \\ 1 & 3 \\ 4 & 5 \end{pmatrix} \quad b_1 = \begin{pmatrix} 4 \\ -1 \\ 3 \end{pmatrix} \quad bb_1 = \begin{pmatrix} 5 \\ 0 \\ 2 \end{pmatrix}$$

$$A_2 = \begin{pmatrix} 3 & 2 & 1 \\ 1 & 3 & 2 \\ 4 & 5 & 3 \end{pmatrix} \quad b_2 = \begin{pmatrix} 6 \\ 5 \\ 11 \end{pmatrix} \quad bb_2 = \begin{pmatrix} 5 \\ 4 \\ 12 \end{pmatrix}$$

Note that A_1 is a 3×2 matrix with rank 2 and A_2 is a 3×3 matrix with rank 2. That is, A_1 is a full rank matrix and A_2 is a rank deficient matrix.

1. Determine $x \in \mathbb{R}^2$ that solves $\min_x \|A_1x - b_1\|$

Sketch of Solution A simple calculation gives the unique solution $x_1 = \begin{pmatrix} 2 \\ -1 \end{pmatrix}$ and this can easily be obtained by solving $A_1 x = b_1$. It can indeed be established that $\|A_1 x_1 - b_1\| = 0$

2. Determine $x \in \mathbb{R}^2$ that solves $\min_x \|A_1 x - b b_1\|$

   Sketch of Solution A simple calculation shows that the range of A_1 is the plane $a + b - c = 0$ in $\mathbb{R}^3$ and obviously bb_1 is not in that plane (Note that b_1 given above is in that plane). However, minimising $\|A_1 x - b b_1\|$ can be done by finding the point b_1' in the range of A_1, that is closest to bb_1 and then find the inverse image of b_1'; if we have many of them, we choose the one with minimum norm. A normal vector to the range of A_1 is $\vec{n} = \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}$. We just need to find a $\lambda \in \mathbb{R}$ such that $bb_1 + \lambda \vec{n}$ is in the range of A_1. In fact, $bb_1 + \lambda \vec{n} = b_1'$. A simple computation gives $\lambda = -1$ and therefore $b_1' = \begin{pmatrix} 4 \\ -1 \\ 3 \end{pmatrix}$. Solving $A_1 x = b_1'$ gives $x_1 = \begin{pmatrix} 2 \\ -1 \end{pmatrix}$, which is the solution to our least squares problem $\min_x \|A_1 x - b b_1\|$. And actually, $\|A_1 x_1 - b b_1\| = |\lambda| \|\vec{n}\| = \sqrt{3}$.

3. Determine $x \in \mathbb{R}^3$ that solves $\min_x \|A_2 x - b_2\|$

   Sketch of Solution A simple calculation shows that the null space of A_2 is spanned by vector $\vec{u} = \begin{pmatrix} 1 \\ -5 \\ 7 \end{pmatrix}$ and that a particular solution to $A_2 x = b_2$ is $x_{2p} = \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}$. The solution to $A_2 x = b_2$ is $x_2 = x_{2p} + \alpha \vec{u} = \begin{pmatrix} 1 + \alpha \\ 2 - 5\alpha \\ -1 + 7\alpha \end{pmatrix}$, with $\alpha \in \mathbb{R}$. For each of these x_2, we have $\|A_2 x_2 - b_2\| = 0$. Moreover, $\|x_2\| = \sqrt{75\alpha^2 - 32\alpha + 6}$ and this is minimised when $\alpha = \frac{16}{75}$ and hence $\|x_2\| = \sqrt{\frac{194}{75}} \simeq 1.6083$.

4. Determine $x \in \mathbb{R}^3$ that solves $\min_x \|A_2 x - bb_2\|$

Sketch of Solution A simple calculation shows that the range of A_2 is the plane $a + b - c = 0$ in $\mathbb{R}^3$, and obviously bb_2 is not in that plane (Note that b_2 given above is in that plane). Therefore there is no vector $x \in \mathbb{R}^3$ such that $A_2 x = bb_2$. However, minimising $\|A_2 x - bb_2\|$ can be done by finding the point b_2' in the range of A_2, that is closest to bb_2 and then find the inverse image of b_2'; if we have many of them, we choose the one with minimum norm. A normal vector to the range of A_2 is $\vec{n} = \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}$. We just need to find a $\lambda \in \mathbb{R}$ such that $bb_2 + \lambda\vec{n}$ is in the range of A_2. In fact, $bb_2 + \lambda\vec{n} = b_2'$. A simple computation gives $\lambda = 1$ and therefore $b_2' = \begin{pmatrix} 6 \\ 5 \\ 11 \end{pmatrix}$. Solving $A_2 x = b_2'$ gives $x_2 = \begin{pmatrix} 1+\alpha \\ 2-5\alpha \\ -1+7\alpha \end{pmatrix}$, with $\alpha \in \mathbb{R}$. For each of these x_2, we have $\|A_2 x_2 - bb_2\| = |\lambda|\|\vec{n}\| = \sqrt{3}$. Moreover, $\|x_2\| = \sqrt{75\alpha^2 - 32\alpha + 6}$ and this is minimised when $\alpha = \frac{16}{75}$ and hence $\|x_2\| = \sqrt{\frac{194}{75}} \simeq 1.6083$.

Note that solving $Ax = b$ or $\min_x \|Ax - b\|$ has at most one solution when A is a full rank matrix, and infinitely many solutions when A is rank deficient. As a direct application to the SALT case, since the SALT actuator-to-heights matrix is rank deficient, this implies that for a given vector of measured relative heights s, there are infinitely many sets of actuator displacements z such that $\|s - Az\|$ is minimised. Now we explore the three main methods relevant to solving (3.4).

The Normal Equations Approach [12, 28] This method is used on SALT. We consider $f(x) = (Ax - b)^T(Ax - b)$. Then we have

$$\begin{aligned} f(x) &= (Ax - b)^T(Ax - b) \\ &= x^T A^T Ax - x^T A^T b - b^T Ax + b^T b \\ &= x^T A^T Ax - 2b^T Ax + b^T b \end{aligned}$$

The gradient of f is given by $\nabla_x(f) = 2A^TAx - 2A^Tb$. The equation $\nabla_x(f) = 0$ or equivalently

$$A^TAx - A^Tb = 0 \tag{3.5}$$

is the set of *normal equations* associated to Problem (3.4).

Provided the matrix A^TA is nonsingular (which by item 2 of Remark 3.1 means that A has full rank), the solution to the normal equations is $x^* = (A^TA)^{-1}A^Tb$. Moreover, x^* is indeed the minimiser of f because the *Hessian* of f is A^TA which is positive definite. This confirms item 3 of Remark 3.1.

Example We reconsider the introductory examples (See page 58).

1. We have $A_1^TA_1 = \begin{pmatrix} 26 & 29 \\ 29 & 26 \end{pmatrix}$ which is a nonsingular matrix, and hence, if we let $C_1 = \left(A_1^TA_1\right)^{-1}A_1^T$, a simple calculation gives $C_1 = \frac{1}{21}\begin{pmatrix} 8 & -7 & 1 \\ -5 & 7 & 2 \end{pmatrix}$, and for the first two questions, we respectively obtain by straightforward substitution

$$C_1b_1 = \left(A_1^TA_1\right)^{-1}A_1^Tb_1 = \frac{1}{21}\begin{pmatrix} 8 & -7 & 1 \\ -5 & 7 & 2 \end{pmatrix}\begin{pmatrix} 4 \\ -1 \\ 3 \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \end{pmatrix}$$

$$C_1bb_1 = \left(A_1^TA_1\right)^{-1}A_1^Tbb_1 = \frac{1}{21}\begin{pmatrix} 8 & -7 & 1 \\ -5 & 7 & 2 \end{pmatrix}\begin{pmatrix} 5 \\ 0 \\ 2 \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \end{pmatrix}$$

2. We have $A_2^TA_2 = \begin{pmatrix} 26 & 29 & 17 \\ 29 & 38 & 23 \\ 17 & 23 & 14 \end{pmatrix}$ which is a singular matrix, and hence, it is impossible to determine $C_2 = \left(A_2^TA_2\right)^{-1}A_2^T$, therefore the normal equations approach is not applicable for the last two questions of the introductory examples.

Remark 3.9. The normal equations approach is fast, but numerically unstable and not recommended for large scale problems and rank deficient problems.

The QR Approach [12, 28] This involves simplifying the problem under consideration (Problem (3.4)) by writing the matrix A as a product of an *orthogonal* matrix Q and an *upper triangular* matrix R.

Theorem 3.10 (See [12])**.** *We consider a full rank $m \times n$ matrix A with $m \geq n$ (that is rank$(A) = n$). Then there exists a unique $m \times n$ orthogonal matrix Q and a unique $n \times n$ upper triangular matrix R with positive diagonals $r_{ii} > 0$ such that $A = QR$.*

Remark 3.11. The QR factorization as given in Theorem 3.10 is not unique if A does not have full rank. Moreover, R is singular when A is rank deficient.

Under the assumption that A has full rank, from the QR factorization of A, we have

$$\begin{aligned} x^* &= (A^T A)^{-1} A^T b \\ &= (R^T Q^T Q R)^{-1} R^T Q^T b \\ &= (R^T R)^{-1} R^T Q^T b \\ &= R^{-1} (R^T)^{-1} R^T Q^T b \\ &= R^{-1} Q^T b \end{aligned}$$

This approach is numerically more stable than the normal equations approach, is computationally more expensive and also applies to large scale problems.

Example We reconsider the introductory examples (See page 58).

1. We have $A_1 = Q_1 R_1 = Q_{11} R_{11}$ with

$$Q_1 = \begin{pmatrix} -\frac{3}{\sqrt{26}} & -\frac{5}{\sqrt{78}} & -\frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{26}} & -\frac{7}{\sqrt{78}} & -\frac{1}{\sqrt{3}} \\ -\frac{4}{\sqrt{26}} & \frac{2}{\sqrt{78}} & \frac{1}{\sqrt{3}} \end{pmatrix} \quad R_1 = \begin{pmatrix} -\sqrt{26} & -\frac{29}{\sqrt{26}} \\ 0 & -7\frac{\sqrt{3}}{\sqrt{26}} \\ 0 & 0 \end{pmatrix}$$

$$Q_{11} = \begin{pmatrix} -\frac{3}{\sqrt{26}} & -\frac{5}{\sqrt{78}} \\ -\frac{1}{\sqrt{26}} & -\frac{7}{\sqrt{78}} \\ -\frac{4}{\sqrt{26}} & \frac{2}{\sqrt{78}} \end{pmatrix} \quad R_{11} = \begin{pmatrix} -\sqrt{26} & -\frac{29}{\sqrt{26}} \\ 0 & -7\frac{\sqrt{3}}{\sqrt{26}} \end{pmatrix}$$

   The QR formula does not apply to Q_1 and R_1 since R_1 is not square. However, a straightforward application of the QR formula on Q_{11} and

R_{11} for the first two questions of the introductory examples gives

$$R_{11}^{-1}Q_{11}^T b_1 = \begin{pmatrix} -\frac{1}{\sqrt{26}} & \frac{29}{7\sqrt{78}} \\ 0 & -\frac{\sqrt{26}}{7\sqrt{3}} \end{pmatrix} \begin{pmatrix} -\frac{3}{\sqrt{26}} & -\frac{1}{\sqrt{26}} & -\frac{4}{\sqrt{26}} \\ -\frac{5}{\sqrt{78}} & -\frac{7}{\sqrt{78}} & \frac{2}{\sqrt{78}} \end{pmatrix} \begin{pmatrix} 4 \\ -1 \\ 3 \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \end{pmatrix}$$

$$R_{11}^{-1}Q_{11}^T bb_1 = \begin{pmatrix} -\frac{1}{\sqrt{26}} & \frac{29}{7\sqrt{78}} \\ 0 & -\frac{\sqrt{26}}{7\sqrt{3}} \end{pmatrix} \begin{pmatrix} -\frac{3}{\sqrt{26}} & -\frac{1}{\sqrt{26}} & -\frac{4}{\sqrt{26}} \\ -\frac{5}{\sqrt{78}} & -\frac{7}{\sqrt{78}} & \frac{2}{\sqrt{78}} \end{pmatrix} \begin{pmatrix} 5 \\ 0 \\ 2 \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \end{pmatrix}$$

2. We have $A_2 = Q_2 R_2$ with

$$Q_2 = \begin{pmatrix} -\frac{3}{\sqrt{26}} & -\frac{5}{\sqrt{78}} & -\frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{26}} & -\frac{7}{\sqrt{78}} & -\frac{1}{\sqrt{3}} \\ -\frac{4}{\sqrt{26}} & \frac{2}{\sqrt{78}} & \frac{1}{\sqrt{3}} \end{pmatrix} \quad R_2 = \begin{pmatrix} -\sqrt{26} & -\frac{29}{\sqrt{26}} & -\frac{17}{\sqrt{26}} \\ 0 & -7\frac{\sqrt{3}}{\sqrt{26}} & -5\frac{\sqrt{3}}{\sqrt{26}} \\ 0 & 0 & 0 \end{pmatrix}$$

The QR formula does not apply to Q_2 and R_2 since R_2 is square but singular. Hence the QR approach does not apply to the last two questions of the introductory examples.

Remark 3.12. The QR approach is about twice as expensive (computationally) as the normal equations approach. It is numerically more stable and can also be applied to large scale problems, but not to rank deficient problems. The SALT system is rank deficient and mirror segments must be fixed via constraints in order to change the system from a rank deficient system to a full rank system.

The Singular Value Decomposition (SVD) Approach [12, 28] This involves simplifying the problem under consideration (Problem (3.4)) by writing the matrix A as a product of two *orthogonal* matrices U and V, and a *diagonal* matrix Σ with nonnegative elements on the main diagonal.

Theorem 3.13 (See [12, 28])**.** *We consider an arbitrary $m \times n$ matrix A with $m \geq n$. Then there exist a unitary $m \times m$ matrix U, a unitary $n \times n$ matrix V and a diagonal $m \times n$ matrix Σ such that $A = U\Sigma V^T$, where*

$$\Sigma = \begin{pmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n \\ 0 & \cdots & 0 \end{pmatrix}$$

with $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$.

Definition 3.14. Let $A = U\Sigma V^T$ be the SVD (Singular Value Decomposition) of A as in Theorem 3.13.

- The columns $u_1, \cdots, u_m$ of U are called *left singular vectors* of A.
- The columns $v_1, \cdots, v_n$ of V are called *right singular vectors* of A.
- The scalars σ_i on the diagonal of Σ are called *singular values* of A.

Remark 3.15. Let $A = U\Sigma V^T$ be the SVD of the $m \times n$ matrix A, with $m \geq n$. The following results can be found in the literature [12]:

1. If A is symmetric, with eigenvalues λ_i and orthogonal eigenvectors u_i (that is, $A = U\Lambda U^T$ is an eigendecomposition of A with the λ_i on the diagonal of Λ), then $A = U\Sigma V^T$ is an SVD of A, where $\sigma_i = |\lambda_i|$ and $v_i = \text{sign}(\lambda_i)u_i$, assuming $\text{sign}(0) = 1$.
2. The eigenvalues of the symmetric matrix A^TA are σ_i^2 and the eigenvectors of A^TA are the right singular vectors v_i of A.
3. The eigenvalues of the symmetric matrix AA^T are σ_i^2 and $m-n$ zeros; and the eigenvectors of AA^T are the left singular vectors u_i of A corresponding to nonzero eigenvalues. One can take any $m-n$ orthogonal vectors as eigenvectors for the eigenvalue 0.
4. If A has full rank, the solution of Problem (3.4) is $x = V\Sigma^{-1}U^Tb$.
5. $\|A\|_2 = \sigma_1$. If A is square and nonsingular, then $\|A^{-1}\|_2^{-1} = \sigma_n$ and the condition number of A is $\kappa(A) = \|A\|_2 \cdot \|A^{-1}\|_2 = \frac{\sigma_1}{\sigma_n}$.
6. Suppose $\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_n = 0$. Then the rank of A is r. The *null space* of A is the space spanned by columns $r+1$ to n of V, that is: $\mathcal{N}(A) = \left\{ \sum_{i=r+1}^{n} \alpha_i v_i \text{ with } \alpha_i \in \mathbb{R} \right\}$. The *range space* of A is the space spanned by columns 1 to r of U, that is, $\mathcal{R}(A) = \left\{ \sum_{i=1}^{r} \alpha_i u_i \text{ with } \alpha_i \in \mathbb{R} \right\}$.
7. Let $A = U\Sigma V^T$ be the SVD of A and let $U = \begin{bmatrix} u_1 & u_2 & \cdots & u_m \end{bmatrix}$ and $V = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix}$. Then $A = \sum_{i=1}^{n} \sigma_i u_i v_i^T$ which is a sum of rank one matrices. Then a matrix of rank $k < n$ closest to A is $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$, and $\|A - A_k\|_2 = \sigma_{k+1}$. This can also be written

as $A_k = U\Sigma_k V^T$ where Σ_k is obtained from Σ by replacing all σ_i for $i > k$ with 0.

Proposition 3.16 (See [12]). *Let A be an $m \times n$ matrix with $m \geq n$ and $rank(A) = r < n$. Then there is an $n - r$ dimensional set of vectors x that minimise $\|Ax - b\|$.*

Proposition 3.17 (See [12]). *Let $\sigma_{\min}$ be the smallest singular value of A and assume $\sigma_{\min} > 0$. Then*

1. *If x minimises $\|Ax - b\|_2$, then $\|x\|_2 \geq \frac{|u_n^T b|}{\sigma_{\min}}$, where u_n is the column of U corresponding to $\sigma_{\min}$ in the SVD of A.*

2. *Changing b to $b + \delta b$ can change x to $x + \delta x$ where $\|\delta x\|_2$ is as large as $\frac{\|\delta b\|_2}{\sigma_{\min}}$.*

In other words, if A is nearly rank deficient, then the solution x is ill-conditioned and possibly very large.

Proposition 3.18 (See [12]). *Let A be an $m \times n$ matrix with rank $r < n$ where $m \geq n$, and $A = U\Sigma V^T$ an SVD of A. This SVD can be written as*

$$A = \begin{bmatrix} U_c & U_u \end{bmatrix} \begin{bmatrix} \Sigma_c & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_c^T \\ V_u^T \end{bmatrix} = U_c \Sigma_c V_c^T$$

where Σ_c is the top left (diagonal and nonsingular) $r \times r$ sub-matrix of Σ, U_c is the left $m \times r$ sub-matrix of U, V_c is the left $n \times r$ sub-matrix of V, U_u and V_u respectively complete the bases of the spaces spanned by U and V.

Let $\sigma = \sigma_{\min}(\Sigma_c)$ be the smallest nonzero singular value of A. Then

1. *All solutions x can be written $x = V_c \Sigma_c^{-1} U_c^T b + V_u z$ where z is an arbitrary vector (set of coefficients, each corresponding to one column vector of V_u).*

2. *The solution x has minimal norm $\|x\|_2$ precisely when $z = 0$, in which case $x = V_c \Sigma_c^{-1} U_c^T b$ and $\|x\|_2 \leq \frac{\|b\|_2}{\sigma}$.*

3. *changing b to $b + \delta b$ can change the minimal norm solution x by at most $\frac{\|\delta b\|_2}{\sigma}$.*

In other words, the norm and condition number of the unique minimal norm solution x depend on the smallest nonzero singular value of A.

Remark 3.19. The equality $A = U_c\Sigma_c V_c^T$ in Proposition 3.18 is called *compact SVD* of A.

Under the assumption that A has full rank, and from the compact SVD of A, we have:

$$\begin{aligned} x^* &= (A^TA)^{-1}A^Tb \\ &= \left((U_c\Sigma_cV_c^T)^TU_c\Sigma_cV_c^T\right)^{-1}(U_c\Sigma_cV_c^T)^Tb \\ &= (V_c\Sigma_cU_c^TU_c\Sigma_cV_c^T)^{-1}(V_c\Sigma_cU_c^T)b \\ &= V_c\Sigma_c^{-2}V_c^TV_c\Sigma_cU_c^Tb \\ &= V_c\Sigma_c^{-1}U_c^Tb \end{aligned}$$

Definition 3.20. We consider an $m \times n$ matrix A with rank $r \leq n$ where $m \geq n$. Let $A = U\Sigma V^T = U_c\Sigma_cV_c^T$ be the SVD and compact SVD of A. Then the pseudo inverse of A is $A^+ = V_c\Sigma_c^{-1}U_c^T$, which can also be written $A^+ = V\Sigma^+U^T$ where Σ^+ is the $n \times m$ matrix

$$\Sigma^+ = \begin{bmatrix} \Sigma_c^{-1} & 0 \\ 0 & 0 \end{bmatrix}$$

So the solution of Problem (3.4) is $x = A^+b$, and when A is rank deficient, x has minimum norm.

Example We reconsider the introductory examples (See page 58).

1. The singular values of A_1 are $\sigma_1 = \sqrt{32+\sqrt{877}}$ and $\sigma_2 = \sqrt{32-\sqrt{877}}$. These are easily obtained as square roots of eigenvalues of $A_1^TA_1$. The right singular vectors v_i of A_1 are obtained as the eigenvectors of $A_1^TA_1$ corresponding to the eigenvalues $\lambda_1 = 32+\sqrt{877}$ and $\lambda_2 = 32-\sqrt{877}$. These vectors are then normailsed to unity. The left singular vectors u_i of A_1 are obtained as the eigenvectors of $A_1A_1^T$ corresponding to the eigenvalues $\lambda_1 = 32+\sqrt{877}$ and $\lambda_2 = 32-\sqrt{877}$ and $\lambda_3 = 0$. These vectors are then normailsed to unity. Before normalisation, we obtain from calculation that

$$u_i = \begin{pmatrix} x \\ y \end{pmatrix} \quad \text{with} \quad y = \frac{\lambda_i - 26}{29}x$$

where the λ_i are the two eigenvalues of $A_1^TA_1$ given above. Similarly,

we obtain from calculation that

$$v_i = \begin{pmatrix} x \\ y \\ z \end{pmatrix} \quad \text{where} \quad y = \frac{19\lambda_i - 49}{22\lambda_i - 49}x$$

$$z = \frac{1}{22}\left[\lambda_i - 13 - 9\frac{19\lambda_i - 49}{22\lambda_i - 49}\right]x$$

where the λ_i are the three eigenvalues of $A_1 A_1^T$ given above. In brief, $A_1 = U_1\Sigma_1 V_1 = U_{11}\Sigma_{11}V_{11}$ where

$$U_1 = \begin{bmatrix} u_1 & u_2 & u_3 \end{bmatrix}, \quad \Sigma_1 = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \\ 0 & 0 \end{bmatrix}, \quad V_1 = \begin{bmatrix} v_1 & v_2 \end{bmatrix}$$

$$U_{11} = \begin{bmatrix} u_1 & u_2 \end{bmatrix}, \quad \Sigma_{11} = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}, \quad V_{11} = V_1$$

After calculation, we obtain

$$V_1\Sigma_1^+ U_1^T = V_{11}\Sigma_{11}^{-1}U_{11}^T = \frac{1}{21}\begin{pmatrix} 8 & -7 & 1 \\ -5 & 7 & 2 \end{pmatrix} = C_1$$

where $\Sigma_1^+ = \begin{bmatrix} \frac{1}{\sigma_1} & 0 & 0 \\ 0 & \frac{1}{\sigma_2} & 0 \end{bmatrix}$. Therefore the solution to the first two questions of the introductory examples is $x_1 = \begin{pmatrix} 2 \\ -1 \end{pmatrix}$ as in the previous two approaches.

2. Following a similar procedure, we have $A_2 = U_2\Sigma_2 V_2^T = U_{22}\Sigma_{22}V_{22}^T$ where

$$U_2 = \begin{pmatrix} -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{3}} \\ -\frac{2}{\sqrt{6}} & 0 & \frac{1}{\sqrt{3}} \end{pmatrix}, \quad U_{22} = \begin{pmatrix} -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{2}} \\ -\frac{2}{\sqrt{6}} & 0 \end{pmatrix}$$

$$\Sigma_2 = \begin{pmatrix} 5\sqrt{3} & 0 & 0 \\ 0 & \sqrt{3} & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \Sigma_{22} = \begin{pmatrix} 5\sqrt{3} & 0 \\ 0 & \sqrt{3} \end{pmatrix}$$

$$V_2 = \begin{pmatrix} -\frac{4}{5\sqrt{2}} & \frac{2}{\sqrt{6}} & -\frac{1}{5\sqrt{3}} \\ -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ -\frac{3}{5\sqrt{2}} & -\frac{1}{\sqrt{6}} & -\frac{7}{5\sqrt{3}} \end{pmatrix}, \quad V_{22} = \begin{pmatrix} -\frac{4}{5\sqrt{2}} & \frac{2}{\sqrt{6}} \\ -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} \\ -\frac{3}{5\sqrt{2}} & -\frac{1}{\sqrt{6}} \end{pmatrix}$$

After calculation, we obtain

$$V_2\Sigma_2^+U_2^T = V_{22}\Sigma_{22}^{-1}U_{22}^T = \frac{1}{75}\begin{pmatrix} 27 & -23 & 4 \\ -10 & 15 & 5 \\ -11 & 14 & 3 \end{pmatrix}$$

where $\Sigma_2^+ = \begin{bmatrix} \frac{1}{5\sqrt{3}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{3}} & 0 \\ 0 & 0 & 0 \end{bmatrix}$. A straightforward application of the SVD formula on U_2, Σ_2 and V_2 (or equivalently on U_{22}, Σ_{22} and V_{22}) for the last two questions of the introductory examples gives

$$V_2\Sigma_2^+U_2^Tb_2 = V_{22}\Sigma_{22}^{-1}U_{22}^Tb_2 = \frac{1}{75}\begin{pmatrix} 27 & -23 & 4 \\ -10 & 15 & 5 \\ -11 & 14 & 3 \end{pmatrix}\begin{pmatrix} 6 \\ 5 \\ 11 \end{pmatrix} = \begin{pmatrix} \frac{91}{75} \\ \frac{70}{75} \\ \frac{37}{75} \end{pmatrix}$$

$$V_2\Sigma_2^+U_2^Tbb_2 = V_{22}\Sigma_{22}^{-1}U_{22}^Tbb_2 = \frac{1}{75}\begin{pmatrix} 27 & -23 & 4 \\ -10 & 15 & 5 \\ -11 & 14 & 3 \end{pmatrix}\begin{pmatrix} 5 \\ 4 \\ 12 \end{pmatrix} = \begin{pmatrix} \frac{91}{75} \\ \frac{70}{75} \\ \frac{37}{75} \end{pmatrix}$$

and this coincides with the solution obtained in questions 3 and 4 of the introductory examples, that is,

$$x_2 = \begin{pmatrix} 1+\alpha \\ 2-5\alpha \\ -1+7\alpha \end{pmatrix} \quad \text{with} \quad \alpha = \frac{16}{75}.$$

Remark 3.21. The SVD approach is about twice as expensive (computationally) as the QR approach. It is the only approach recommended for rank deficient problems, and can also be applied to large scale problems. Mirror segments need not be fixed via constraints for the only purpose to change the system from a rank deficient system to a full rank system.

3.2.2 Assessment of Computing Time, Mirror Displacement and Alignment Accuracy

In this section, we use 100000 trials of randomly sampled (generated by simulation) initial configurations of the SALT primary mirror under the assumption that each actuator position is within a given range $z_{\max}$ from the ideal position, and the mirror may be outside of acceptable alignment. Since

the SALT control system measures the relative heights and estimates the actuator positions from those relative heights, we will assess the difference between simulated actuator positions provided, and the estimated actuator positions from the corresponding relative heights, using normal equations, QR and SVD approaches respectively. Thus we may assess critical requirements for control of typical algorithm execution time and the precision demanded by the algorithm for initial alignment of mirrors by CCAS. Note that normal equations and QR are used when the system is of full rank, that is, when there are constraints to the system, in the unique purpose that from a given set of relative heights, we have a unique corresponding set of actuator displacements. SVD is used when the system is rank deficient, that is, it corresponds to the original system without constraints on any segment. In the rank deficient case, there is no unique set of actuator displacements from a given set of relative heights, and SVD chooses one of the sets of actuator displacements, with minimum norm. For our simulated experiments, we choose the maximum actuator displacement from zero to be $10^{-6} \leq z_{\max} \leq 10^{-4}$ metre. Each simulated actuator position can take any value between $-z_{\max}$ and $z_{\max}$, with equal probability (uniform probability distribution). Algorithmic precision is acceptable when RMS tip/tilt errors are less than 0.1 arcsecond.

Mirror Displacements

In this scenario, we use $z_{\max} = 10^{-4}$ metre, and we randomly generate actuator positions between $-z_{\max}$ and $z_{\max}$. Figures 3.2 and 3.3 illustrate, in a histogram, the RMS of actuator displacements in metres and the RMS of corresponding tip/tilts in arcseconds. RMS actuator positions vary between 5.06×10^{-5} and 6.47×10^{-5} metre, with an average of 5.77×10^{-5} metre and a high concentration around 5.8×10^{-5} metre. On the other hand, RMS tip/tilts vary between 35.73 and 51.74 arcseconds, with an average of 43.88 arcseconds and a high concentration around 44 arcseconds. This indicates that for each of 100000 trials, the mirror is out of acceptable alignment, and therefore has to be brought within acceptable alignment.

Computing Time

We give detailed examination of simulations in the case $z_{\max} = 10^{-4}$ metre. Table 3.1 gives information about computing times for estimating actuator

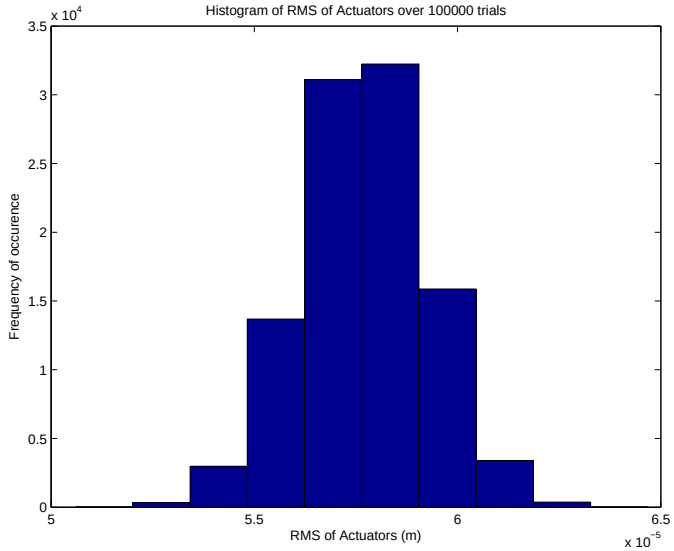

Figure 3.2: Histogram of RMS actuator displacements: with each initial actuator position between -10^{-4} metre and 10^{-4} metre, the probability of misalignment is very high

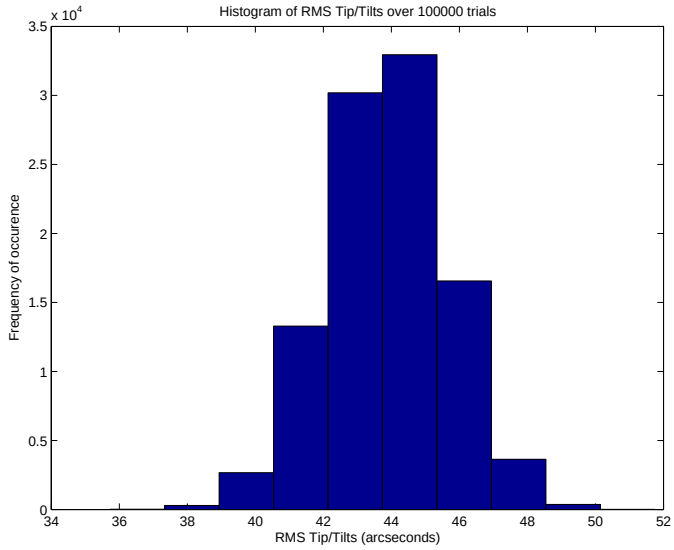

Figure 3.3: Histogram of RMS tip/tilts: with each initial actuator position between -10^{-4} metre and 10^{-4} metre, the probability of the primary mirror being out of acceptable alignment is very high

Table 3.1: Computing times (time to estimate actuator positions from a set of relative heights corresponding to a set of randomly generated actuator positions) using different methods over 100000 trials: Compared to the approximately 0.1 second actuator response time, computing time is not a concern

	Normal Equations	QR	SVD
Minimum time (s)	1.29×10^{-4}	1.22×10^{-4}	1.24×10^{-4}
Maximum time (s)	0.00471	0.00357	0.00479
Average time (s)	1.375×10^{-4}	1.336×10^{-4}	1.342×10^{-4}
$F\left(t \geq 10^{-4}\right)$	100000	100000	100000
$F\left(t \geq 2 \times 10^{-4}\right)$	248	171	206
$F\left(t \geq 3 \times 10^{-4}\right)$	52	42	33
$F\left(t \geq 4 \times 10^{-4}\right)$	28	28	13
$F\left(t \geq 5 \times 10^{-4}\right)$	21	19	9
$F\left(t \geq 10^{-3}\right)$	7	8	3

positions from a set of relative heights corresponding to a random set of simulated actuator positions, using different methods. The methods analysed are the normal equations method, the QR method and the SVD method. In each method, the pseudo inverse of the actuator-to-heights matrix is given and the computation reduces to a matrix-vector multiplication. In this table, time is given in seconds, $F\left(t \geq T\right)$ indicates how many times over 100000 trials the computing time is greater than the given value T. With the normal equations method, the computing time varies from 1.29×10^{-4} second to 0.00471 second, with an average of 1.375×10^{-4} second and a probability 2.48×10^{-3} of being greater than 2×10^{-4} second. With the QR method, the computing time varies from 1.22×10^{-4} second to 0.00357 second, with an average of 1.336×10^{-4} second and a probability 1.71×10^{-3} of being greater than 2×10^{-4} second. With the SVD method, the computing time varies from 1.24×10^{-4} second to 0.00479 second, with an average of 1.342×10^{-4} second and a probability 2.06×10^{-3} of being greater than 2×10^{-4} second. Thus, compared to the actuator response time which is approximately 0.1 second, the computing time is not a constraint on control time interval.

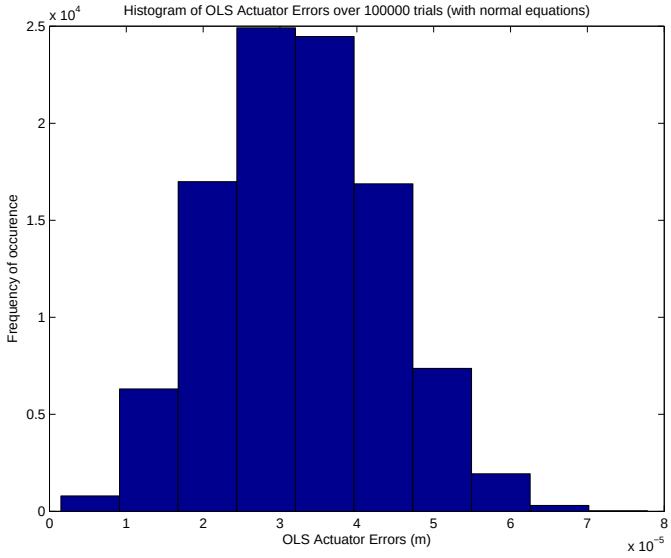

Figure 3.4: Histogram of RMS actuator errors using normal equations: algorithmic accuracy is not ideal as the actuator errors are a considerable fraction of the actuator displacements

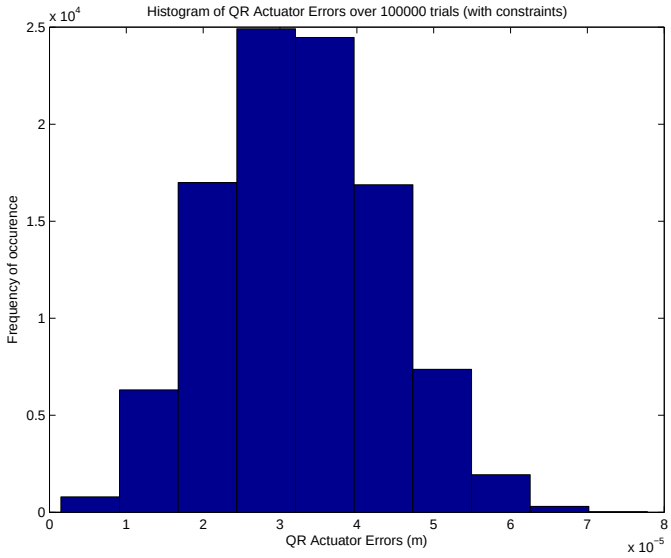

Figure 3.5: Histogram of RMS actuator errors using QR: algorithmic accuracy is not ideal as the actuator errors are a considerable fraction of the actuator displacements

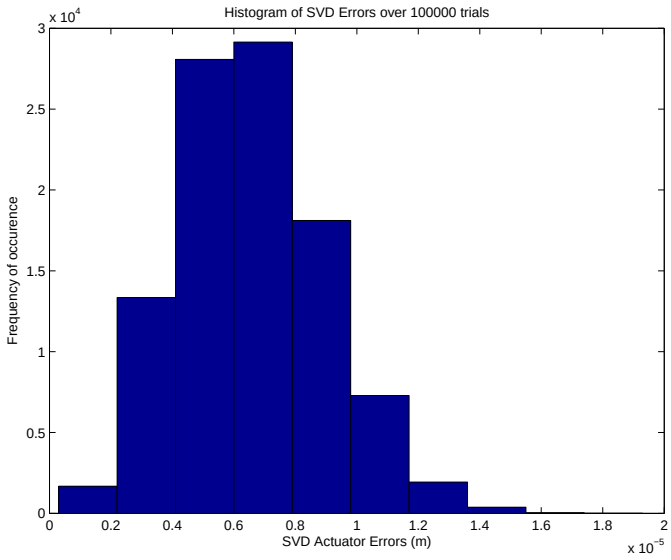

Figure 3.6: Histogram of RMS actuator errors using SVD: algorithmic accuracy is not ideal but much better than that of normal equations and QR

Actuator Displacement Errors

Again, we set $z_{\max} = 10^{-4}$ metre. Figures 3.4, 3.5 and 3.6 illustrate each in a histogram, the actuator displacement errors from the same simulated data of the previous paragraph. These errors are given as the RMS of the differences between the simulated values of actuator displacements, and the estimation obtained from computation. Of course, perfect algorithmic accuracy is the case of zero difference. For normal equations and QR methods, these actuator displacement errors vary between 1.48×10^{-6} and 7.78×10^{-5} metre, with an average of 3.26×10^{-5} metre, and a high concentration around 3×10^{-5} metre. For the SVD method, these actuator displacement errors vary between 3.01×10^{-7} and 1.93×10^{-5} metre, with an average of 6.58×10^{-6} metre and a high concentration around 6×10^{-6} metre. This indicates that for this simulation, SVD is able to estimate the actuator displacements more accurately than the normal equations and QR methods, and therefore suggests use of the SVD method and a mirror without physical constraints.

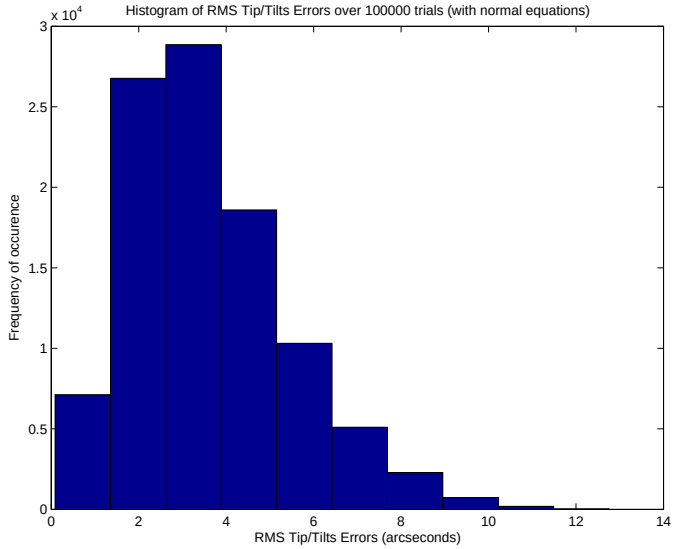

Figure 3.7: Histogram of RMS tip/tilt errors using normal equations: the probability is very high that the mirror is out of acceptable alignment (probability is about 10^{-4} for the mirror to be in the 0.1 arcsecond range) while SAMS believes the mirror is under control

RMS of Tip/Tilt Errors

We use, as in the previous paragraphs, $z_{\max} = 10^{-4}$ metre. Recall image quality is sensitive to tip/tilts and much less sensitive to pistons. Image quality is acceptable when the RMS of tip/tilts is less than 0.1 arcsecond. Figures 3.7, 3.8 and 3.9 illustrate each in a histogram, the tip/tilt errors from the same simulations. This is, again, the RMS tip/tilts corresponding to the difference between simulated and estimated actuator positions. For the normal equations and QR methods, RMS tip/tilt errors vary between 0.09 and 12.76 arcseconds, with an average of 3.61 arcseconds and a high concentration around 3 arcseconds. For the SVD method, RMS tip/tilt errors vary between 0.0044 and 3.0268 arcseconds, with an average of 0.71 arcsecond and a high concentration around 0.5 arcsecond. Once again, SVD method is better than normal equations and QR methods. However, it is a striking result that for $z_{\max} = 10^{-4}$ metre, only 9 of 100000 simulations give acceptable RMS tip/tilt errors for normal equations and QR methods with 480 of 100000 simulations acceptable for SVD method. It is clear that maximum allowable actuator displacements $z_{\max} = 10^{-4}$ metre do not lead to

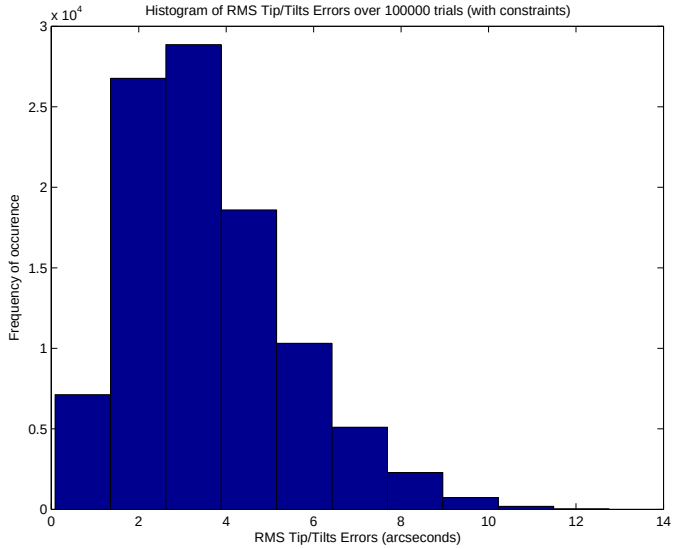

Figure 3.8: Histogram of RMS tip/tilt errors using QR: the probability is very high that the mirror is out of acceptable alignment (probability is about 10^{-4} for the mirror to be in the 0.1 arcsecond range) while SAMS believes the mirror is under control

acceptable control because all control algorithms fail with high probability.

3.2.3 Assessment of $z_{\max}$ for Acceptable Controllability

We have carried out the above simulations for various values of $z_{\max}$. In Table 3.2, we give the numbers of acceptable solutions in 100000 simulations. As before, we also give computing time in seconds per individual simulation.

Table 3.2: Rate of configurations with a 0.1 arcsecond RMS tip/tilt error over 100000 trials

	QR method	SVD method
$z_{\max} = 10^{-4}$m	9	480
$z_{\max} = 5 \times 10^{-5}$m	33	3372
$z_{\max} = 10^{-5}$m	3154	80261
$z_{\max} = 5 \times 10^{-6}$m	19092	99543
$z_{\max} = 10^{-6}$m	99714	100000

From Table 3.2, we note that SVD is more reliable than QR (and equiv-

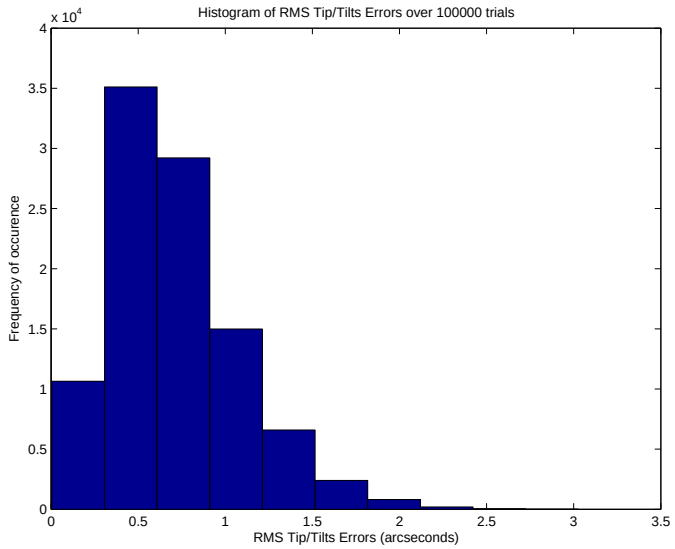

Figure 3.9: Histogram of RMS tip/tilt errors using SVD: the probability is high (but not as in normal equations and QR) that the mirror is out of acceptable alignment (probability is about 5×10^{-3} for the mirror to be in the 0.1 arcsecond range) while SAMS believes the mirror is under control

alently normal equations) for every value of $z_{\max}$. When $z_{\max} = 5 \times 10^{-6}$ metre, we see that for approximately 0.5% of trials, SVD will fail while QR remains unacceptable. At $z_{\max} = 10^{-6}$ metre, approximately 0.3% of trials would fail for QR while none were found to fail for SVD. Clearly, actuator displacements must be less than a micron for the control algorithm with QR to be reliable, and less than five microns for the control algorithm with SVD to be reliable.

Figures 3.10 and 3.11 illustrate in a histogram, the distributions of simulated actuator displacements and corresponding tip/tilts that allow achievement of acceptable controllability with a 99.5% probability using SVD, that is, when $z_{\max} = 5 \times 10^{-6}$ metre. We can see in this case that the primary mirror, again, is misaligned and out of focus but, as to be expected, in a smaller range than when $z_{\max} = 10^{-4}$ metre.

If the algorithms require a precision for individual actuator displacements of 10^{-6} metre, then it follows that the precision with which the drive motors of each actuator operate, must be within this same limit. This mechanical constraint is to be communicated to SALT engineers.

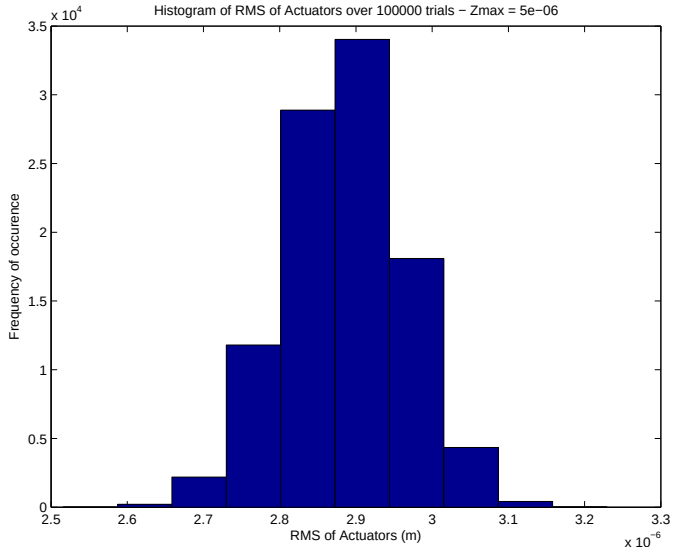

Figure 3.10: Histogram of acceptable RMS actuator displacements using SVD: with $z_{\max} = 5 \times 10^{-6}$ metre, acceptable controllability (less than 0.1 arcsecond RMS tip/tilt errors) is achieved via SVD with a 99.5% probability

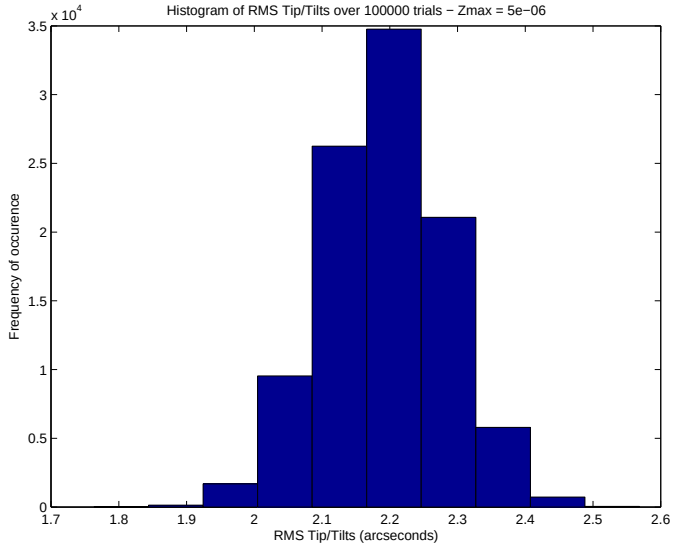

Figure 3.11: Histogram of acceptable RMS tip/tilts using SVD: with $z_{\max} = 5 \times 10^{-6}$ metre, acceptable controllability (less than 0.1 arcsecond RMS tip/tilt errors) is achieved via SVD with a 99.5% probability

Finally, we may note that computing time per simulation remains much less than the actuator response time (2.48×10^{-4} second on average for QR and 1.51×10^{-4} second on average for SVD) and is therefore not a constraint on controllability.

This section is of great importance because it sets limits to acceptable controllability of the segments, in terms of the degree of misalignment of the segments.

3.2.4 Filtering Data

This is necessary because of the noise in the sensor measurements. The measurements of relative heights are given with a frequency f_m and the control is performed with a frequency f_c. Moreover, f_m and f_c are chosen in such a way that there exists a positive integer P such that $f_m = Pf_c$. Therefore the averaging will be performed using the last P measurements including the current one. The filtering process is meant to estimate the value of the measurement at the time the control is about to be performed. Note that most of the results in this section can be derived by straightforward calculation.

Exploration of filters involves the concept of *z-transform*.

The z-transform and its Inverse [7, 8]

We consider a time series, or a discrete-time signal, or more generally a sequence $(x_n)_{n\geq 0}$. The (unilateral) *z-transform* is defined as follows:

$$X(z) = \mathcal{Z}\left(x_n\right) = \sum_{n=0}^{\infty} x_n z^{-n}$$

The *Region of Convergence (ROC)* is the set of points z in the complex plane for which the z-transform summation converges, that is:

$$ROC = \left\{ z : \left| \sum_{n=0}^{\infty} x_n z^{-n} \right| < \infty \right\}$$

The *inverse z-transform* is defined as follows:

$$x_n = \mathcal{Z}^{-1}\left(X(z)\right) = \frac{1}{2\pi i} \oint_{\gamma} X(z) z^{n-1} dz \qquad \forall n \geq 0$$

where γ is a counterclockwise closed path encircling the origin of the complex plane and entirely in the region of convergence. The contour path γ

must encircle all the poles of $X(z)$. The x_n are exactly the coefficients of the expansion of $X(z)$ in powers of z^{-1}, as long as z is in the region of convergence.

A simple example is when $x_n = ab^n$ for all $n \geq 0$. In this case the z-transform gives

$$X(z) = \mathcal{Z}(x_n) = \sum_{n=0}^{\infty} x_n z^{-n} = \frac{a}{1 - bz^{-1}}$$

provided z is in the region of convergence, which from some basic properties of z-transforms and geometric series, is given by: $ROC = \{z : |z| > |b|\}$.

Linear-Time-Invariant digital filters [4, 7, 8]

Background We consider a time series $(x_t)_{t\in\mathbb{N}}$ and we denote by s_t the image of x_t after a filtering process.

Overview Linear-Time-Invariant (LTI) digital filters are specific cases of linear filters and are characterised by their *transfer function* or by a difference equation involving the current and some past measurements as well as some past estimations. We choose two sets of factors $(a_n)_{0\leq n\leq M}$ and $(b_n)_{0\leq n\leq N}$ where M and N are positive integers and $a_0 = 1$. Then (s_t) is defined as follows:

$$s_t = -\sum_{n=1}^{M} a_n s_{t-n} + \sum_{n=0}^{N} b_n x_{t-n}$$

or equivalently

$$\sum_{n=0}^{M} a_n s_{t-n} = \sum_{n=0}^{N} b_n x_{t-n}.$$

If $a_n = 0$ for all $n > 0$ then the filter is a *Finite Impulse Response (FIR)* filter; otherwise it is an *Infinite Impulse Response (IIR)* filter.

Stability, Impulse response and Frequency response of the filter The *transfer function* of the filter is the ratio of the *Z-transform* of the output to that of the input and is given by

$$H(z) = \frac{B(z)}{A(z)} = \frac{\sum_{n=0}^{N} b_n z^{-n}}{1 + \sum_{n=1}^{M} a_n z^{-n}}$$

and the frequency response is $H\left(e^{i\omega}\right)$ for all frequencies ω.

The filter is considered to be stable if all its poles are in the unit circle (on the complex plane).

The impulse response is the inverse Z-transform of the transfer function. This can be obtained by decomposing the transfer function into partial fractions and summing up the inverse Z-transforms of the obtained simple expressions using the properties of the Z-transform, or by the long division of the numerator by the denominator, both in ascending order of z^{-1}. The long division gives an equation of the form

$$H(z) = \sum_{n=0}^{\infty} h_n z^{-n}$$

and the h_n give the impulse response. However, the impulse response can also be determined recursively as follows:

$$\begin{cases} h_0 = b_0 \\ h_n = \sum\limits_{k=0}^{M-1} b_k \delta_{n-k} - \sum\limits_{l=1}^{N-1} a_l h_{n-l} \quad n > 0 \end{cases}$$

where δ_n is the *Kronecker Delta impulse* defined on $\mathbb{Z}$ as follows:

$$\delta_n = \begin{cases} 1 & \text{if } n = 0 \\ 0 & \text{if } n \neq 0 \end{cases}$$

Simple moving average (used on SALT) This is an example of *FIR* filter.

Overview The simple moving average consists of averaging the values of x for the last k steps including the current step. Therefore we have

$$s_t = \frac{1}{k} \sum_{n=0}^{k-1} x_{t-n} = \frac{x_t + x_{t-1} + \cdots + x_{t-k+1}}{k} = s_{t-1} + \frac{x_t - x_{t-k}}{k}.$$

One of the disadvantages of the simple moving average is that s_t cannot be determined for $t < k$. However, in that case we can define s_t as follows:

$$s_t = \frac{1}{t} \sum_{n=0}^{t-1} x_{t-n} = \frac{1}{t} \sum_{n=1}^{t} x_n.$$

An alternative, as used in many softwares, is to keep the window size constant (size k) precede the available data with zeros to fill the window, and compute the corresponding average. This gives:

$$s_t = \frac{1}{k}\sum_{n=0}^{t-1} x_{t-n} = \frac{1}{k}\sum_{n=1}^{t} x_n.$$

Stability and Impulse response of the filter The transfer function of the filter is given by

$$H(z) = \frac{1}{k}\left(1 + z^{-1} + \cdots + z^{-k+1}\right) = \frac{1}{k}\left(\frac{1 + z + z^2 + \cdots + z^{k-1}}{z^{k-1}}\right).$$

It is clear that the zeros of the transfer function are complex numbers on the unit circle (all the k^{th} roots of unity except 1) and the only pole of the transfer function is 0, which is in the unit circle. Hence the filter is stable. This can be illustrated in the so called *pole-zero diagram* of the transfer function.

It can easily be established that the impulse response of the filter is given by $h_n = \frac{1}{k}$ for $0 \leq n \leq k-1$.

Weighted moving average This is also an example of *FIR* filter.

Overview We choose a set of weighting factors $(w_n)_{1\leq n\leq k}$ such that $\sum_{n=1}^{k} w_n = 1$ and we define (s_t) as follows:

$$s_t = \sum_{n=1}^{k} w_n x_{t+1-n} = w_1 x_t + w_2 x_{t-1} + \cdots + w_k x_{t-k+1}.$$

Like the simple moving average, the weighted moving average has the disadvantage that s_t cannot be determined for $t < k$. However, in that case we can choose a set of weighting factors $(w_n)_{1\leq n\leq t}$ such that $\sum_{n=1}^{t} w_n = 1$ and define

$$s_t = \sum_{n=1}^{t} w_n x_{t+1-n}.$$

An alternative, as in the previous case, is to keep the window size constant (size k) precede the available data with zeros to fill the window, and compute the corresponding average. This means, using a similar method as in the

simple moving average case, we consider the same set of weighting factors $(w_n)_{1\leq n\leq k}$ such that $\sum_{n=1}^{k} w_n = 1$ and this time we define (s_t) as follows:

$$s_t = \sum_{n=1}^{t} w_n x_{t+1-n} = w_1 x_t + w_2 x_{t-1} + \cdots + w_t x_1.$$

In practice, and to give a better meaning to the case of incomplete available data (case when $t < k$ as illustrated in the equation above), it is preferable to choose the set of weighting factors $(w_n)_{1\leq n\leq k}$ as a positive decreasing sequence so as to give more weights to the most recent measurements in the data from the time series.

Stability and Impulse response of the filter The transfer function of the filter is given by

$$\begin{aligned} H(z) &= w_1 + w_2 z^{-1} + \cdots + w_k z^{-k+1} \\ &= \frac{w_k + w_{k-1} z + w_{k-2} z^2 + \cdots + w_1 z^{k-1}}{z^{k-1}} \end{aligned}$$

It can be verified that the zeros of the transfer function are complex numbers in the unit circle and the only pole of the transfer function is 0, which is in the unit circle. Hence the filter is stable. This can be illustrated in the *pole-zero diagram* of the transfer function.

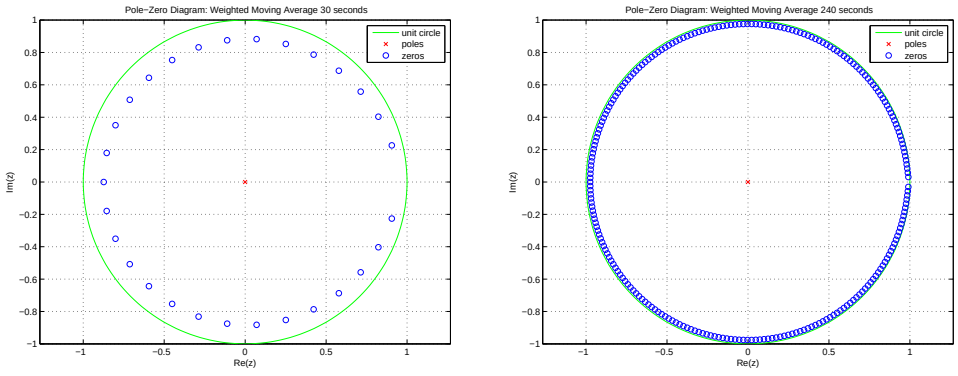

Figure 3.12: Pole-zero diagram of the weighted moving average filter using the last 30 measurements (left) and 240 measurements (right)

The pole-zero diagram in Figure 3.12 (where the corresponding filter uses a positive decreasing arithmetic sequence of coefficients with a total

sum equal to 1) suggests that the bigger the number of past measurements used, the closer the weighted moving average is to the simple moving average.

It can also be established that the impulse response of the filter is given by $h_n = w_{n+1}$ for $0 \leq n \leq k-1$.

Exponential moving average This is an example of *IIR* filter.

Overview We choose a real number α (called *smoothing factor*) such that $0 < \alpha < 1$ and (s_t) is defined as follows:

$$s_0 = x_0 \quad \text{and} \quad s_t = \alpha x_t + (1-\alpha)s_{t-1} = s_{t-1} + \alpha\left(x_t - s_{t-1}\right), \quad t > 0.$$

It can be established that

$$s_t = (1-\alpha)^t x_0 + \alpha \sum_{n=1}^{t} (1-\alpha)^{n-1} x_{t+1-n}.$$

Stability and Impulse response of the filter The transfer function of the filter is given by

$$\begin{aligned} H(z) &= \frac{\alpha}{1+(\alpha-1)z^{-1}} = \frac{\alpha}{1-(1-\alpha)z^{-1}} \\ &= \frac{\alpha z}{z+(\alpha-1)} = \frac{\alpha z}{z-(1-\alpha)} \\ &= \sum_{n=0}^{\infty} \alpha(1-\alpha)^n z^{-n}. \end{aligned}$$

It is clear that the zero of the transfer function is at the origin and the only pole of the transfer function is $1-\alpha$, which is in the unit circle since $0 < \alpha < 1$. Hence the filter is stable.

The impulse response of the (IIR) filter is given by $h_n = \alpha(1-\alpha)^n$ for all integers $n \geq 0$.

Progressive Linear Fit This method is to be considered only when the measurements follow a linear pattern (with some noise). Progressive linear fit will not be explored in the scope of this thesis, since by inspection the measurements do not follow a linear pattern.

3.2.5 Conclusion

To summarise, we are concerned that SALT minimisation and filtering are not robust. We suggest singular value decomposition to handle the minimisation problem, and weighted moving average and exponential moving average with appropriate parameters (a suitable set of weighting factors and therefore time interval for the weighted moving average, and suitable smoothing factor and time interval for the exponential moving average) to handle the filtering process, to be added in the SALT software. In the case of weighted moving average, we suggest a 4 minutes correction time and an arithmetic sequence of positive decreasing weighting factors with total sum equal to one while for exponential moving average, we suggest a 30 seconds correction time and a smoothing factor $\alpha = 0.5$. This has been done and experiments have been performed for comparison purposes.

3.3 Effect of Least Squares and Filtering on the Control of SALT

Here we give results from observations conducted before and after improvement of numerical algorithms, with comments and analyses. For comparison purposes, we choose samples from nights when environmental conditions are reasonable and comparable. Note that results obtained before numerical corrections are from experiments conducted on the telescope with the entire primary mirror, that is 91 segments and 480 sensors. However, during experiments involving numerical corrections, we only worked with the central 7 segments, that is, the central segment and the first ring. Remaining segments were disabled for work on edge sensors.

Remark 3.22. The legends on illustrative figures have different interpretations before and after numerical corrections.

1. Before corrections, SAMS refers to output from SAMS using the original approach; CCAS refers to output from CCAS; CCAS-All indicates how many segments are used for CCAS measurements, that is, how many segments are not obstructed; CAM refers to output from our approach, for comparison with the original SAMS output and CCAS output when it is possible.

2. After corrections, since our approach is implemented in SAMS algorithm, SAMS refers to output from our approach, for comparison with output from CCAS when it is possible; CCAS refers to output from CCAS; CCAS-CORR refers to output from CCAS after GRoC correction, that is, after adjustment of all the segments in tip, tilt and piston, resulting from the change in radius of curvature; CCAS-All indicates how many segments are used for CCAS measurements, that is, how many segments are not obstructed.

3.3.1 Existing SALT Results Before Least Squares and Filtering Corrections

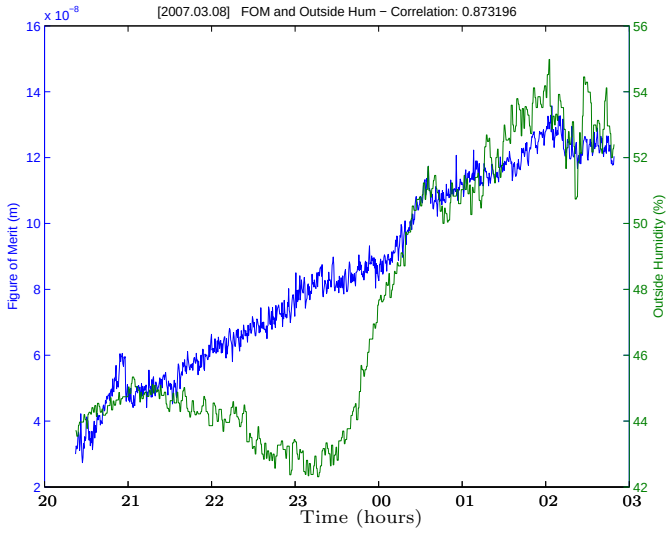

Figure 3.13: FoM and humidity before corrections (08 March 2007): the behavior of figure of merit shows concern about mirror controllability while humidity, although strongly correlated to figure of merit, displays a different pattern

Figure 3.13 illustrates how the figure of merit and humidity evolve with time. It shows the figure of merit growing almost linearly from about 30nm to about 140nm, though it would have been preferable for it to remain below 60nm at all times. Outside humidity is increasing on average and is highly correlated to the figure of merit (correlation coefficient 0.87). Moreover, in this case, it is part of the significant explanatory variables. The behavior of the figure of merit shows concern about the efficiency of the control

algorithm.

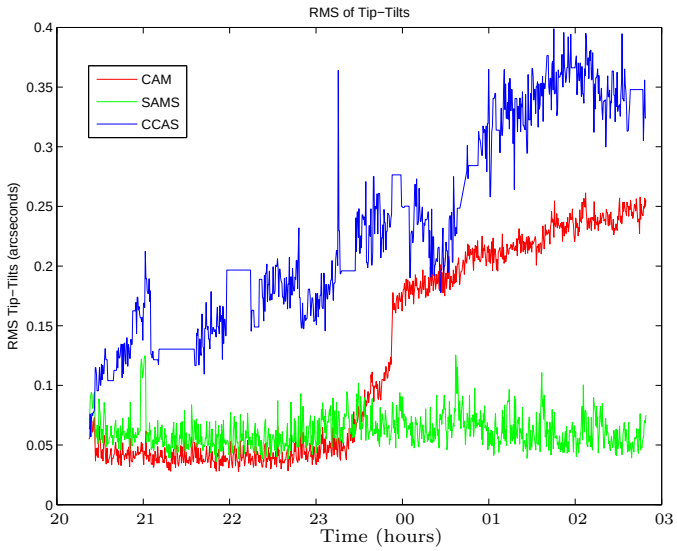

Figure 3.14: RMS of tip/tilts before corrections (08 March 2007): SAMS believes the mirror is under control (green), which is in disagreement with CCAS (blue); our calculations (red) confirms that there might be a problem with SAMS; the system seems to undergo a sudden disturbance around 11pm, which is not captured by SAMS

Figure 3.14 illustrates how the Root Mean Square of tip/tilts evolves with time. According to SAMS, the mirror alignment (emphasising on focus) is acceptable since the RMS of tip/tilts is almost always below 0.1 arcsecond. However, CCAS measurements of tip/tilts reveal that the RMS grows almost linearly with time (from 0.05 to about 0.35 arcsecond), which means the system is going out of focus as time goes, which is not in agreement with SAMS. We show a time-series labelled CAM. This is our code, using the normal equations approach of SAMS, applied to the SALT data of 08 March 2007 for time-series of relative heights. Moreover, our approach implemented for comparison agrees with SAMS to some extent at the beginning, and then gets closer to the CCAS output. The sudden change between 11pm and midnight suggests the system has been disturbed. This period corresponds to a swing in humidity from decreasing to rapidly increasing (in our available time-series), which might supply a physical reason for the disturbance. Our encoding of the normal equations approach (CAM) shows an abrupt adjustment towards measured CCAS tip/tilts around 11pm. We

also note increasing tip/tilts from our approach, in agreement with CCAS while SAMS remains stable. Also note that the RMS tip/tilts from CCAS goes out of the range of 0.1 arcsecond within a few minutes from the beginning, and never manages to come below 0.1 arcsecond again. This confirms that there is a flaw in reliability of the control system of SAMS, even with the normal equations method.

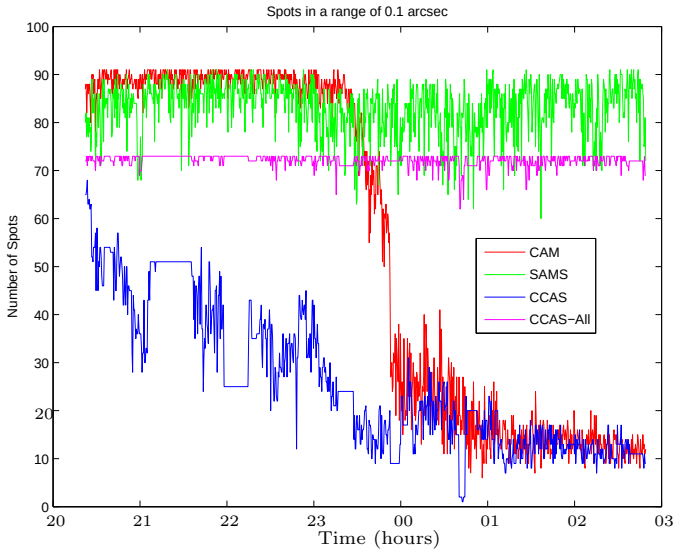

Figure 3.15: Spots in acceptable range before corrections (08 March 2007): SAMS (green) believes that almost all the time, at least 80 segments are in acceptable range (0.1 arcsecond) while CCAS indicates that out of 73 segments in general (purple), at most 60 and decreasing (blue), are in acceptable range; our approach (red) agrees with SAMS, then switches to CCAS around 11pm, indicating a possible disturbance

Figure 3.15 indicates how many segments are close to the ideal position in terms of tips and tilts. According to SAMS, almost all the time, at least 80 segments are within the range of 0.1 arcsecond. This directly contradicts the CCAS output revealing that at most 60 segments are within the 0.1 arcsecond range, and decreasing to about 10 segments within six to seven hours. Note that due to obstruction, CCAS measurements are not provided for all the 91 segments. However, measurements are provided from CCAS for about 73 segments at all times, as illustrated by the curve labelled CCAS-All. A slight difference between the outputs of SAMS and CCAS would be understandable due to the fact that some segments are obstructed, but not

Table 3.3: Significant Environmental Variables (08 March 2007)

Variable	Minimum	Maximum	Mean	Variance
ttr (^{o}C)	14.5	17.5	15.92	0.81
windin (m/s)	0	3	1.24	0.27
wind30 (m/s)	4	8	5.69	0.63
hum30 (%)	42	55	47.42	16.07

up to about 80 segments at some point. On the other hand, our code (the CAM implementation of the normal equations approach) is in agreement with SAMS for the first three hours, and then drastically switches to almost match CCAS output between 11pm and midnight. This is again strong evidence of an error in the SAMS implementation of the normal equations approach. SAMS thinks the system is under control, while CCAS output suggests otherwise.

The observations were conducted on 08 March 2007. From section 2.1.2, stepwise regression (assuming we discard dgroc and tref for reasons mentioned there) reveals time, temperature of truss, outside humidity, inside and outside wind speeds to be the main explanations for the figure of merit. This suggests computational and environmental conditions are the causes, and indeed these results are supported by Figures 3.13, 3.14 and 3.15. The environmental variables that are significant for 08 March 2007 are summarised in Table 3.3.

3.3.2 After Least Squares Correction

In July 2010, we visited SALT in order to include our code, the QR/SVD approach to least squares problems, in the SAMS software using a multiplexer. In this section, we therefore give diagrams (Figures 3.16, 3.17 and 3.18) corresponding to the respective diagrams in Figures 3.13, 3.14 and 3.15 of the previous section, but for the data of July 2010.

Indeed, we wrote and implemented a new code in Labview (SALT operational software for the control algorithm) and included this in the SALT original software using a multiplexer. We were then given one week (12 to 15 July 2010), to conduct our experiments on the telescope, and then the following week, in order to increase the number of experiments for efficient comparison with previous results (before numerical corrections). However,

in this time range, observations were possible only for six nights, for a few hours per night, due to environmental restrictions such as rain and high humidity. Note that we were working with 7 segments, and not all the 91 segments, and the results in this section are for this reduced configuration.

The least squares method is used in computing actuator positions from relative heights. The original SAMS algorithm used the normal equations approach. We used the QR approach when the actuator-to-heights matrix has full rank, and the SVD approach when the actuator-to-heights matrix is rank deficient. The dependence on the rank of the actuator-to-heights matrix is due to the fact that the configuration of this actuator-to-heights matrix changes with how many sensors and how many segments are used. Specifically, the entries in the actuator-to-heights matrix corresponding to disabled sensors or unused segments will change to zero. Note that the initial actuator-to-heights matrix with all the sensors and all the segments operational, is rank deficient and additional constraints are used to transform the system from a rank deficient system to a full rank system, under the initial assumption that everything works perfectly well.

We give results for 13 July 2010. In that night, humidity was low. At SALT, there is mistrust of the performance of capacitive edge sensors in high humidity.

Figure 3.16 illustrates how the figure of merit and humidity evolve with time. It shows a sudden jump of the figure of merit from 20nm to about 80nm within about 15 minutes. Then it stabilises for about 90 minutes, and continues to increase up to about 120nm within 90 minutes. On the other hand, although the humidity curve follows that of the figure of merit, their correlation is reduced (correlation coefficient 0.68), compared to Figure 3.13, which indicates that control has improved.

In Figure 3.17, SAMS denotes the results from the QR/SVD approach, CCAS is as usual the output from CCAS, CCAS-CORR is the output from CCAS after software adjustments compensating for the change in radius of curvature. Figure 3.17 illustrates how the Root Mean Square of tip/tilts evolves with time. Note that according to our implementation of the least squares approach using QR/SVD included in the SAMS algorithm, the root mean square of tip/tilts is almost always around 0.05 arcsecond. This is not in agreement with CCAS showing a linear increase from 0.05 to about 0.2 arcseconds within three hours. This disagreement is dramatic and suggests

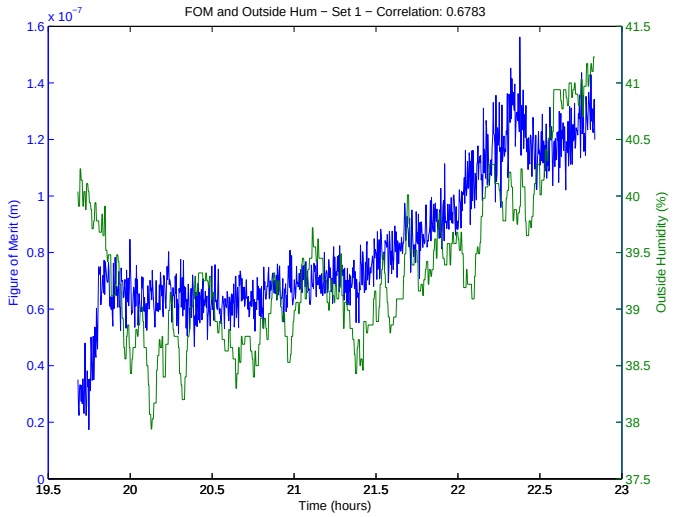

Figure 3.16: FoM and humidity after least squares corrections (13 July 2010): figure of merit increases rapidly at start, then stabilises, but still gets out of range; its curve follows that of humidity, but with reduced correlation compared to the case before corrections

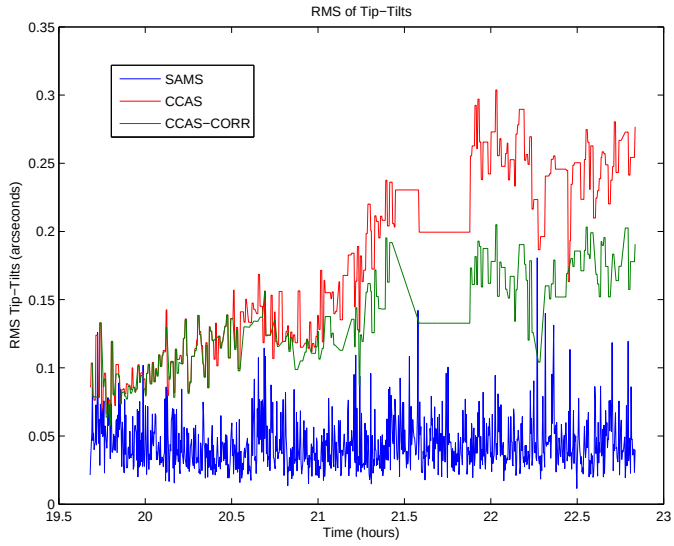

Figure 3.17: RMS of tip/tilts after least squares corrections (13 July 2010): from our approach (QR/SVD), SAMS believes the primary mirror is under control (blue), which is in disagreement with CCAS (red before GRoC corrections and green after GRoC corrections); this suggests further improvements should follow

a need of improvement in the control system, that is, by itself the QR/SVD approach is not sufficient for the control of SALT primary mirror. We note that the normal equations method will agree with our method (the QR/SVD method implemented in SAMS control algorithm) when the least squares problem is a full rank problem.

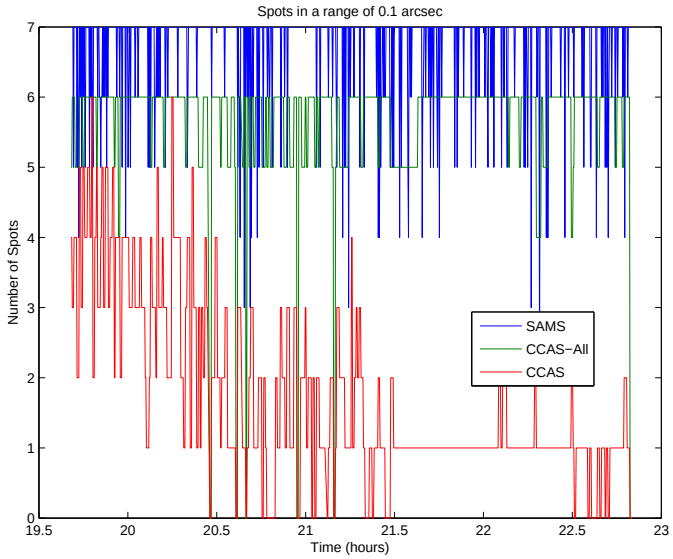

Figure 3.18: Spots in acceptable range after least squares corrections (13 July 2010): according to SAMS (blue), at least 5 of 7 segments are in the 0.1 arcsecond range almost all the time, while from CCAS, out of 5 or 6 segments, at most 5 (and decreasing) are in the 0.1 arcsecond range; this confirms need of further improvements

Figure 3.18 indicates how many segments are close to the ideal position in terms of tips and tilts. Recall that CCAS-All having values less than 7 means that some of the segments are obstructed. According to SAMS, almost all the time, at least 5 of the 7 segments are within the range of 0.1 arcsecond and according to CCAS, almost all the times, 5 or 6 segments are provided finite values from measurements, and the number of segments within the 0.1 arcsecond range decreases from 5 to 1, and sometimes reaches 0, within 90 minutes to two hours. Again, we have a dramatic disagreement. This suggests that improvement needs to be made in the control algorithm.

The experiments were conducted on 13 July 2010. From section 2.1.2, stepwise regression reveals time, temperature of truss and temperature of

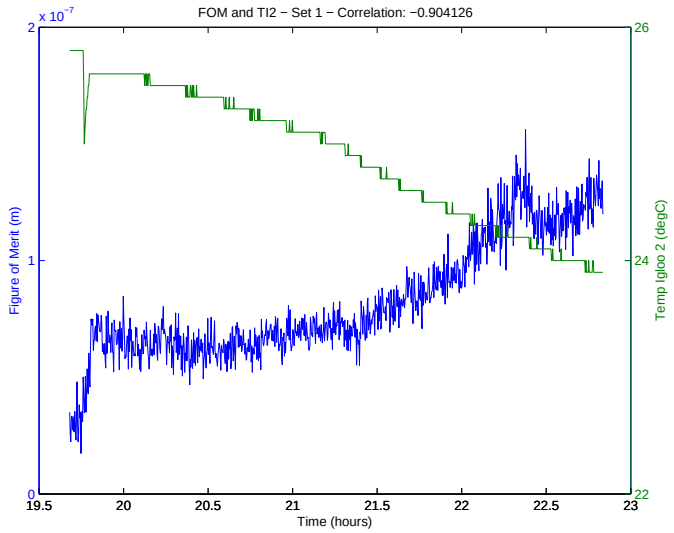

Figure 3.19: Figure of Merit and Temperature of Igloo 2 after least squares corrections (13 July 2010): temperature of igloo 2 decreases while FoM increases and they are strongly correlated; lack of stability of temperature of igloo can have a serious impact on the outcome of the electronics

igloo 2 to be the main explanation for the figure of merit. The time variable suggests computation be improved. Moreover, temperature of igloo 2 is not kept as stable as desired (illustration in Figure 3.19), and can have a serious impact on the outcome from the electronics. The environmental variables that are significant for 13 July 2010 are summarised in Table 3.4.

Table 3.4: Significant Environmental Variables (13 July 2010): temperature of truss is relatively stable, even more stable than temperature of igloo 2, yet temperature of igloos should be kept very stable for reliability of outcome from the electronics

Variable	Minimum	Maximum	Mean	Variance
ttr (^{o}C)	2.4	3.7	3.07	0.15
ti2 (^{o}C)	23.9	25.8	24.89	0.32

Experiments conducted in July 2010 reveal that after least squares correction using QR or SVD (instead of normal equations) depending on whether the system is full rank or rank deficient, computation is still to be improved in the control algorithm.

3.3.3 Filtering Data for Correction in the Control Process

On SALT, the standard filtering technique is a simple moving average over 4 minutes. We investigated the power of this filter, in search of a fix for the disagreement discovered in the previous section.

Weighted Moving Average

The results given here are for a weighted moving average filter with a 4 minutes correction time and positive linearly decreasing weighting factors. This specification, by inspection of all the conducted experiments using the weighted moving average, gave the best experimental performance.

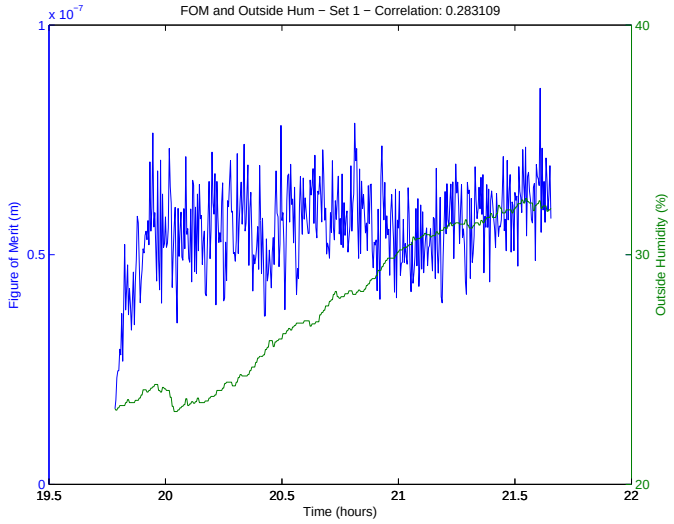

Figure 3.20: FoM and humidity after [weighted moving average] filtering corrections (01 March 2011): humidity increases almost linearly but remains relatively low while figure of merit rapidly increases from 20 nm to about 60 nm and tends to stbilise at that level for the remaining time; this indicates improvement over simple moving average previously used

Figure 3.20 illustrates how the figure of merit and humidity evolve with time. By simple inspection, the figure of merit grows from 20 to 60 nanometres in about 15 minutes and tends to stabilise at that level for the remaining time. Meanwhile, the humidity grows almost linearly from 23% to about 32% in 90 minutes, which means humidity is relatively low. The system is under control in the time range of this experiment. The figure of merit is stable

around 60 nanometres, which clearly shows an improvement over the experiment illustrated in Figure 3.16. Also note that the correlation between the figure of merit and humidity is low (correlation coefficient 0.28).

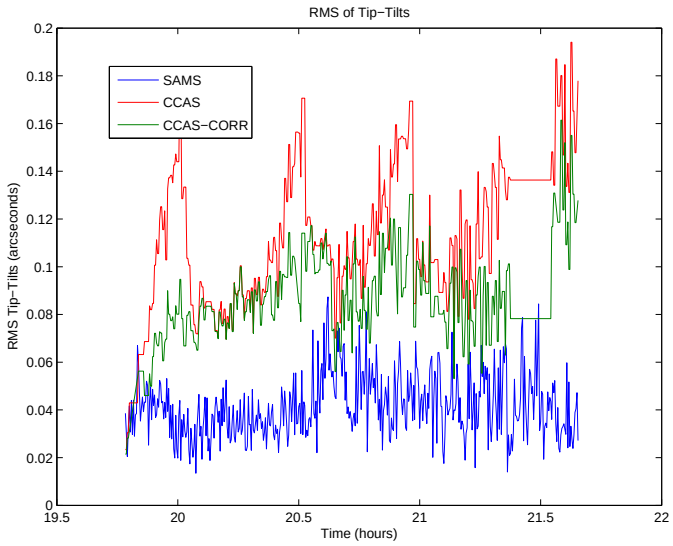

Figure 3.21: RMS of tip/tilts after [weighted moving average] filtering corrections (01 March 2011): according to SAMS (blue) the system is under control, which is to some extent in agreement with CCAS after GRoC corrections (green); output from CCAS before GRoC corrections goes out of range from time to time but attempts to recover; this confirms weighted moving average as an improvement over simple moving average

Figure 3.21 illustrates how the Root Mean Square of tip/tilts evolves with time. According to SAMS, the RMS of tip/tilts is within acceptable range the whole time. Output from CCAS, especially after GRoC correction (CCAS-CORR), reveals that the RMS of tip/tilts is also within acceptable range. This was not the case in Figure 3.17 and again we see improvement.

Figure 3.22 indicates how many segments are close to the ideal position in terms of tips and tilts. According to SAMS, almost all the time, at least 6 of the 7 segments are within acceptable range. CCAS has measurements for 6 segments almost all the time and out of these 6 segments providing measurements from CCAS, the number of segments within acceptable range seems to decrease almost linearly from 6 to 1, and sometimes reaches zero, but most of the time this number is between 2 and 4. Compared to Figure 3.18, the disagreement between SAMS and CCAS has been reduced (from

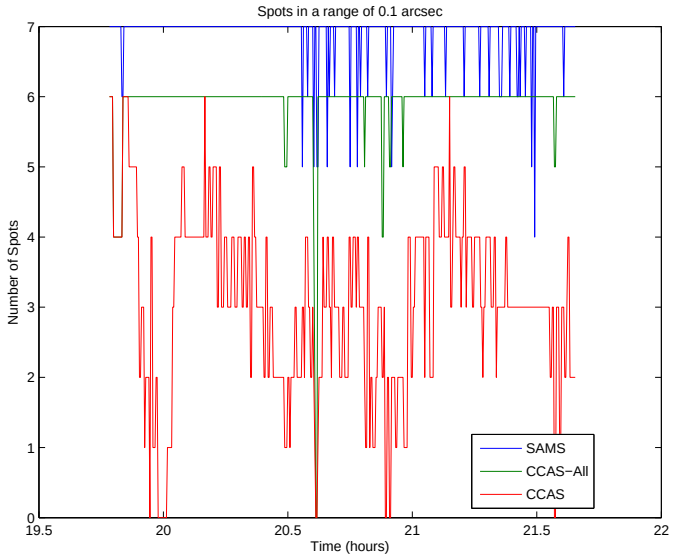

Figure 3.22: Spots in acceptable range after [weighted moving average] filtering corrections (01 March 2011): according to SAMS (blue), at least 6 of the 7 segments are in acceptable range while from CCAS, out of 6 segments (green), most of the time 2 to 4 segments are in acceptable range (red)

Table 3.5: Significant Environmental Variables (WMAVG 01 March 2011)

Variable	Minimum	Maximum	Mean	Variance
hum30 (%)	23.17	32.45	27.79	10.17

about 6 to about 4) and thus we again have improvement over the previous method.

The experiments were conducted on 01 March 2011. From section 2.1.2, stepwise regression reveals time and outside humidity as the main explanation of the figure of merit. This suggests computation and humidity as the main explanation for the figure of merit. Note that humidity is relatively low and also that the figure of merit is stable around 60 nanometres after growing quickly from about 20 nanometres in the first 15 minutes of the experiment. The environmental variables that are significant for 01 March 2011 using the weighted moving average as a filtering process are summarised in Table 3.5.

Although humidity is the only significant explanation to the figure of

merit, it is not highly correlated to the figure of merit, and it is also relatively low, which makes the measurements obtained from the capacitive edge sensors more reliable. This suggests an improvement has indeed been achieved in the control algorithm.

Exponential Moving Average

Here we consider the exponential moving average filter with 30 seconds correction time and smoothing factor $\alpha = 0.5$. This combination, by inspection of all the conducted experiments using the exponential moving average, gave the best experimental performance.

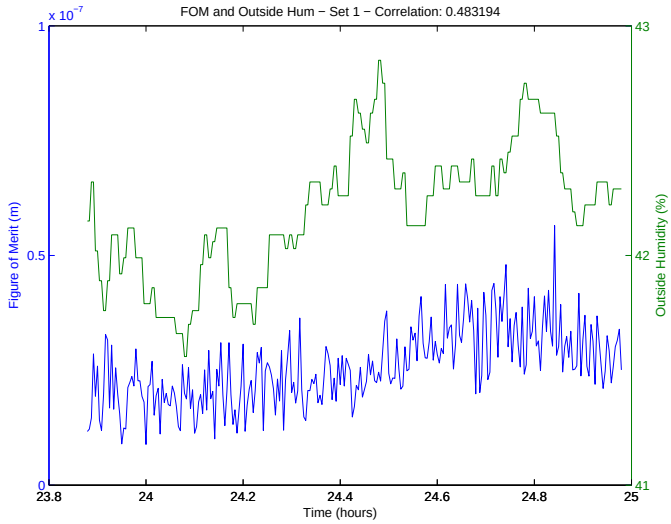

Figure 3.23: FoM and humidity after [exponential moving average] filtering corrections (01 March 2011): figure of merit is stable and below 50 nm, and not highly correlated to humidity which is relatively low and follows a pattern similar to that of figure of merit; this suggests exponential moving average is an improvement over weighted moving average

Figure 3.23 illustrates how the figure of merit and humidity evolve with time. There is no sudden jump in the figure of merit and this figure of merit is almost always below 50 nanometres. Humidity is low (between 41.5% and 43%) and has a pattern similar to that of the figure of merit. However, their correlation is low (correlation coefficient 0.48). The figure of merit has now fallen (improved) by a factor of 2 over that of the weighted moving average over 4 minutes illustrated in the previous section. This indicates the system

is under control, and that improvement has been achieved over the weighted moving average method.

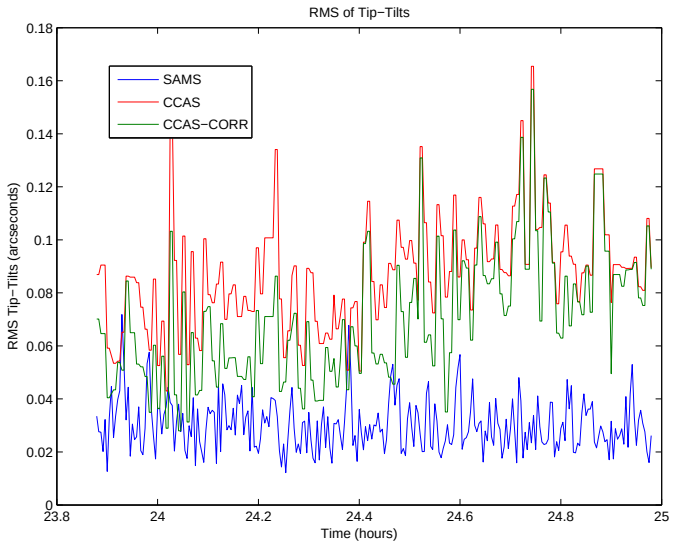

Figure 3.24: RMS of tip/tilts after [exponential moving average] filtering corrections (01 March 2011): according to SAMS (our approach using QR/SVD and exponential moving average, in blue) the system is under control; this is also the case with the output from CCAS before (red) and after GRoC corrections (green); this confirms an improvement over previous methods

Figure 3.24 illustrates how the Root Mean Square of tip/tilts evolves with time. According to SAMS (our version involving the QR/SVD method and the exponential moving average filtering), the system is under control. Note that both SAMS and CCAS are within acceptable range of RMS tip/tilts which is below 0.1 arcsecond.

Figure 3.25 indicates how many segments are close to the ideal position in terms of tips and tilts. SAMS reveals that all the 7 segments are in acceptable range of tip/tilts except for very few measurements where 6 of the 7 segments are in acceptable range. CCAS reveals for all time, that measurements are provided from exactly 6 segments, and out of these 6 segments, at least 3 are in acceptable range of tip/tilts almost all the time, with a bigger concentration between 4 and 5. This again shows improvement has been achieved.

The experiments were conducted on 01 March 2011. Stepwise regression reveals only time as a significant explanatory variable for the figure of merit

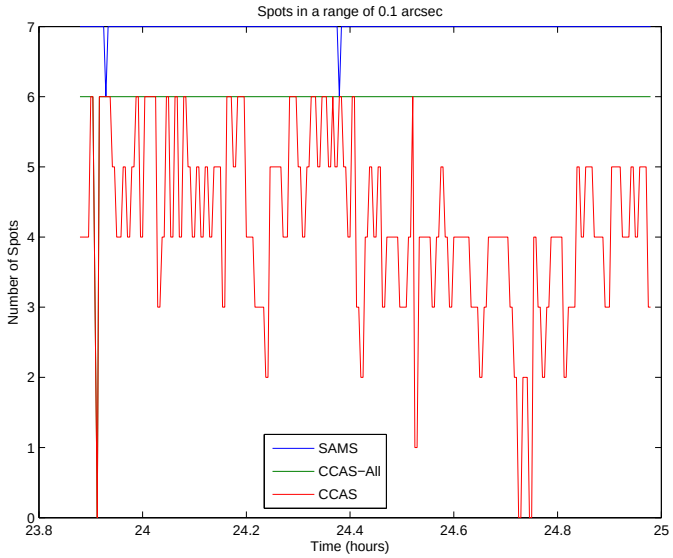

Figure 3.25: Spots in acceptable range after [exponential moving average] filtering corrections (01 March 2011): according to SAMS (blue) all the segments are in acceptable range while from CCAS out of 6 segments (green) at least 3 are in acceptable range almost all the time (red)

as response. Note that the figure of merit is always below 50 nanometres. This means computation would be the only explanation to the figure of merit if this figure of merit was out of reasonable range. It also suggests that, in the context of the current analysis, humidity is not a serious concern and that the main aspect to explore is computation.

3.4 Conclusion

The control algorithm for SALT has been explored. Especially, the least squares method has been explored as it is a part of the control algorithm. To solve least squares problems using numerical methods, the normal equations approach was the method initially used on SALT but we find the QR approach or the SVD approach depending on the rank of the actuator-to-heights matrix, to be computationally more reliable.

Our approach (the QR/SVD approach) gives us the power to choose between the QR and the SVD approaches as desired, depending on the rank of the actuator-to-heights matrix which also depends on how many

segments and sensors are used, since the whole system is not expected to work perfectly well at all times. In general, SVD should be used.

It is an important result that random errors in actuator displacements must be of order one micron for algorithms of any method to give acceptable tip/tilts.

We investigated filtering by simple moving average as it is used on SALT, and compared this with weighted moving average and exponential moving average methods. Recall that control intervals on SALT were set at four minutes. We find superior control by filtering with exponential moving average which allows control intervals of thirty seconds.

It is of a great importance that our results be adopted by SALT, because we have essentially shown that for the full 91 segments, the original normal equations approach will not be reliable and that simple moving average filtering is inefficient. Our code is multiplexed into SALT software and should be used in the future.

We wish to make clear that perfect edge sensors (sensors providing zero measurement error) will not guarantee that SALT control system will work. This is because besides the temperature of truss and humidity, time is the most significant explanation of the figure of merit, which means computation is the most likely main cause of poor image quality. The numerics proposed in this chapter provide an improvement on the implementation of the current control system. In particular, we found unacceptable accumulation of numerical errors unless SVD was implemented with exponential moving average (see Figure 3.23, page 96). With improved numerics, we found that RMS actuator precision must be stringently chosen at better than one micron (Table 3.2). SALT staff will be informed of this. In addition, we identified errors and omissions in SALT software (that is, deviations from specification and documentation of SALT). A trivial example was the inconsistency in the dimensions of tips, tilts and pistons. This chapter provides consistent documentation with our implementation of the re-designed SALT control system. Finally, we note that should the edge sensors be replaced with more accurate sensors, our software (or any other software used on SALT to implement the control algorithm), should be retested.

Moreover, improvements can still be made, in particular, to control the primary mirror from an arbitrary initial misalignment. This involves more sophisticated mathematical techniques that are explored in the next chapter.

Chapter 4

Theoretical Approaches to Control of Segmented Mirrors for Arbitrary Initial Misalignment

In this chapter, we explore some theoretical approaches to the mathematical control of a multi-element telescope, based on a background in optimal control theory. These approaches can be applied in discrete or in continuous time, and most importantly, can consider the option of aligning the segments in time under arbitrary initial configuration.

4.1 Fast Alignment by Control

The results from section 3.2.3 clearly show that the normal equations and QR methods, which give similar results, are unreliable for mirror misalignment described by actuator displacements exceeding one micron each, while SVD, on the other hand, is unreliable when the mirror misalignment is described by each actuator displacement exceeding five microns. From Figure 4.1, we see that such a restriction on actuator displacements corresponds to RMS actuator displacements of less than 0.65 micron. Such algorithmic restriction is consistent with RMS tip/tilt errors in acceptable range. If we suppose that at the start of each night's viewing, we rely on algorithmic control, then errors of actuator displacements that have accumulated over the day may only be controlled if by nightfall, the mirror misalignment is described by RMS actuator displacements corresponding to individual displacements of order one micron. If this is not the case, the algorithms will fail as will image quality. We note that for $z_{\max} = 10^{-4}$ metre, RMS actua-

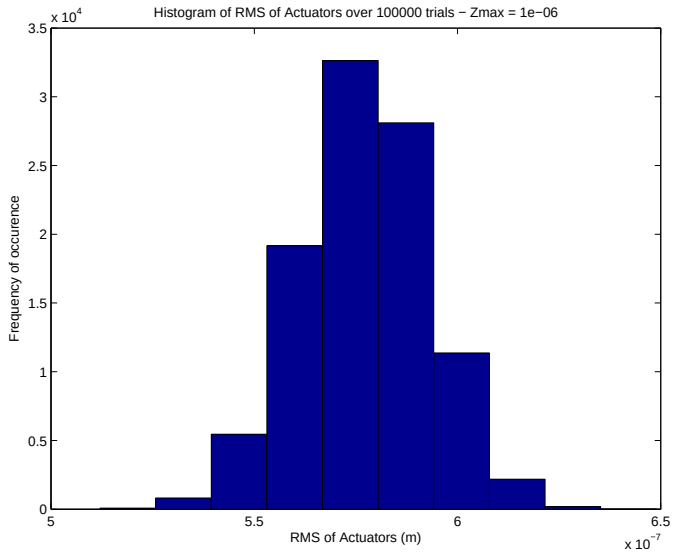

Figure 4.1: Histogram of RMS actuators for $z_{\max} = 10^{-6}$ metre: RMS actuator displacements varies between 0.5 and 0.65 microns with a distribution close to normal, and the primary mirror can be controlled with QR or SVD

tor displacements exceeds 50 microns, as is seen from Figure 4.2, and RMS tip/tilt is unacceptable as it exceeds 34 arcseconds (between 34 and 52 arcseconds, with a high concentration around 44 arcseconds, as is seen from Figure 4.3). Moreover, RMS tip/tilt errors, as illustrated in the histogram of Figure 4.4, is unacceptable as the probability for these errors to be within the 0.1 arcsecond range is very low (less than 1/10).

4.2 General Formulation of the Control Problem: Linear Quadratic Problems

To proceed, we must investigate control algorithms independent of the above methods. In the following, we will find that the so-called gradient flow and optimality condition methods can be applied. The former method will be favoured as it yields a stable control. It applies to the general case of linear quadratic problems which include the SALT control problem that can be formulated as given later in (4.1) or (4.2) in discrete time, and (4.3) or (4.4) in continuous time. However, we will later focus on formulation (4.4) for our approach to the solution of SALT problem. We will begin this section

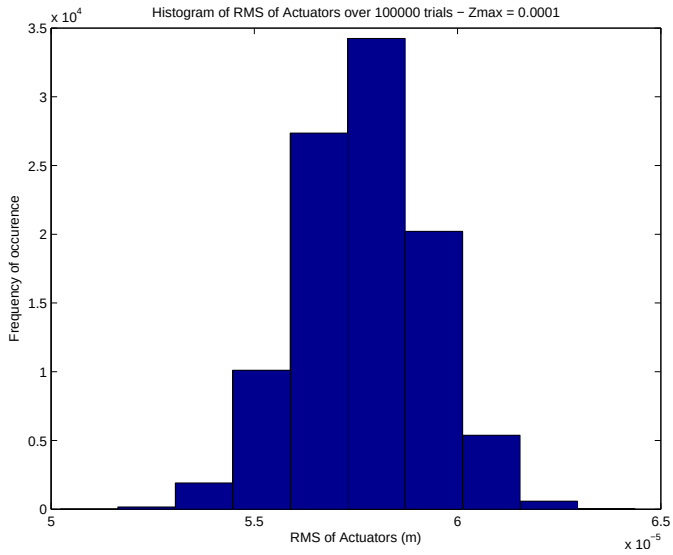

Figure 4.2: Histogram of RMS actuators for $z_{\max} = 10^{-4}$ metre: RMS actuator displacements varies between 50 and 65 microns with a distribution close to normal; this suggests a linear relationship between $z_{\max}$ and RMS actuator displacements; the primary mirror cannot be controlled with QR or SVD

with the general formulation of linear quadratic problems.

Linear quadratic problems [3, 22, 40, 43] are a specific case of optimal control problems mentioned in Chapter 1 (See page 8). They can be formulated in discrete time as well as in continuous time. When we don't have access to the values of the state variables at all time, the problem is classified as a linear quadratic problem with imperfect state information [3]. In this case, the optimal control, if it exists, is most likely an output feedback control.

4.2.1 Discrete time Formulation

We will only consider problems with a finite number of steps, that is, the discrete time version of finite horizon problems. However, problems with infinite number of steps are explored in [5] for nonlinear quadratic problems, where a state feedback control is investigated. Problems with perfect state

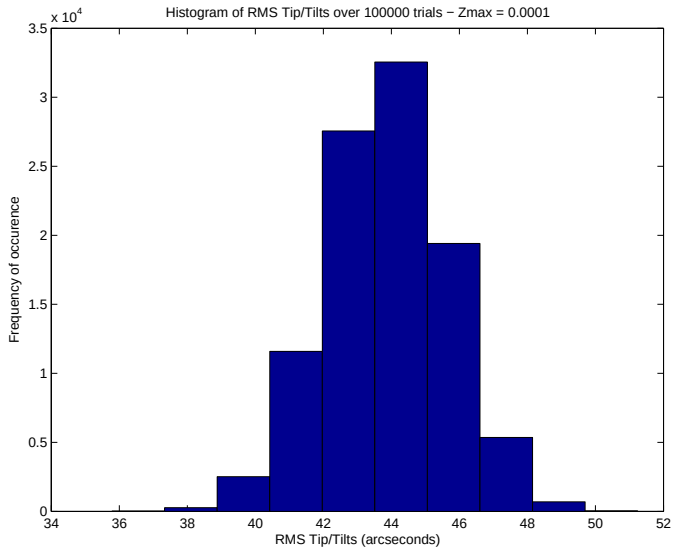

Figure 4.3: Histogram of RMS tip/tilt for $z_{\max} = 10^{-4}$ metre: this clearly shows that the primary mirror is far out of acceptable alignment since RMS tip/tilts is far greater than 0.1 arcsecond; this corresponds to an uncontrollable primary mirror configuration

information are formulated in general as follows:

$$\begin{aligned} &\min_{u} \left((z_N)^T Q_T z_N + \sum_{k=0}^{N-1} (z_k)^T Q z_k + (u_k)^T R u_k \right) \\ &\text{subject to } \begin{cases} z_{k+1} &= M z_k + N u_k \qquad (\text{and } z_0 \text{ is given}) \\ s_k &= A z_k \end{cases} \end{aligned} \tag{4.1}$$

On the other hand, problems with imperfect state information are formulated as follows:

$$\begin{aligned} &\min_{u} \left[\mathbb{E} \left\{ (z_N)^T Q_T z_N + \sum_{k=0}^{N-1} (z_k)^T Q z_k + (u_k)^T R u_k \right\} \right] \\ &\text{subject to } \begin{cases} z_{k+1} &= M z_k + N u_k \qquad (\text{and } z_0 \text{ is given}) \\ s_k &= A z_k \end{cases} \end{aligned} \tag{4.2}$$

where $\mathbb{E}$ stands for the mathematical expectation.

Both formulations (4.1) and (4.2) above are adapted from [3] where in both cases, M, N, Q and R depend on the step k. The adjustment of the

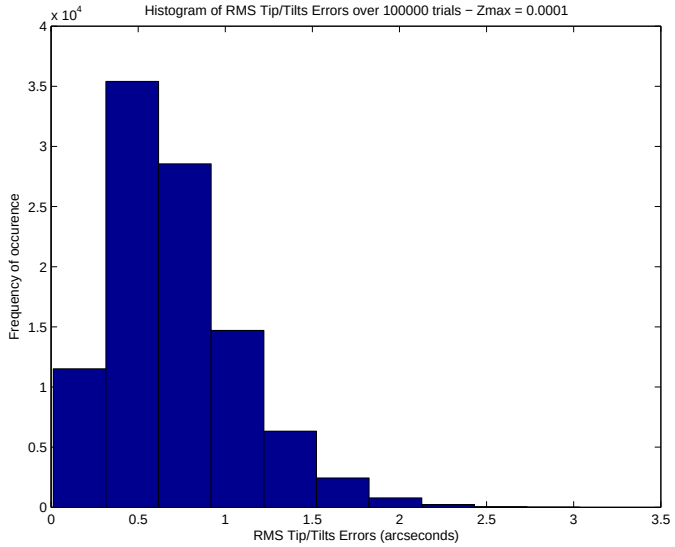

Figure 4.4: Histogram of RMS tip/tilt errors for $z_{\max} = 10^{-4}$ metre using SVD: this indicates that the primary mirror is not controllable using SVD since the probability for RMS tip/tilt errors to be within the 0.1 arcsecond is very low (less than 1/10)

state z (in the SALT case actuator displacements) from step k to step $k+1$ involves a disturbance w_k. In formulation (4.2), the relationship between the state z (in the SALT case actuator displacements) and the output s (in the SALT case relative heights) at step k involves an observation noise vector v_k with a known probability distribution. The matrix A in this relationship depends on k and is known for each value of k.

In both formulations (4.1) and (4.2), M is an $n \times n$ matrix, N is an $n \times p$ matrix, A is an $m \times n$ matrix, Q_T and Q are symmetric positive semidefinite matrices and R is a symmetric positive definite matrix. These linear quadratic problems can be solved using the discrete time dynamic programming approach [3].

Remark 4.1. Note that in the SALT case, $Q_T = Q = A^T A$, $R = \sigma I$ where A is the actuator-to-heights matrix, σ is a positive real number and I is the identity matrix of appropriate size. Also note that if $\sigma = 0$ in the SALT case, then the problem reduces to minimising the relative heights without considering how much effort (energy) is put in the process of controlling the system. The term $(u_k)^T R u_k$ is interpreted as a *potential energy* and we

thereby seek a *minimum energy* control.

Controllability (Kalman Condition)

- The controllability matrix of Problem (4.1) [it also applies to Problem (4.2)] is the $n \times np$ matrix

$$\mathcal{C} = \begin{bmatrix} N & MN & M^2N & \cdots & M^{n-1}N \end{bmatrix}$$

- Problem (4.1) [and similarly, Problem (4.2)] is controllable if its controllability matrix has rank n [40].

Output Controllability

- The output controllability matrix of Problem (4.2) [it also applies to Problem (4.1)] is the $m \times np$ matrix

$$\mathcal{OC} = \begin{bmatrix} AN & AMN & AM^2N & \cdots & AM^{n-1}N \end{bmatrix}$$

- Problem (4.2) [and similarly, Problem (4.1)] is output controllable if its output controllability matrix has rank m [40].

Remark 4.2. Note that since Remark 4.1 holds, due to the fact that A is 480×273 with rank 269, a simple substitution reveals that the SALT system is controllable but not output controllable. And since the control is performed based on the output, constraints are needed to satisfy output controllability.

4.2.2 Continuous time Formulation

There are different ways of formulating linear quadratic problems, depending on the time range, or if the objective function to minimise is an expected value. These (continuous time linear quadratic problems) can be solved using the *continuous time dynamic programming* approach when the optimal control is to be determined in the state feedback form.

Finite Horizon

We will consider problems of the form:

$$\begin{aligned} &\min_u J = z(T)^T Q_T z(T) + \int_0^T \left[z(t)^T Q z(t) + u(t)^T R u(t)\right] dt \\ &\text{subject to } \begin{cases} \dot{z}(t) = Mz(t) + Nu(t), \quad z(0) = z_0 \\ s(t) = Az(t) \end{cases} \end{aligned} \tag{4.3}$$

where Q and Q_T are symmetric positive semidefinite matrices, R is symmetric positive definite and A is of full rank. A being of full rank guarantees the equivalence $s(t) = Az(t) \iff z(t) = Bs(t)$ where B is the pseudo inverse of A.

Infinite Horizon

We will consider problems of the form:

$$\begin{aligned} &\min_u J = \mathbb{E}\left(\frac{1}{2}\int_0^\infty \left[z(t)^T Q z(t) + u(t)^T R u(t)\right] dt | z_0\right) \\ &\text{subject to } \begin{cases} \dot{z}(t) = Mz(t) + Nu(t), \quad z(0) = z_0 \\ s(t) = Az(t) \end{cases} \end{aligned} \tag{4.4}$$

where the assumptions are the same as in the finite horizon case but the final time is infinite.

4.3 Solution to Linear Quadratic Problems

4.3.1 In Discrete Time

There is a difference, depending on whether we are solving a problem with perfect state information, or a problem with imperfect state information. In both cases, we apply the discrete time dynamic programming approach [3].

Case with Perfect State Information

For Problem (4.1), as adapted from a similar formulation in [3] (where Q and R depend on the step k and the adjustment of the state z from step k to step $k+1$ involves a random disturbance), we define the *cost-to-go* function

as follows:

$$\begin{cases} J_N(z_N) = z_N^T Q_T z_N \\ J_i(z_i) = z_N^T Q_T z_N + \sum_{k=i}^{N-1} z_k^T Q z_k + u_k^T R u_k \quad \forall i \leq N-1 \end{cases} \tag{4.5}$$

and the optimal cost-to-go function by $J_i^*(z_i) = \min_{u_i} J_i(z_i)$. Then the *Bellman Equation of Dynamic Programming* for Problem (4.1) becomes

$$J_i^*(z_i) = \min_{u_i} \left[z_i^T Q z_i + u_i^T R u_i + J_{i+1}(M z_i + N u_i) \right] \quad \forall i \leq N-1 \tag{4.6}$$

Therefore we have:

$$\begin{aligned} J_{N-1}^*(z_{N-1}) &= \min_{u_{N-1}} \Big[z_{N-1}^T Q z_{N-1} + u_{N-1}^T R u_{N-1} \\ &\qquad + (M z_{N-1} + N u_{N-1})^T Q_T (M z_{N-1} + N u_{N-1}) \Big] \\ &= \min_{u_{N-1}} \left[u_{N-1}^T R u_{N-1} + u_{N-1}^T N^T Q_T N u_{N-1} + 2 z_{N-1}^T M^T Q_T N u_{N-1} \right] \\ &\quad + z_{N-1}^T Q z_{N-1} + z_{N-1}^T M^T Q_T M z_{N-1} \end{aligned} \tag{4.7}$$

The above minimisation problem yields

$$u_{N-1}^* = -(R + N^T Q_T N)^{-1} N^T Q_T M z_{N-1}$$

and then

$$\begin{aligned} J_{N-1}^*(z_{N-1}) &= z_{N-1}^T K_{N-1} z_{N-1} \quad \text{where} \\ K_{N-1} &= M^T \left(Q_T - Q_T N (N^T Q_T N + R)^{-1} N^T Q_T \right) M + Q \end{aligned} \tag{4.8}$$

and proceeding the same way for $k = N-2, N-3, \ldots, 1, 0$, we get

$$\begin{aligned} u_k^* = \mu_k^*(z_k) &= L_k z_k \\ L_k &= -(N^T K_{k+1} N + R)^{-1} N^T K_{k+1} M \\ K_N &= Q_T \\ K_k &= M^T \left(K_{k+1} - K_{k+1} N (N^T K_{k+1} N + R)^{-1} N^T K_{k+1} \right) M + Q \end{aligned} \tag{4.9}$$

then $J_k^*(z_k) = z_k^T K_k z_k$ and therefore $J^* = J_0^*(z_0) = z_0^T K_0 z_0$ where K_0 can be obtained by solving the above *matrix Riccati difference equation*. Note that the above matrix Riccati difference equation can be solved only if the matrix $N^T K_{k+1} N + R$ is nonsingular $\forall k \geq 0$.

Case with Imperfect State Information

This case is similar to the perfect state information case, except that at stage k, the controller does not have access to the value of the state variable z_k, but has the observed values s_k linked to z_k by the relation $s_k = Az_k$. This case is closer to the SALT case than the perfect state information case. Instead of Problem (4.1), the problem to solve is Problem (4.2), adapted from a similar formulation in [3] where Q and R depend on k, the adjustment in z involves a random disturbance w_k at step k, the relationship between the state z (actuator displacements) and the output s (relative heights) at each step k involves an observation noise vector v_k with a known probability distribution, and the A matrix in that relationship (the actuator-to-heights matrix) also depends on k, and is known for each value of k. Using the same (discrete time dynamic programming) process as in [3], we define the *information vector* as follows: $I_0 = s_0$ and $\forall k \geq 1,\ I_k = (s_0, s_1, \ldots, s_k, u_0, u_1, \ldots, u_{k-1})$. Note that $\forall k \geq 0,\ I_{k+1} = (I_k, s_{k+1}, u_k)$. The cost-to-go function is now defined as follows:

$$\begin{cases} J_N(I_N) = \underset{z_N}{\mathbb{E}} \left(z_N^T Q_T z_N | I_N \right) \\ J_i(I_i) = \underset{z_i}{\mathbb{E}} \left(z_N^T Q_T z_N + \sum_{k=i}^{N-1} z_k^T Q z_k + u_k^T R u_k | I_i, u_i \right) \quad \forall i \leq N-1 \end{cases} \tag{4.10}$$

and the optimal cost-to-go function by $J_i^*(I_i) = \min_{u_i} J_i(I_i)$. The Bellman Equation for Problem (4.2) becomes

$$J_i^*(I_i) = \min_{u_i} \left[\underset{z_i, s_{i+1}}{\mathbb{E}} \left\{ z_i^T Q z_i + u_i^T R u_i + J_{i+1}(I_{i+1}) | I_i, u_i \right\} \right] \tag{4.11}$$

Therefore we have:

$$\begin{aligned} J_{N-1}^*(I_{N-1}) = \min_{u_{N-1}} \Bigg[& \underset{z_{N-1}}{\mathbb{E}} \Big\{ z_{N-1}^T Q z_{N-1} + u_{N-1}^T R u_{N-1} \\ & + (M z_{N-1} + N u_{N-1})^T Q_T (M z_{N-1} + N u_{N-1}) | I_{N-1} \Big\} \Bigg] \\ = & \underset{z_{N-1}}{\mathbb{E}} \left[z_{N-1}^T (M^T Q_T M + Q) z_{N-1} | I_{N-1} \right] \\ & + \min_{u_{N-1}} \left[u_{N-1}^T (N^T Q_T N + R) u_{N-1} + 2\mathbb{E}(z_{N-1}|I_{N-1})^T M^T Q_T N u_{N-1} \right] \end{aligned} \tag{4.12}$$

The above minimisation problem yields

$$u_{N-1}^* = -(N^T Q_T N + R)^{-1} N^T Q_T M \mathbb{E}(z_{N-1}|I_{N-1})$$

and then

$$J^*_{N-1}(I_{N-1}) = \underset{z_{N-1}}{\mathbb{E}} \left(z^T_{N-1} K_{N-1} z_{N-1} | I_{N-1}\right) + \underset{z_{N-1}}{\mathbb{E}} \left[\left(z_{N-1} - \mathbb{E}\{z_{N-1}|I_{N-1}\}\right)^T P_{N-1} \left(z_{N-1} - \mathbb{E}\{z_{N-1}|I_{N-1}\}\right) | I_{N-1} \right] \tag{4.13}$$

where K_{N-1} and P_{N-1} are given by:

$$\begin{aligned} P_{N-1} &= M^T Q_T N \left(R + N^T Q_T N\right)^{-1} N^T Q_T M \\ K_{N-1} &= M^T Q_T M - P_{N-1} + Q. \end{aligned}$$

Moreover,

$$\begin{aligned} J^*_{N-2}(I_{N-2}) &= \min_{u_{N-2}} \left[\underset{z_{N-2}, s_{N-1}}{\mathbb{E}} \left\{ z^T_{N-2} Q z_{N-2} + u^T_{N-2} R u_{N-2} \right.\right. \\ &\qquad \left.\left. + J_{N-1}(I_{N-1}) | I_{N-2}, u_{N-2} \right\} \right] \\ &= \mathbb{E} \left[\left(z_{N-1} - \mathbb{E}\{z_{N-1}|I_{N-1}\}\right)^T P_{N-1} \left(z_{N-1} - \mathbb{E}\{z_{N-1}|I_{N-1}\}\right) | I_{N-2}, u_{N-2} \right] \\ &+ \underset{z_{N-2}}{\mathbb{E}} \left(z^T_{N-2} Q z_{N-2} | I_{N-2}\right) \\ &+ \min_{u_{N-2}} \left[u^T_{N-2} R u_{N-2} + \mathbb{E}\left(z^T_{N-1} K_{N-1} z_{N-1} | I_{N-2}, u_{N-2}\right) \right] \end{aligned} \tag{4.14}$$

The above minimisation problem yields

$$u^*_{N-2} = -\left(R + N^T K_{N-1} N\right)^{-1} N^T K_{N-1} M \mathbb{E}\left(z_{N-2} | I_{N-2}\right).$$

In a similar way for lower values of k, we have:

$$\begin{aligned} u^*_k = \mu^*_k(I_k) &= L_k \mathbb{E}\left(z_k | I_k\right) \\ L_k &= -\left(R + N^T K_{k+1} N\right)^{-1} N^T K_{k+1} M \\ K_N &= Q_T \\ K_k &= M^T K_{k+1} M - P_k + Q \\ P_k &= M^T K_{k+1} N \left(R + N^T K_{k+1} N\right)^{-1} N^T K_{k+1} M \end{aligned} \tag{4.15}$$

Remark 4.3. Note that since Remark 4.1 holds, the expression of the solution given above can be rewritten in a simpler form:

$$\begin{aligned} u^*_k = \mu^*_k(I_k) &= L_k \mathbb{E}\left(z_k | I_k\right) \\ L_k &= -\left(R + K_{k+1}\right)^{-1} K_{k+1} \\ K_N &= Q_T \\ K_k &= K_{k+1} - P_k + Q \\ P_k &= K_{k+1} \left(R + K_{k+1}\right)^{-1} K_{k+1} \end{aligned} \tag{4.16}$$

4.3.2 In Continuous Time

Finite Horizon

State Feedback The goal is to find an optimal control $u^* = U(t)z(t)$ for Problem (4.3) where the *gain matrix* $U(t)$ is to be determined. Note that in the SALT case, $z(t)$ is the vector of actuator positions at time t.

Theorem 4.4 below (see [32, 33, 40]) is a characterisation of controllable linear quadratic systems:

Theorem 4.4 (Kalman Condition)**.** *The system* (4.3) *is controllable if and only if its* controllability matrix *has rank* n*, the controllability matrix being the* $n \times np$ *matrix*

$$\mathcal{C} = \begin{bmatrix} N & MN & M^2N & \cdots & M^{n-1}N \end{bmatrix}.$$

Theorem 4.5 below [22] indicates how to determine the optimal state feedback control of Problem (4.3).

Theorem 4.5. *The following statements hold:*

- *The optimal control* u^* *of Problem* (4.3) *is given by*

$$u^*(t) = -R^{-1}N^T K(t) z^*(t)$$

  where K *is a symmetric matrix that solves the* matrix Riccati differential equation

$$\begin{aligned} &\dot{K}(t) = -K(t)M - M^T K(t) + K(t)NR^{-1}N^T K(t) - Q \\ &K(T) = Q_T \end{aligned} \tag{4.17}$$

- *The optimal response (optimal trajectory)* z^* *satisfies*

$$\dot{z}^*(t) = \left(M - NR^{-1}N^T K(t)\right) z^*(t), \qquad z^*(0) = z_0 \tag{4.18}$$

- *The optimal cost is*

$$J(u^*) = \frac{1}{2} z_0^T K(0) z_0 \tag{4.19}$$

Output Feedback The goal is to find an optimal control $u^* = -F(t)s(t)$ for Problem (4.3) where the *gain matrix* $F(t)$ is to be determined. Note that in the SALT case, $s(t)$ is the vector of relative heights at time t.

Theorem 4.6 below (see [32, 33, 40]) characterises output controllable linear quadratic systems.

Theorem 4.6. *The system* (4.3) *is output controllable if and only if its output controllability matrix has rank* m*, the output controllability matrix being the* $m \times np$ *matrix*

$$\mathcal{OC} = \begin{bmatrix} AN & AMN & AM^2N & \cdots & AM^{n-1}N \end{bmatrix}.$$

Remark 4.7. We suggest a few options to find the optimal control of Problem (4.3) in the output feedback form.

1. A simple option is to consider $z(t)$ in terms of $s(t)$ using the least squares approach. Considering the pseudo inverse of A denoted by B, from $s(t) = Az(t)$ we have $z(t) = Bs(t)$ and the new expression of u^* becomes $u^*(t) = R^{-1}K(t)Bs(t)$ with the same conditions on K as in Equation (4.17) above.

2. Another approach is to transform the problem into an equivalent problem, with s as the new state variable. Still considering the fact that $z(t) = Bs(t)$, and that $\dot{s}(t) = A\dot{z}(t)$ from $s(t) = Az(t)$, Problem (4.3) becomes

$$\begin{aligned} &\min_u J = s(T)^T B^T Q_T B s(T) + \int_0^T \left[s(t)^T B^T Q B s(t) + u(t)^T R u(t) \right] dt \\ &\text{subject to} \quad \dot{s}(t) = AMBs(t) + ANu(t), \quad s(0) = Az_0 \end{aligned} \tag{4.20}$$

   Let $\widetilde{Q}_T = B^T Q_T B$, $\widetilde{Q} = B^T Q B$, $\widetilde{R} = R$, $\widetilde{M} = AMB$ and $\widetilde{N} = AN$. Then Problem (4.3) becomes

$$\begin{aligned} &\min_u J = s(T)^T \widetilde{Q}_T s(T) + \int_0^T \left[s(t)^T \widetilde{Q} s(t) + u(t)^T \widetilde{R} u(t) \right] dt \\ &\text{subject to} \quad \dot{s}(t) = \widetilde{M} s(t) + \widetilde{N} u(t), \quad s(0) = Az_0 \end{aligned} \tag{4.21}$$

   Note that if Q and Q_T are symmetric positive semidefinite, so are $\widetilde{Q}$ and $\widetilde{Q}_T$; and if R is symmetric positive definite, so is $\widetilde{R}$. Therefore we

can apply continuous time dynamic programming to Problem (4.21) to have the solution $u^*(t) = \widetilde{R}^{-1}\widetilde{K}(t)s(t)$ where $\widetilde{K}$ is a symmetric matrix that solves the matrix Riccati differential equation

$$\begin{aligned} &\dot{\widetilde{K}}(t) + \widetilde{K}(t)\widetilde{N}\widetilde{R}^{-1}\widetilde{N}^T\widetilde{K}(t) + \widetilde{K}(t)\widetilde{M} + \widetilde{M}^T\widetilde{K}(t) - \widetilde{Q} = 0 \\ &\widetilde{K}(T) = -\widetilde{Q}_T \end{aligned} \tag{4.22}$$

3. Note that the two options explored above are valid only when A has full rank.

Infinite Horizon

State Feedback The goal is to find an optimal control $u^* = Uz(t)$ for Problem (4.4) where the *gain matrix* U is to be determined. Note that in the SALT case, $z(t)$ is the vector of actuator positions at time t.

The controllability of Problem (4.4) is valid under the conditions given in Theorem 4.4.

Theorem 4.8 below [22] indicates how to determine the optimal state feedback control of Problem (4.4).

Theorem 4.8. *We assume the system in Problem* (4.4) *is controllable. Then*

- *The optimal control u^* is given by $u^*(t) = -R^{-1}N^TKz^*(t)$ where K is the unique symmetric positive definite matrix that solves the* matrix continuous time algebraic Riccati equation

$$-M^TK - KM + KNR^{-1}N^TK = Q \tag{4.23}$$

- *The optimal response (optimal trajectory) z^* satisfies*

$$\dot{z}^*(t) = \left(M - NR^{-1}N^TK\right)z^*(t), \qquad z^*(0) = z_0 \tag{4.24}$$

- *The optimal cost is*

$$J(u^*) = \frac{1}{2}z_0^TKz_0 \tag{4.25}$$

Output Feedback The goal is to find an optimal control $u^* = -Fs(t)$ for Problem (4.4) where the *gain matrix* F is to be determined. Note that in the SALT case, $s(t)$ is the vector of relative heights at time t.

The output controllability of Problem (4.4) is valid under the conditions given in Theorem 4.6.

Remark 4.9. Similar to Remark 4.7, under the assumption that A has full rank and that B is the pseudo inverse of A, a few suggestions for an optimal output feedback control for Problem (4.4) are:

1. $u^*(t) = -R^{-1}N^TKBs^*(t)$ where s^* is the optimal trajectory for s.

2. Problem (4.4) can also be rewritten as

$$\begin{aligned} &\min_u J = \mathbb{E}\left(\frac{1}{2}\int_0^\infty \left[s(t)^T\widetilde{Q}s(t) + u(t)^T\widetilde{R}u(t)\right]dt|s_0\right) \\ &\text{subject to } \dot{s}(t) = \widetilde{M}z(t) + \widetilde{N}u(t), \quad s(0) = s_0 = Az_0 \end{aligned} \tag{4.26}$$

   where $\widetilde{Q} = B^TQB$, $\widetilde{R} = R$, $\widetilde{M} = AMB$ and $\widetilde{N} = AN$. Note that if Problem (4.4) is output controllable, then Problem (4.26) is controllable. And by applying the continuous time dynamic programming technique to Problem (4.26), we obtain $u^*(t) = -\widetilde{R}^{-1}\widetilde{N}^T\widetilde{K}s^*(t)$ where $\widetilde{K}$ is the unique symmetric positive definite matrix that solves the matrix continuous time algebraic Riccati equation

$$-\widetilde{M}^T\widetilde{K} - \widetilde{K}\widetilde{M} + \widetilde{K}\widetilde{N}\widetilde{R}^{-1}\widetilde{N}^T\widetilde{K} = \widetilde{Q}. \tag{4.27}$$

   In this case, the optimal cost is

$$J(u^*) = \frac{1}{2}\mathbb{E}\left(s_0^T\widetilde{K}s_0\right). \tag{4.28}$$

4.3.3 A new Approach to Problem (4.4)

Preliminary Result from Literature

When in Problem (4.4), A is a full rank $m \times n$ matrix with $m \leq n$, the following result applies.

Theorem 4.10 (See [24])**.** *Assuming that the initial state z_0 is a random vector uniformly distributed on the surface of the n-dimensional unit sphere, the optimal output feedback gain is given by*

$$F = R^{-1}N^TK\lambda A^T(A\lambda A^T)^{-1} \tag{4.29}$$

where

$$K = \int_0^\infty e^{M_0^Ts}\left[Q + A^TF^TRFA\right]e^{M_0s}ds \tag{4.30}$$

$$\lambda = \int_0^\infty e^{M_0s}e^{M_0^Ts}ds \tag{4.31}$$

provided $M_0 = M - NFA$ is stable.

Remark 4.11. The K and λ from Theorem 4.10 are solutions to the following equations:

$$KM_0 + M_0^T K + Q + A^T F^T RFA = 0 \tag{4.32}$$

$$\lambda M_0^T + M_0 \lambda + I_n = 0 \tag{4.33}$$

The Gradient Flow Approach

The gradient flow technique is a relatively recent mathematical technique. This technique has been applied to a few classes of optimal control problems, including nonlinear quadratic optimal control problems in discrete time, linear quadratic optimal control problems in continuous time with stochastic jump parameters [5, 44, 47, 48]. This technique can be adapted to the problems under study, provided we are dealing with infinite horizon problems, in discrete or in continuous time. This method is stable and robust with respect to observation errors (or measurement errors), provided the probability distribution of the error term is known. The main idea behind the gradient flow approach is to transform an optimal control problem (in continuous time) into an ordinary differential equation problem whereby solving the ODE gives the solution to the original optimal control problem. A standard formulation of a linear output feedback optimal control problem has the form

$$\begin{aligned} \min_{u} \quad & J\left(t, x(t), u(t)\right) \\ \text{subject to} \quad & \begin{cases} \dot{x}(t) = Ax(t) + Bu(t); \quad x(0) = x_0 \\ y(t) = Cx(t) \\ u(t) = -Fy(t) \end{cases} \end{aligned} \tag{4.34}$$

In this formulation, x is the state variable; y is the output variable; u is the control variable; the function J to minimise is called the objective function; C is the interaction matrix (relationship between the state and the output), and F is the linear output feedback gain matrix. The solution to the original optimal control problem is entirely determined by the computation of F. The gradient flow algorithm determines the F matrix by the addition of a differential equation for F, of the form

$$\dot{F} = -\frac{\partial J}{\partial F} \tag{4.35}$$

called the gradient flow associated with the objective function J. This is done after J is made a function of F by a transformation

$$J\left(t, x(t), u(t)\right) \to J(F, P) \tag{4.36}$$

where $P = \mathbb{E}\left(x_0 x_0^T\right)$, in our case, as in equations (4.39) below, such that $\dot{F} \to 0$ as $t \to \infty$. Intuitively, we place the gain F of the standard control in a potential well defined on J. Clearly, equation (4.35) finds the value of F that minimises J. The solution to the new problem gives us the solution to the original problem. Note that given our original optimal control problem, computation of F is executed, and then holds throughout the standard control process (4.34). Moreover, the transformation (4.36) can always be found. Furthermore, we can ensure controllability of (4.34), given F (see Theorem 4.18 below). Gradient flow is then no more expensive than other methods, and is robust.

Problem Transcription Note that the goal is to find a linear output feedback optimal control, that is, an optimal control in the form $u^*(t) = -Fs(t)$ that minimises the objective function in Problem (4.4), where F is the gain matrix to be determined. This will be adapted to the SALT case where $s(t)$ is the vector of relative heights at time t, also known in the context of control theory as the output variable. Recall that $z(t)$ is the vector of actuator displacements at time t, also known in the context of control theory as the state variable. The result given below in Lemma 4.12 is inspired by a similar result from [47].

Lemma 4.12. *The index function given in* (4.4) *can be reduced to*

$$J = \frac{1}{2} trace\left(KP^T\right) \tag{4.37}$$

where K is the unique positive definite solution to the following Lyapunov equation

$$\begin{aligned} &KM_0 + M_0^T K + Q + A^T F^T RFA = 0 \\ &M_0 = M - NFA \\ &P = \mathbb{E}\left(z_0 z_0^T\right) \end{aligned} \tag{4.38}$$

provided M_0 is stable.

Proof. We have

$$\dot{z}(t) = Mz(t) + Nu(t) = (M - NFA)z(t) = M_0 z(t).$$

Let $v(t) = z^T(t)Kz(t)$ and $\bar{v}(t) = \mathbb{E}\{v(t)\} = \mathbb{E}\left\{z^T(t)Kz(t)\right\}$. Then

$$\begin{aligned}\dot{\bar{v}}(t) &= \frac{d}{dt}\mathbb{E}\left\{z^T(t)Kz(t)\right\} \\ &= \mathbb{E}\left\{\dot{z}^T(t)Kz(t) + z^T(t)K\dot{z}(t) + z^T(t)\dot{K}z(t)\right\} \\ &= \mathbb{E}\left\{z^T(t)\left(M_0^TK + KM_0 + \dot{K}\right)z(t)\right\};\end{aligned}$$

but $\dot{K} = 0$, thus

$$\dot{\bar{v}}(t) = \mathbb{E}\left\{z^T(t)\left(M_0^TK + KM_0\right)z(t)\right\}.$$

So,

$$\begin{aligned}J(F,P) &= \mathbb{E}\left\{\frac{1}{2}\int_0^\infty \left[z^T(s)Qz(s) + u^T(s)Ru(s)\right]ds|P\right\} \\ &= \mathbb{E}\left\{\frac{1}{2}\int_0^\infty \left[z^T(s)Qz(s) + \left(FAz(s)\right)^T R\left(FAz(s)\right)\right]ds|P\right\} \\ &= \mathbb{E}\left\{\frac{1}{2}\int_0^\infty \left[z^T(s)Qz(s) + z^T(s)A^TF^TRFAz(s)\right]ds|P\right\} \\ &= \mathbb{E}\left\{\frac{1}{2}\int_0^\infty \left[z^T(s)\left(Q + A^TF^TRFA\right)z(s)\right]ds|P\right\} \\ &= \mathbb{E}\left\{\frac{1}{2}\int_0^\infty \left[z^T(s)\left(-KM_0 - M_0^TK\right)z(s)\right]ds|P\right\} \\ &= \mathbb{E}\left\{\frac{1}{2}\int_0^\infty -\dot{v}(s)ds|P\right\} \\ &= \frac{1}{2}\mathbb{E}\left\{z_0^TKz_0\right\} - \lim_{t\to\infty}\mathbb{E}\left\{v(t)|P\right\} \\ &= \frac{1}{2}\text{trace}\left[K\left(\mathbb{E}\left\{z_0z_0^T\right\}\right)^T\right] - \lim_{t\to\infty}\mathbb{E}\left\{v(t)|P\right\} \\ &= \frac{1}{2}\text{trace}\left(KP^T\right) - \lim_{t\to\infty}\mathbb{E}\left\{v(t)|P\right\}\end{aligned}$$

But $\dot{z}(t) = M_0z(t)$, $z(0) = z_0$ and M_0 stable imply $\lim_{t\to\infty} z(t) = 0$ and therefore $\lim_{t\to\infty} v(t) = 0$ since $v(t) = z^T(t)Kz(t)$. Henceforth $J(F,P) = \frac{1}{2}\text{trace}\left(KP^T\right)$. □

Remark 4.13. Note that

- A sufficient condition for M_0 to be stable is for $\Lambda = -\frac{1}{2}\left(M_0 + M_0^T\right)$ to be positive definite.
- From *Sylvester's criterion*, Λ is positive definite if and only if the determinants g_j of all its leading principal minors are such that $g_j \geq \varepsilon$ for a small positive real number ε.

We have $J = J(F, P)$; let $\Xi = \{F \in \mathbb{R}^{m\times r} \text{ such that } M_0 \text{ is stable}\}$.

Lemma 4.14 (See [47])**.** *Assume P is positive definite. Then*

$$S(\eta) = \{F \in \Xi \text{ such that } J(F, P) \leq \eta\}$$

is compact, $\forall \eta \geq 0$

Proof. Similar to that of Lemma 1 part (ii) in [48] □

Remark 4.15. From Lemmas 4.12 and 4.14, it can be established [47] that the set $\bar{S} = \bigcup_{\eta\geq 0} S(\eta)$ is open in $\mathbb{R}^{m\times r}$. As a result, all global and local minima of $J(F, P)$ are interior points of $\bar{S}$. In addition, $F \in \bar{S}$ if $M_0 = M - NFA$ is stable.

Our original problem formulated in (4.4) then becomes

$$\min J(F, P) = \frac{1}{2}\text{trace}\left(KP^T\right)$$
$$\text{subject to } \begin{cases} g_j \geq \varepsilon, \quad \varepsilon > 0 \\ KM_0 + M_0^T K + Q + A^T F^T RFA = 0 \\ M_0 = M - NFA \\ P = \mathbb{E}\left(z_0 z_0^T\right) \end{cases} \tag{4.39}$$

Note that if $P = \mathbb{E}\left(z_0 z_0^T\right)$ is known and z_0 is not, this is a generalisation of Problem (4.4). Also note that the gradient flow approach gives the optimal control in terms of the output while the dynamic programming approach gives the optimal control in terms of the *expected value* of the state.

We define the *Hamiltonian* as follows:

$$H = \frac{1}{2}\text{trace}\left(KP^T\right) + \text{trace}\left[\lambda(KM_0 + M_0^T K + Q + A^T F^T RFA)\right] \tag{4.40}$$

where the co-state λ is an $n \times n$ symmetric matrix satisfying the adjoint equation $\frac{\partial H}{\partial K} = 0 \iff \frac{1}{2}P + \lambda M_0^T + M_0\lambda = 0$; this follows from the following property of matrices: $\frac{\partial}{\partial A}\text{trace}(AB) = B^T$, and some basic properties of the matrix trace.

Theorem 4.16 below (inspired by a similar result in [47]) shows how to compute the gradient of J (and similarly the gradient of every g_j) with respect to F.

Theorem 4.16. *The gradient of J with respect to F is given by*

$$\frac{\partial J}{\partial F} = \frac{\partial H}{\partial F} = 2RFA\lambda A^T - 2N^T K\lambda A^T$$

where K and λ satisfy the following Lyapunov equations:

$$KM_0 + M_0^T K + Q + A^T F^T RFA = 0 \tag{4.41}$$

$$\frac{1}{2}P + \lambda M_0^T + M_0\lambda = 0 \tag{4.42}$$

Proof. Similar to that of Theorem 3.1 in [46]. □

Remark 4.17. When A is a full rank $m \times n$ matrix with $m \leq n$, solving the equation $\frac{\partial J}{\partial F} = 0$ gives $F = R^{-1}N^T K\lambda A^T(A\lambda A^T)^{-1}$ as in Theorem 4.10.

The gradient flow associated with $J = J(F, P)$ is

$$\dot{F} = -\frac{\partial J}{\partial F} = -\frac{\partial H}{\partial F} = 2N^T K\lambda A^T - 2RFA\lambda A^T \tag{4.43}$$

So any global minimum must be an equilibrium of (4.43). The following theorem (Theorem 4.18, see [47]) gives several properties associated with the gradient flow (4.43).

Theorem 4.18. *Given the initial condition $F(0) = F^0$ such that $F^0 \in \bar{S}$*

1. *The gradient flow* (4.43) *has a unique solution $F(t) \in \bar{S}$ defined on $[0, \infty)$.*

2. *The index function $J(F)$ is non-increasing along the solution $F(t)$ with*

$$J\left(F(t)\right) = J\left(F^0\right) - \int_0^t \|\dot{F}(t)\|_F^2 dt \tag{4.44}$$

   where $\|\cdot\|_F$ is the Frobenius norm defined by $\|A\|_F^2 = trace\left(A^T A\right)$.

3. $\lim_{t\to\infty} \dot{F}(t) = 0$.

4. *There exists a convergent subsequence of $\{F(t)\}$ as $t \to \infty$ and any such subsequence converges to an equilibrium of* (4.43) *in $\bar{S}$.*

Proof. Similar to that of Theorem 2.1 in [48]. □

Remark 4.19. Solving our approximate problem involves solving the system (4.41) - (4.42) - (4.43), but we don't have the initial condition $F(0)$ yet.

Determining $F(0)$. Sylvester's criterion says we must find an F such that $M_0 = M - NFA$ is stable, that is, all the leading principal minors of $\Lambda = -\frac{1}{2}\left(M_0 + M_0^T\right)$ are greater or equal to a small positive real number ε_0. If Λ is an $n \times n$ matrix, from [18], a constraint transcription of Sylvester's criterion is as follows:

$$\min J = \sum_{i=1}^{n} g_i(F) \tag{4.45}$$

where $g_i(F) = \varphi_\varepsilon\left(\det\left(\Lambda_i\right) - \varepsilon_0\right)$ with Λ_i standing for the top left corner $i \times i$ sub-matrix of Λ and φ_ε defined on the set $\mathbb{R}$ of real numbers (and actually a *smoothing* of the function $x \mapsto \max(-x, 0)$ around $x_0 = 0$) as follows:

$$\varphi_\varepsilon\left(x\right) = \begin{cases} -x & \text{if } x < -\varepsilon \\ 0 & \text{if } x > \varepsilon \\ \frac{(x-\varepsilon)^2}{4\varepsilon} & \text{if } -\varepsilon \leq x \leq \varepsilon \end{cases}$$

Problem (4.45) can be solved using any unconstrained optimisation technique, such as the Quasi-Newton method. Note that $F \mapsto M_0 = M - NFA$, $M_0 \mapsto \Lambda = -\frac{1}{2}\left(M_0 + M_0^T\right)$, $\Lambda \mapsto \Lambda_i = P^T \Lambda P$, $\Lambda_i \mapsto y_i = \det\left(\Lambda_i\right) - \varepsilon_0$ and $y_i \mapsto z_i = \varphi_\varepsilon\left(y_i\right)$, where P is the $n \times i$ matrix with one on the main diagonal and zero everywhere else. An intuitive idea is to use the chain rule to explicitly find the gradient of the map $F \mapsto z_i$. This, however, might not be recommended for large scale problems. Moreover, since the best approach to solving system (4.41)-(4.42)-(4.43) is the numerical approach, a better idea is to arbitrarily choose $F(0)$ in such a way that $M_0 = M - NF(0)A$ is stable. A few algorithms to solve our new problem are investigated below and are based on the following lemma (lemma 4.20) that assumes the initial state is a random vector uniformly distributed on the surface of the n-dimensional unit sphere, but can be adapted to a more general case where the initial state is not subject to this restriction.

Lemma 4.20 (See [24]). *For any positive integer n, let F_{n-1} be the solution of (4.29) with $K = K_{n-1}$ and $\lambda = \lambda_{n-1}$, that is,*

$$F_{n-1} = R^{-1} N^T K_{n-1} \lambda_{n-1} A^T (A \lambda_{n-1} A^T)^{-1} \tag{4.46}$$

where K_n is the solution of (4.41) with $F = F_{n-1}$, that is,

$$K_n(M - NF_{n-1}A) + (M - NF_{n-1}A)^T K_n + Q + A^T F_{n-1}^T R F_{n-1} A = 0 \tag{4.47}$$

and λ_{n-1} *is the solution of* (4.42) *with* $F = F_{n-1}$*, that is,*

$$\lambda_{n-1}(M - NF_{n-1}A)^T + (M - NF_{n-1}A)\lambda_{n-1} + I = 0 \tag{4.48}$$

1. *Then, assuming* Q *is positive definite and* $M - NF_{n-1}A$ *is stable, a unique and positive definite* K_n *exists.*

2. *Furthermore, assuming there exists a positive definite solution* λ_{n-1} *to equation* (4.48)*, then*

$$trace(K_n) \leq trace(K_{n-1}) \tag{4.49}$$

Now we give three algorithms that can be independently used to solve our new problem.

Algorithm 1 (See [26])

1. Choose K_0 arbitrarily (may be from the full state feedback solution, or solve equation (4.41) using the F obtained from solving the equation $M - NFA = M_s$ where M_s is an arbitrarily chosen known stable matrix)

2. At step n:

   (a) Solve (4.29) and (4.42) simultaneously in λ_{n+1} and F_{n+1} for fixed K_n

   (b) Actualize K_n into K_{n+1} where K_{n+1} is the solution of (4.41) for fixed λ_{n+1} and F_{n+1}

3. Iterate step 2 until convergence

Algorithm 1 is computationally expensive and is not considered hereafter. The next algorithm is the so-called *optimality condition* algorithm and will be implemented below

Algorithm 2 (Optimality Condition - See [26])

1. Choose F_0 arbitrarily such that $M - NF_0A$ is stable

2. At step n:

   (a) Solve (4.41) in K_{n+1} for fixed F_n

(b) Solve (4.42) in λ_{n+1} for fixed F_n

(c) Actualize F_n into F_{n+1} through (4.29) for fixed λ_{n+1} and K_{n+1}

3. Iterate step 2 until convergence

Algorithm 3: (The Gradient Flow Algorithm - See [47])

1. Choose F_0 arbitrarily such that $M - NF_0A$ is stable

2. At step n:

(a) Solve (4.41) in K_{n+1} for fixed F_n

(b) Solve (4.42) in λ_{n+1} for fixed F_n

(c) Actualize F_n into F_{n+1} through (4.43) for fixed λ_{n+1} and K_{n+1}

3. Iterate step 2 until convergence

Remark 4.21. Since Algorithm 1 is computationally expensive, we will be more interested in algorithms 2 and 3.

If the results in this section (section 4.3.3) applied to the SALT case, a straightforward substitution would give us the results we need. Unfortunately, in the SALT case, the A matrix is a rank deficient $m \times n$ matrix with $m > n$, precisely, $m = 480$ and $n = 273$, with rank 269. Thus we need to transform the initial problem to meet the requirements in this section. This process is described in the next section.

4.4 A Formulation of Linear Quadratic Problems in the case $m \geq n$

We consider Problem (4.4) with $m \geq n$ and A not necessarily a full rank matrix. We use the SVD of A to simplify the problem by reformulating it in the *mode space*. We have $s = Az = U\Sigma V^T z \iff U^T s = \Sigma V^T z$, since by definition of SVD (Theorem 3.13), U and V are square unitary matrices, that is, $U^T U = I$ and $V^T V = I$ where I is the identity matrix of appropriate size. Now let $\widetilde{z} = V^T z$ and $\widetilde{s} = U^T s$. This is equivalent to $z = V\widetilde{z}$ and $s = U\widetilde{s}$. Then we have $\widetilde{s} = \Sigma\widetilde{z}$. We also have $z^T Q z = \widetilde{z}^T V^T Q V \widetilde{z}$. Moreover, $\dot{z} = Mz + Nu \iff V^T\dot{z} = V^T Mz + V^T Nu \iff \dot{\widetilde{z}} = V^T MV\widetilde{z} + V^T Nu$.

But the output s is in the range of A. Hence $\widetilde{s}_i = 0$ for all $i > r$, where r is the rank of A. So $\widetilde{s} = \Sigma\widetilde{z}$ can be rewritten as

$$\begin{bmatrix} \widetilde{s}_c \\ 0 \end{bmatrix} = \begin{bmatrix} \Sigma_r \\ 0 \end{bmatrix} \widetilde{z} \iff \widetilde{s}_c = \Sigma_r \widetilde{z}$$

where Σ_r is the top sub-matrix of Σ containing the r rows with nonzero singular values. Note that Σ_r is a full rank $r \times n$ matrix with $r \leq n$. Also note that

$$\widetilde{s} = U^T s \iff \begin{bmatrix} \widetilde{s}_c \\ 0 \end{bmatrix} = \begin{bmatrix} U_c^T \\ U_u^T \end{bmatrix} s$$

where U_c is the left sub-matrix of U composed of the r column vectors corresponding to nonzero singular values of A, and U_u is the remaining sub-matrix of U. It follows that $\widetilde{s}_c = U_c^T s$ (and indeed $U_u^T s = 0$).

If we let $\widetilde{Q} = V^T Q V$, $\widetilde{M} = V^T M V$, $\widetilde{N} = V^T N$, $\widetilde{z}_0 = V^T z_0$, our problem becomes

$$\begin{aligned} &\min_u J = \mathbb{E}\left(\frac{1}{2}\int_0^\infty \left[\widetilde{z}(t)^T \widetilde{Q} \widetilde{z}(t) + u(t)^T R u(t)\right] dt | \widetilde{z}_0\right) \\ &\text{subject to } \begin{cases} \dot{\widetilde{z}}(t) = \widetilde{M}\widetilde{z}(t) + \widetilde{N}u(t), \quad \widetilde{z}(0) = \widetilde{z}_0 \\ \widetilde{s}_c(t) = \Sigma_r \widetilde{z}(t) \end{cases} \end{aligned} \tag{4.50}$$

Hence from previous work, the problem above can be reduced to:

$$\begin{aligned} &\min_u J = \frac{1}{2}\text{trace}(K\widetilde{P}^T) \\ &\text{subject to } \begin{cases} K\widetilde{M}_0 + \widetilde{M}_0^T K + \widetilde{Q} + \Sigma_r^T \widetilde{F}^T R \widetilde{F} \Sigma_r = 0 \\ \widetilde{M}_0 = \widetilde{M} - \widetilde{N}\widetilde{F}\Sigma_r \\ \widetilde{P} = \mathbb{E}(\widetilde{z}_0 \widetilde{z}_0^T) \\ \widetilde{M}_0 \text{ is stable} \end{cases} \end{aligned}$$

Therefore we have $u^*(t) = -\widetilde{F}\widetilde{s}_c(t)$ where

$$\widetilde{F} = R^{-1}\widetilde{N}^T K \lambda \Sigma_r^T (\Sigma_r \lambda \Sigma_r^T)^{-1}$$

with

$$\begin{aligned} &K\widetilde{M}_0 + \widetilde{M}_0^T K + \widetilde{Q} + \Sigma_r^T \widetilde{F}^T R \widetilde{F} \Sigma_r = 0 \\ &\lambda \widetilde{M}_0^T + \widetilde{M}_0 \lambda + \frac{1}{2}\widetilde{P} = 0 \\ &\widetilde{M}_0 = \widetilde{M} - \widetilde{N}\widetilde{F}\Sigma_r \\ &\widetilde{P} = \mathbb{E}(\widetilde{z}_0 \widetilde{z}_0^T) \end{aligned}$$

And finally, $u^*(t) = -\widetilde{F}U_c^T s(t)$, which means $F = \widetilde{F}U_c^T$.

Remark 4.22. Note that

- Unlike Problem (4.4), Problem (4.50), is more likely to be output controllable.
- Applying Sylvester's criterion to $M - NFA$ is equivalent to applying it to $\widetilde{M} - \widetilde{N}\widetilde{F}\Sigma_r$.
- Solving Problem (4.4) is handled by solving Problem (4.50) obtained from transforming Problem (4.4) via SVD, and then transforming the solution from the mode space back to the original space via the inverse transformation.
- If the A matrix in Problem (4.4) is rank deficient, then so is the square matrix $\Sigma_r \lambda \Sigma_r^T$. Hence, the $\widetilde{F}$ matrix as given just above this remark, cannot be determined since the matrix $\Sigma_r \lambda \Sigma_r^T$ is singular. Therefore the optimality condition algorithm fails for all rank deficient systems.

4.5 Illustrative Examples

We consider SALT primary mirror alignment problem as an optimal control problem, more precisely as a linear quadratic problem as formulated in Problem (4.4). The formulation of Problem (4.4) is recalled below:

$$\begin{aligned} &\min_u J = \mathbb{E}\left(\frac{1}{2}\int_0^\infty \left[z(t)^T Q z(t) + u(t)^T R u(t)\right] dt | z_0\right) \\ &\text{subject to } \begin{cases} \dot{z}(t) = Mz(t) + Nu(t), \quad z(0) = z_0 \\ s(t) = Az(t) \end{cases} \end{aligned} \tag{4.51}$$

where $s(t)$ is the vector of relative heights as read by sensors at time t, $z(t)$ is the vector of actuator positions at time t, A is the SALT 480×273 actuator-to-heights matrix, $Q = A^T A$, $R = \sigma I$ with I being the 273×273 identity matrix and σ a positive real number (we chose $\sigma = 0.2$ in our computation to emphasise more on minimising the overall relative heights and less on the effort or energy involved in the control process), M is the 273×273 zero matrix, N is the 273×273 identity matrix. Recall that in the SALT case, the problem is controllable but not output controllable and yet the control is performed based on the output since we don't have information about the state (actuator positions) unless we estimate from the output (relative

heights). The goal is to find an output feedback optimal control, that is, an optimal control in the form $u^* = -Fs(t)$ where F is the gain matrix to be determined. Two approaches involving problem transcription and numerical techniques are the optimality condition and the gradient flow approaches. In the optimality condition approach, our original problem reduces to solving (via algorithm 2, page 120) the system (4.41) - (4.42) - (4.29) recalled below:

$$\begin{cases} KM_0 + M_0^T K + Q + A^T F^T RFA = 0 \\ \frac{1}{2}P + \lambda M_0^T + M_0 \lambda = 0 \\ F = R^{-1} N^T K \lambda A^T (A \lambda A^T)^{-1} \end{cases} \tag{4.52}$$

In the gradient flow approach, our original problem reduces to solving (via algorithm 3, page 121) the system (4.41) - (4.42) - (4.43) recalled below:

$$\begin{cases} KM_0 + M_0^T K + Q + A^T F^T RFA = 0 \\ \frac{1}{2}P + \lambda M_0^T + M_0 \lambda = 0 \\ \dot{F} = 2N^T K \lambda A^T - 2RFA\lambda A^T; \qquad F(0) = F_0 \end{cases} \tag{4.53}$$

where $M - NF_0A$ is a known arbitrarily chosen stable matrix. In both cases, $P = \mathbb{E}\left(z_0 z_0^T\right)$, $M_0 = M - NFA$, and when the gain matrix F is numerically determined, the dynamics of the state (actuator displacements) and the output (relative heights) over time are determined as well. Then at any time t, the root-mean square (RMS) of actuator displacements $z(t)$ is determined (it can also apply to relative heights and more generally to any vector). The RMS of a vector $v = (v_1, v_2, \ldots, v_n)$ is determined as follows:

$$\text{RMS}\,(v) = \sqrt{\frac{v_1^2 + v_2^2 + \cdots + v_n^2}{n}} \tag{4.54}$$

As mentioned before (end of section 4.3.3), results from literature would apply if the actuator-to-heights matrix was a full rank $m \times n$ matrix with $m \leq n$, which in this case is not true. Therefore we convert the problem in the *mode space* using singular value decomposition, and convert the solution back to the initial space, as described in the previous section (Section 4.4).

4.5.1 Assessment of Optimality Condition and Gradient Flow on SALT

Lyapunov equations are inconsistent (for optimality condition and gradient flow approaches in equations (4.52) and (4.53)) when A is rank deficient. If

the system is unconstrained, the matrix $Q + A^T F^T RFA$ (in the Lyapunov equations) is square and rank deficient, hence singular. This is not the case when A is constrained or regularised. We need to constrain A as in SALT or regularise via SVD. SVD regularisation is done by changing the zero singular values using a simple rule such as $\sigma_{i+1} = \xi\sigma_i$ for $r \leq i \leq n-1$ where $r = 269$ is the rank of A, $n = 273$ is the number of actuators and number of columns of A, and ξ is a *regularisation factor*, with $0 < \xi \leq 1$. This method's efficiency depends on ξ and σ_r (the smallest nonzero singular value of A). The closer to 1 the regularisation factor ξ is, the stronger the regularisation. The regularised system is closer to the original system, compared to the piston constrained system. Recall that A is 480×273. Examples of regularisation factors are:

1. $\xi = 1$.
2. $\xi = \frac{\sigma_r}{\sigma_{r-1}}$ (last ratio) in this case 1. This choice is unreliable if the last ratio is too small.
3. $\xi = \frac{1}{r-1} \sum_{i=1}^{r-1} \frac{\sigma_{i+1}}{\sigma_i}$ (average ratio) in this case 0.9835.
4. $\xi = \sqrt{\frac{1}{r-1} \sum_{i=1}^{r-1} \left(\frac{\sigma_{i+1}}{\sigma_i}\right)^2}$ (RMS ratio) in this case 0.9846.

Note in Table 4.1 that in cases (1), (2), (3) and (4), despite the fact that the rank of the residual is 202, only four of the singular values are of order 10^{-2} and the remaining singular values are of order 10^{-14} or less.

Whether the system is regularised, constrained or unconstrained, the optimality condition algorithm systematically fails to converge. This is because in the iterative process of algorithm 2, the matrix $A\lambda A^T$ (even in the mode space) is ill-conditioned, due to the eventual rank deficiency of the λ matrix. Indeed, the λ matrix is not necessarily a full rank matrix at every step of algorithm 2.

After regularisation (and similarly after piston constraints), the gradient flow algorithm converges when $z_{\max} \leq 2.5 \times 10^{-4}$ metres (250 microns). This result is obtained from a simulation where the initial configuration of the primary mirror (z_0 in our problem formulation) is generated. Each actuator displacement is randomly chosen between $-z_{\max}$ and $z_{\max}$; 200 trials are conducted for each value of $z_{\max}$, and $z_{\max}$ varies between 100 and 1000 microns with a 10 microns step. However, it is important that

Table 4.1: The A matrix with constraints and regularisation

A **matrix**	**Size**	**Rank**	**Norm**	**Maximum**	**Minimum**
Unconstrained	480×273	269	4.4844	1.3979	-1.3979
SALT constr.	484×273	273	4.4844	1.3979	-1.3979
Diff. (SALT)	N/A	N/A	N/A	N/A	N/A
Regularised (1)	480×273	273	4.4844	1.3986	-1.3988
Diff. (1)	480×273	202	0.0347	0.0012	-0.0014
Regularised (2)	480×273	273	4.4844	1.3986	-1.3988
Diff. (2)	480×273	202	0.0347	0.0012	-0.0014
Regularised (3)	480×273	273	4.4844	1.3985	-1.3987
Diff. (3)	480×273	202	0.0341	0.0012	-0.0013
Regularised (4)	480×273	273	4.4844	1.3985	-1.3987
Diff. (4)	480×273	202	0.0342	0.0012	-0.0013

the F matrix be computed fast enough. Time to compute the F matrix is smaller than 1 second when $z_{\max} \leq 240$ microns. This time is always above 0.5 second, and the probability that it is below 0.7 second is greater than 95%. Note that when A is regularised, the gradient flow algorithm can be implemented without conversion to and from the mode space via SVD. Time to compute the F matrix directly is on average 0.1 second greater than when the computation of the F matrix is done via conversion to and from the mode space. This is due to the fact that in the mode space, the matrix and vector dimensions are substantially reduced and matrices are in simpler forms. After 200 trials, when $z_{\max} = 10^{-4}$ metres (100 microns), computing time for the F matrix via gradient flow is between 0.584541 and 0.974725 second. Sampled probabilities are given in Table 4.2, where $P(t \geq T)$ is the probability that the computing time t of the F matrix is greater than a given value T (where T is given in seconds):

A pseudo-code to explain how the gradient flow algorithm is implemented, is given below:

Gradient Flow Pseudo-code

(a) Read relative heights $s(0)$ from sensors at start

Table 4.2: Sampled probabilities of computing time of the F matrix (computing time in seconds) via gradient flow when $z_{\max} = 100$ microns

Probability	Value
$P(t \geq 0.5)$	200/200
$P(t \geq 0.6)$	102/200
$P(t \geq 0.7)$	11/200
$P(t \geq 0.8)$	5/200
$P(t \geq 0.9)$	2/200

(b) Estimate actuator positions z_0 using SVD (either directly or via exponential moving average over 30 seconds to compensate for noise)

(c) Compute the output feedback gain matrix F via algorithm 3 (given in page 121)

(d) At each step (one second or less – time step for numerical integration of an ODE – since this is a continuous time process), adjust the actuator positions using the equation $\dot{z}(t) = Mz(t)+Nu(t) = (M-NFA)z(t) \iff z(t) = \exp[(M-NFA)t]\, z_0$ – or any standard numerical technique for solving ODEs

4.5.2 Simulation Results on SALT

We start in a random configuration of the primary mirror. The initial configuration (z_0 in our problem formulation) is generated by randomly sampling each actuator displacement between $-z_{\max}$ and $z_{\max}$ via uniform distribution, for a given value of $z_{\max}$. One case scenario is when each actuator displacement is within the range of $z_{\max} = 10^{-4}$ metre (see Table 3.2), that is, when the control method of Chapter 3 fails.

We illustrate the gradient flow results in Figure 4.5 for the specific case $z_{\max} = 10^{-4}$ metre (100 microns). We see that the RMS actuator displacement (describing the misalignment of the primary mirror) decays exponentially. It takes about 4.7 seconds to have all the actuator displacements (and consequently the RMS actuator displacements) below one micron. This is, with this initial configuration, how long it should take for the SALT primary mirror to be safely controlled by fast alignment via QR after gradient

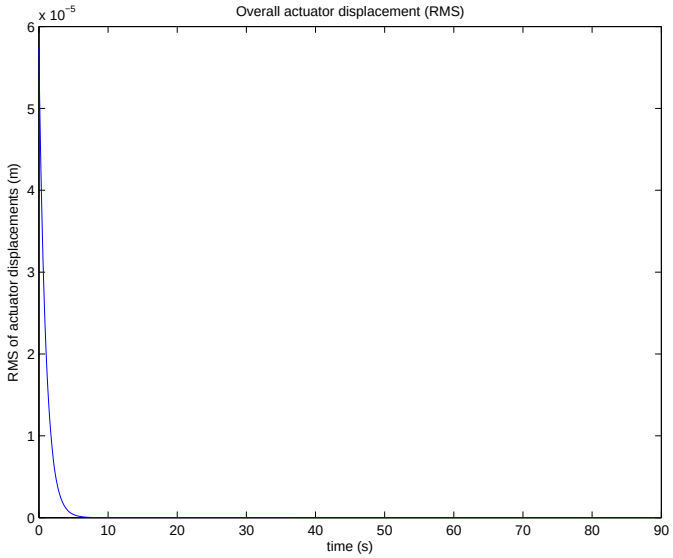

Figure 4.5: RMS actuator displacements over time: the gradient flow approach brings all the actuator displacements below 1 micron within 4.7 seconds, and brings all the actuator displacements below 5 microns within 3 seconds; then fast alignment via SVD (or even via QR) can take over after about 5 seconds

flow (4.7 seconds). On the other hand, the gradient flow approach brings all the actuator displacements below five microns within about 3 seconds. This is, again, with this initial configuration, how long it should take for the SALT primary mirror to be safely controlled by fast alignment via SVD after gradient flow (3 seconds). In Figure 4.6, we simulate for 200 initial mirror configurations. We get from simulations that the gradient flow approach takes between 4.6 and 4.7 seconds to get all the actuator displacements below one micron (acceptable safe control by fast alignment via QR), and about 3 seconds to get all the actuator displacements below five microns (acceptable safe control by fast alignment via SVD). We then have the important result that gradient flow algorithm is a very efficient alignment process. That is, after about five seconds, actuator displacements are in the range of controllability as in Chapter 3. Of course, we assume that the precision of actuator drive motors is acceptable.

The gradient flow mirror alignment system takes about five seconds to bring the primary mirror in a configuration where fast alignment (by SVD or even by QR) can safely take over. It is clear that the gradient flow approach

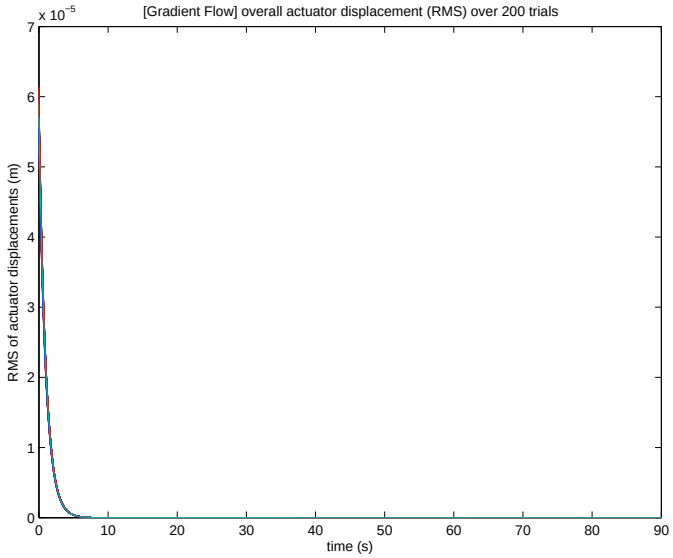

Figure 4.6: Overall actuator displacements over time using 200 trials (Gradient flow): RMS actuator displacements are illustrated. All actuator displacements are brought below 1 micron within 4.6 to 4.7 seconds (then fast alignment via QR can safely take over after 64.7 seconds), and below 5 microns within 3 seconds (then fast alignment via SVD can safely take over after 3 seconds)

is a reasonable option to consider. RMS actuator displacements of Figure 4.5 is approximately fitted by the function $y = 10^{-4}e^{-t}$ where t is given in seconds. Assuming this model holds for all initial amplitudes, we note that when these are order 10^{-6} metre, gradient flow is in successful long time control of the mirror segments (RMS actuator displacements order one micron). In this case, gradient flow control should apply on SALT without need of SVD control, if it is left active for all time. Recall that gradient flow will work with constraints on the physical mirror such as are needed in QR control (Section 4.3), or, with regularisation on the system to guarantee consistency of the equations to solve. Note that regularisation is a better approximation to the unconstrained system than piston constraints.

4.6 Conclusion

It has been established that for a multi-element telescope, especially for SALT, in fast alignment, computing time is not a constraint in the control

process since this computing time is most likely below 10^{-3} second, which is negligible compared to the actuator response time which is around 0.1 second. However, the fast alignment process is not reliable when the mirror is misaligned in such a way that individual actuator displacements might exceed one micron each (fast alignment with QR), or five microns each (fast alignment with SVD). This suggests exploration of alternate control techniques.

Theoretical approaches to the control system of a multi-element telescope have been explored, using background in control theory. The control problem has been formulated as a linear quadratic problem. Discrete time as well as continuous time formulations have been explored and in either case, optimal control can be applied. When the optimal control is a state feedback control, the most common tool to achieve it is dynamic programming. However, our control problem is an output feedback control and other techniques were found useful, involving problem transcription and numerical techniques. We have found that the system can best be controlled from an arbitrary initial configuration by the gradient flow method, at least when $z_{\max} \leq 2.4\times10^{-4}$ metre, $z_{\max}$ being the maximum displacement from zero for each actuator, in the positive or negative direction. With $z_{\max} \leq 2.4 \times 10^{-4}$ metre, the time it takes to compute the F matrix is smaller than one second, and between 0.5 and 0.7 second with a probability greater than 95%. Note that the bigger the value of $z_{\max}$, the longer it takes for the system to reach acceptable alignment, but acceptable alignment is achieved nonetheless. Actuator displacements and therefore RMS actuator displacements (describing mirror misalignment), can be fitted by decreasing exponential functions. When $z_{\max} = 10^{-4}$ metre, the gradient flow approach achieves acceptable alignment within five seconds for fast alignment via QR, and within about three seconds for fast alignment via SVD. The time constant of Figure 4.6 is of importance in deciding controllability of the mirror. For example, the mirror distorts because of environmental factors that might change over a time of minutes but if the numerics and actuators respond on times of hours, control might be impossible. In Figure 4.6, we note a time scale of about 6 seconds. This is a factor 5 faster than SVD control (with exponential moving average over 30 seconds) and applies to worse initial configurations (misalignments); thus gradient flow improves controllability over the existing SALT code. The significant variable of temperature

changes much more slowly than this. All this suggests that in the case of SALT, alignment by Gradient Flow is a reasonable option to consider when the misalignment of the primary mirror is out of the QR/SVD acceptable range.

Chapter 5

Conclusion

5.1 Assessment of Results Relevant to SALT

In 2007, image quality on SALT could be judged from direct photographic evidence of star images. Such images are not available to us but from discussions with SALT staff, it is clear that they were not satisfactory. At the time, a considerable number of measurements were taken of environmental factors. From these, it was found that humidity was often high while it was known that capacitive edge sensors were sensitive to high humidity. Examination was also made of the spherical aberration corrector. Corrective decisions were taken to replace edge sensors with inductive devices that are not humidity-sensitive and to return the spherical aberration corrector to its manufacturer for realignment. At the present time, tenders have been called for the edge sensors and the spherical aberration corrector is now indeed repaired and reinstalled on SALT. Yet, we note that in 2007 data, the control algorithm indicated that the mirror was under good control. It was known that CCAS contradicted the SAMS output.

1. In Chapter 2, we performed a detailed statistical analysis on the environmental data of 2007 in order to objectively decide the significant environmental factors affecting image quality. Stepwise regression clearly showed that in order of significance, time, truss temperature and humidity were the relevant variables. Humidity presumably was significant owing to capacitive edge sensors. Management of truss temperature is not performed (the dome is open to the sky) and consequent distortions of the truss must of course be managed by active control of mirror segments. Finally, the appearance of time as the dominant variable immediately suggested that computational errors could be accumulating over time. Together with the above-mentioned contradiction between CCAS and SAMS outputs, it was natural that we should re-examine the control algorithm.

2. In Chapter 3, an important initial result is deduced from Figure 3.14 where we run the 2007 SALT algorithm (SAMS – green line) and an independently written algorithm (CAM – red line), both using normal equations, where we see that SAMS was unresponsive to change in environment while our algorithm was responsive and, trended with CCAS (blue line). It follows that there was indeed an error in the coding of the original SALT algorithm.

   We then considered three control methods. Our findings are as follows: Normal equations and QR methods both require additional constraints to the primary mirror configuration in order to give the actuator-to-heights matrix full rank for the methods to be applicable. Besides the inconvenience of setting constraints to the mirror, the normal equations and QR methods were found to be less reliable than the SVD method. We note from Chapter 1 that the Hobby-Eberly Telescope uses the SVD method and our conclusion is thus supported.

   We implemented QR/SVD control on SALT in July 2010. Consider now the behavior of the figure of merit with time, recalling that this should be within 60 microns. Figure 3.13 most clearly shows that the original SAMS software of March 2007 failed to control because the figure of merit rises throughout the night and goes above 60 microns. In Figure 3.16 where QR/SVD is implemented, we find steady control for about 1.5 hour before control fails. From this it was clear that problems remain with the control algorithm. In Figure 3.16, the data was filtered using simple moving average over a four minutes period. This long averaging period is by itself a source of concern because alignment errors grow specifically during this period. Any filtering method that can reduce this period is of interest. We showed that exponential moving average allows us to reduce the filtering period to 30 seconds and furthermore gives excellent control with SVD method (Figure 3.23) because stable figure of merit at 20 microns is discovered. Finally, we see from Figure 3.24 that RMS tip/tilts are well within acceptable range (0.1 arcsecond) over the viewing period.

   In assessing the SVD method for SALT (Section 3.2.3), we found that individual actuator displacements had to be aligned in the CCAS process to within an accuracy of one micron in order for the algorithms to be reliable. This result implies that the actuator drive motors are

required to have an accuracy of better than one micron. Reliability of normal equations and QR methods is inferior to that of SVD and should be avoided.

3. In Chapter 4, we are concerned with domain of controllability. The degree of misalignment that can lead to acceptable control by QR/SVD is measured by the range of an actuator displacement (order one micron - Section 3.2.3). It is clear that during the night, if the alignment is severely perturbed above this level, QR/SVD may fail to restore alignment (Table 3.2). If we are confident that this will not happen, QR/SVD will always serve SALT.

   If we are not confident that the mirror is misaligned in such a way that actuator displacements are always within one micron, a more powerful algorithm is demanded. If during the night a perturbation of mirror alignment leading to RMS actuator displacements of a few microns occurs, the gradient flow algorithm will indeed restore controllability in reasonable time. It has been established that if each actuator displacement is within the range of 10^{-4} metre, which is equal to 100 microns (in the positive or negative direction), the gradient flow method can bring the whole system under control (bring all actuator displacements below one micron) within five seconds, so that fast alignment (by SVD or even by QR) can safely take over. This is very important for basic operation procedure on SALT.

5.2 Future Work

Concerning future SALT operations, our testing of July 2010 was limited to seven mirror segments with capacitive edge sensors, over less than a two hours period (we took measurements only when humidity was known to be sufficiently low that we could trust the edge sensors). If SALT management decides to continue with standard SVD numerics (as on HET), it is essential that SALT retests the QR/SVD control algorithm with 91 mirror segments and inductive edge sensors over a full night of observation before the mirror and the new software can finally, be safely commissioned.

It may yet be the case that the one micron requirement on actuator displacement precision cannot be met by existing actuator drive motors. Actuator displacements are of the same order of magnitude as relative heights as

measured by edge sensors. It then follows that edge sensors as well, should be sensitive to displacements of order one micron. If either precision cannot be achieved, SALT may yet have to turn to the gradient flow method. Indeed, we have found that gradient flow gives improved feedback time (less than order 5 seconds, compared to order 30 seconds for SVD with exponential moving average) and is robust against measurement errors. We recommend that gradient flow be multiplexed into the SALT control software, along with the (corrected) normal equations and SVD (with exponential moving average) code, and tested on SALT.

Concerning the future of multi-element telescopes, it is of obvious interest to investigate the controllability requirements. We note from Chapter 1 the actuator-to-heights matrix has dimensions (number of sensors by number of actuators) of 480×273 for SALT. For the proposed large telescopes, these dimensions are 2772×1476 (TMT) and approximately 6000×2952 (E-ELT). We ran simulations on larger versions of the SALT primary mirror where the size of the actuator-to-heights matrix is the closest to those of TMT and E-ELT respectively (13 rings for TMT and 18 rings for E-ELT). Compared to SALT, the time it takes to compute the pseudo inverse of the actuator-to-heights matrix increases by a factor of approximately 120 for TMT, and approximately 660 for E-ELT. Indeed, the respective computation times are approximately 0.0572 second for SALT, 6.7245 seconds for TMT and 37.5371 seconds for E-ELT. However, this pseudo inverse is assumed to be already provided, and what is left to assess is the matrix-vector operations. We expect matrix operations on these large dimension matrices to be satisfactory under SVD, given sufficient precision in actuator displacements. From our simulations, compared to SALT, the algorithm execution time increases by a factor of approximately 40 for TMT and approximately 140 for E-ELT. Indeed, considering the computation of actuator positions from relative heights, compared to SALT where the average computation time is around 0.0673 millisecond, we obtain an average computation time of 2.6755 milliseconds for TMT and 9.1657 milliseconds for E-ELT. This remains small in comparison to actuator response time. However, we have not investigated precisions required of edge sensors and actuator drive motors for these mirrors. On TMT, the proposed radius of curvature is 90 metres, and assuming the acceptable tip/tilts of SALT mirror, we estimate by simple geometry, that the acceptable tip/tilts on TMT must improve by one order

of magnitude. In particular, edge sensors and drive motors should work with precision better than 0.1 micron. If TMT and E-ELT fail to achieve this, the control algorithm based on SVD will not work. In this case, gradient flow method will be required for satisfactory image quality.

Bibliography

[1] A. F. Beardon, *Algebra and geometry*, Cambridge University Press, 2005.

[2] M. Berthold and D. J. Hans (eds.), *Intelligent data analysis: An introduction*, second ed., Springer-Verlag, Berlin, Heidelberg, 2007.

[3] D. P. Bertsekas, *Dynamic programming and optimal control*, third ed., vol. 1, Athena Scientific, 2005.

[4] P. M. T. Broersen, *Automatic autocorrelation and spectral analysis*, Springer-Verlag, London, 2006.

[5] M. W. Cantoni, K. L. Teo, and V. Sreeram, *A gradient flow approach to computing a nonlinear quadratic optimal feedback gain matrix for discrete time systems*, Conference on Decision and Control, vol. 2, December 1994, pp. 1400–1405.

[6] G. Chanan, D. G. MacMartin, J. Nelson, and T. Mast, *Control and alignment of segmented-mirror telescopes: Matrices, modes and error propagation*, Applied Optics **43** (2004), no. 6, 1223–1232.

[7] C. Chatfield, *The analysis of time series: An introduction*, fifth ed., Chapman & Hall/CRC, 1995.

[8] ———, *The analysis of times series: An introduction*, sixth ed., Chapman & Hall/CRC, 2004.

[9] S. Chatterjee and A. S. Hadi, *Regression analysis by example*, fourth ed., Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., 2006.

[10] E. K. P. Chong and S. H. Zak, *An introduction to optimization*, second ed., Wiley-Interscience, 2001.

[11] T. F. Cox, *An introduction to multivariate data analysis*, Oxford University Press, 2005.

[12] J. W. Demmel, *Applied numerical linear algebra*, Society for Industrial and Applied Mathematics, 1997.

[13] M. Dimmler, T. Erm, B. Bauvir, B. Sedghi, H. Bonnet, M. Müler, and A Wallander, *E-ELT primary mirror control system*, vol. 7012, SPIE, no. 1O, 2008.

[14] N. R. Draper and H. Smith, *Applied regression analysis*, third ed., Wiley-Interscience, 1998.

[15] B. S. Everitt and G. Dunn, *Applied multivariate data analysis*, first ed., 1991.

[16] R. Fenn, *Geometry*, Springer, London, 2001.

[17] E. W. Frees, *Regression modeling with actuarial and financial applications*, Cambridge University Press, 2010.

[18] C. J. Goh and K. L. Teo, *Alternative algorithms for solving nonlinear function and functional inequalities*, Applied Mathematics and Computation **41** (1991), no. 2, 159–177.

[19] R. A. Horn and C. R. Johnson, *Matrix analysis*, Cambridge University Press, 1990.

[20] A. J. Izenman, *Modern multivariate statistical techniques: Regression, classification and manifold learning*, Springer Texts in Statistics, Springer Science and Business Media LLC, 2008.

[21] S. M. Kay, *Fundamentals of statistical signal processing*, Prentice-Hall signal processing series, Englewood Cliffs, N. J.: Prentice-Hall PTR, 1998.

[22] G. Knowles, *An introduction to applied optimal control*, Academic Press, Inc., New York, 1981.

[23] V. Krishnan, *Probability and random processes*, John Wiley & Sons, Inc, 2006.

[24] W. S. Levine and M. Athans, *On the determination of the optimal constant output feedback gains for linear multivariable systems*, IEEE Transactions on Automatic Control **AC-15** (1970), no. 1.

[25] B. Luong, *SALT–SAMS data analysis report*, Tech. report, February 2006.

[26] M. Mariton and P. Bertrand, *Output feedback for a class of linear systems with stochastic jump parameters*, IEEE Transactions on Automatic Control **AC-30** (1985), no. 9.

[27] A. D. R. McQuarrie and C. L. Tsai, *Regression and time series model selection*, World Scientific Publishing, 1998.

[28] C. D. Meyer, *Matrix analysis and applied linear algebra*, Society for Industrial and Applied Mathematics, 2000.

[29] J. Nelson and G. H. Sanders, *The status of the thirty meter telescope project*, vol. 7012, SPIE, no. 1A, 2008.

[30] P. Nerin, *SALT–SAMS analysis report, geometrical modelisation PDR*, Tech. report, September 2002.

[31] ———, *SALT–SAMS analysis report, geometrical modelisation CDR*, Tech. report, January 2003.

[32] K. Ogata, *Modern control engineering*, first ed., Prentice-Hall, 1970.

[33] ———, *Modern control engineering*, fourth ed., T. Robbins and Prentice-Hall, Inc., 2002.

[34] R. L. Ott, *An introduction to statistical methods and data analysis*, fifth ed., Wadsworth Group, 2001.

[35] F. L. Ramsey and D. W. Schafer, *The statistical sleuth: a course in methods of data analysis*, Belmont, Calif.: Duxbury Press, 1997.

[36] C. R. Rao and H. Toutenburg, *Linear models: Least squares and alternatives*, Springer-Verlag, New York, 1995.

[37] J. O. Rawlings, S. G. Pantula, and D. A. Dickey, *Applied regression analysis: A research tool*, Springer-Verlag, New York, Inc., 1998.

[38] N. Sessions, *SALT primary mirror array configuration*, Tech. report, November 2002.

[39] R. H. Shumway and D. S. Stoffer, *Time series analysis and its applications*, Springer Science and Business Media, 2006.

[40] E. D. Sontag, *Mathematical control theory: Deterministic finite dimensional systems*, Springer-Verlag, New York, 1998.

[41] J. Swiegers, *SALT primary mirror alignment system specification*, Tech. report.

[42] ———, *SALT primary mirror segment alignment requirements*, Tech. report, April 2001.

[43] K. L. Teo, C. J. Goh, and K. H. Wong, *A unified computational approach to optimal control problems*, John Wiley & Sons, Inc., New York, 1991.

[44] K. L. Teo, K. H. Wong, and W. Y. Yan, *Gradient-flow approach for computing a nonlinear-quadratic optimal-output feedback gain matrix*, Journal of Optimization Theory and Applications **85** (1995), no. 1, 75–96.

[45] S. Weerahandi, *Exact statistical methods for data analysis*, Springer-Verlag, New York, 1995.

[46] K. H. Wong, N. Lock, and K. Kaji, *A computational method for a class of jump linear quadratic systems*, Austral. Math. Soc. **B 36** (1995), 414–423.

[47] K. H. Wong, C. Myburgh, and L. Omari, *A gradient flow approach for computing jump linear quadratic optimal feedback gains*, Discrete and Continuous Dynamical Systems **6** (2000), no. 4, 803–808.

[48] W. Y. Yan, K. L. Teo, and J. B. Moore, *A gradient flow approach to computing LQ optimal output feedback gains*, Optimal Control Applications & Methods **15** (1994), 67–75.

[49] X. Yan and X. G. Su, *Linear regression analysis: Theory and computing*, World Scientific Publishing, 2009.