A statistical model to describe the spread of wafer lenses

Figueroa Burbano, Y.A.

Award date:
2011

Link to publication
A Statistical Model to Describe the Spread of Wafer Lenses.

Author:
Yeisson Augusto Figueroa Burbano

Supervisor:
Alessandro Di Bucchianico (TU/e)
Edwin Wolterink (Anteryon)
Bart van Loenen (Anteryon)
“Action is the foundational key to all success”.

Pablo Picasso
Abstract

A lens is an optical device with optical symmetry that can transmit, filter, block, diffract or refract light. Lens production based on the wafer technique can be done in several stages; most relevant ones are mastering, replication, bonding and dicing. A lens module consists of multiple combinations of such components stacked one above the other according as the optical requirements.

The description for a lens profile based on curve fitting procedure yield, among others, desirable properties which can be used in later process to test the whole module lens performance. In order to predict such performance of a lens module, a good understanding of different approaches to describe the inherent variability present in a lens profile description is needed. An incorrect interpretation of such description can generate undesirable effects when a lens module is placed in an optical device.

This work presents several approaches on the curve fitting procedure to describe a lens profile and a well-known statistical approach called Principal Component Analysis is used to describe the sources of most variability in the lens profile.

Results covers various curve fitting procedures on the lens profile, statistical inference on such fittings, variability description based on the parameter estimation under the linear regression framework and dimensionality reduction under the PCA technique.
It is in the sense of the words of my favorite writer, Oscar Wilde, “Success is a science; if you have the conditions, you get the result”, that I would like to thank all people involved in the satisfactory culmination of this master thesis.

I would like to express my immense gratitude to my supervisor at the University, Dr. Alessandro Di Bucchianico. It was him who suggested this project for my thesis. I am very thankful to him for his constant encouragement and support throughout the mathematical understanding on the course of the project.

I am also very thankful to Bart van Loenen, my supervisor at Anteryon, who was constantly monitoring my progress and always willing to provide me with information whenever I needed it.

I must also acknowledge Edwin Wolterink, from Anteryon. His knowledge on the project, discussions and feedback on the course of the project proved to be fruitful in developing the object of this thesis.

Let me also thank Jules Kierkels and Michel Werts, my colleagues at Anteryon, for their friendship and encouragement during my stay in Anteryon.

Last, but not least, I thank Mrinal Shenoy, my colleague from the University, who made this project a totally enjoyable experience.
# Contents

1 Introduction  
1.1 Background ................................................. 1  
1.2 Manufacturing Process ...................................... 2  
1.3 Literature Review .......................................... 3  
1.4 Aim and structure of Thesis ................................. 4  

2 Mathematical Principles  
2.1 Change of basis transformations .......................... 7  
2.2 The Linear Model and Linear Regression ................. 11  
2.3 Summary of the Chapter .................................... 25  

3 Data Description ............................................. 26  

4 Estimation in the Cross-Section of a Lens  
4.1 Estimation Procedure ...................................... 31  
4.2 Basis Transformation versus Direct Estimation ......... 36  
4.3 Diagnostics .................................................. 41  
4.4 Shifting to the Origin ..................................... 51  
4.5 Flipping ...................................................... 55  
4.6 Summary of the Chapter .................................... 62  

5 Understanding Variation  
5.1 Variance-Covariance from Linear Regression .......... 64  
5.2 Variance-Covariance from a set of Estimated Coefficients .................................. 67  
5.3 Principal Component Analysis ............................... 69  

i A Statistical Model to Describe the Spread of Wafer Lenses.
6 Conclusions and Recommendations 78

6.1 Summary of the Work ........................................ 78
6.2 Recommendations ........................................... 80

7 Future Research 82

7.1 Estimation on the Surface of a Lens ....................... 82
7.2 Models for Components of Variance ....................... 83
7.3 Optimal Designs ........................................... 86
7.4 Summary .................................................. 89

A Background Literature 90

A.1 Zernike Polynomials ....................................... 90
A.2 Notation and Zemax ....................................... 92

B Tables and Lists 95

Bibliography 103

ii A Statistical Model to Describe the Spread of Wafer Lenses.


1 Introduction

1.1 Background

Anteryon produces optical devices at the wafer scale using the WaferOptics® technology\textsuperscript{1}. Based on this technology, Anteryon combines high volume production, lens shape flexibility and low-costs. WaferOptics® is an innovative technique to produce micro-lenses in mass production under restrictive conditions. Manufacturing lenses at the wafer scale not only combines the above mentioned characteristics but also results in products with optical functionalities equally comparable with lenses produced under traditional manufacturing processes.

The main components of a lens are commonly glass substrate and polymers, which both lead to desired optical properties and shape. These two components are, at the final stage, mounted on a glass substrate which consists in a multiple combinations of such parts stacked one above the other; such combination is usually called lens module. Figure 1.1 presents a cross-section of a lens module. When the final product is manufactured, because of physical and chemical reactions among others during the manufacturing, the final lens shape deviates from the original design. A lens shape

\begin{figure}[h]
    \centering
    \includegraphics[width=0.5\textwidth]{fig1}
    \caption{Lens module, cross-section}
\end{figure}

\textsuperscript{1}WaferOptics® is a registered trademark of Anteryon.

\textsuperscript{1}A Statistical Model to Describe the Spread of Wafer Lenses.
design is basically made of a radial polynomial description with symmetry assumptions so the total shape of the lens is created. This shaped design made without considering variation and effects in the final products can lead to severe deviations from the desired optical purposes in terms of beam dispersion, blurring, and in general terms, performance of the lens. Finally, an entire production run may become useless.

It is then fundamental to quantify the variability of the final lens shape so appropriate compensations in the design phase can be introduced and extensive post production analysis can be done in the evaluation of the lens performance.

### 1.2 Manufacturing Process

The wafer technique based on the replication method consists in several stages (see Jeong et al. (2006)). The most relevant ones, which in fact are used by Anteryon, are mastering, replication, bonding and dicing. Figure 1.2 depicts the wafer process while Figure 1.3 presents the detailed replication process. Based on this technique, Anteryon is able to produce integrated optical devices using 8 inch wafer level packaging, so-called lens modules. Lens modules are made on glass substrates carrying large arrays of lenses which are arranged by Anteryon's replication technology. Basically there is no restriction on the shape. The final lens can be (a)spherical, (a)cylindrical or even free shape. There are some intermediate steps which allow a lens module to be of less than 2 mm in thickness. Although production of lenses with the above technique is relatively new and still in development, production is done in high volume. Finally, Anteryon

2 A Statistical Model to Describe the Spread of Wafer Lenses.
produces optical lenses at the wafer scale which has applications in several areas, the predominant one being in mobile phone cameras.

1.3 Literature Review

Lens production operated under the Wafer technique is an industrial batch process where each lens is manufactured as a part of wafer which represents a thin disk. Each wafer contains some hundreds of lenses. Only at the very final stage, a lens is obtained by dicing the respective wafer. Because of that, we can say that lenses are produced by lots, i.e. by wafers. In some similar production processes, the wafers as well can be handled by lots. In such production processes detecting the occurrence of assignable causes is fundamental so preventive actions may be taken. The variation within a group of measurements is traceable to various causes, and therefore, needs to be decomposed into relevant components of variation for an effective equipment monitoring and diagnosis. Therefore this project is aim to investigate a process in which the variance on specific parameters is attributed to different sources of variability.

Models to investigate different source of variability in an industrial process are quite extensive in literature. Due to the nature of the manufacturing process, the variation of the final shape of a lens may come from deviations produced within a wafer and/or among wafers. Czitrom and Spagon (1997) provides, besides a broad description of the semiconductor industry (which shares general similarities with the optical industry), a compilation of several case studies in the semiconductor industry which includes successful implementation of statistical techniques such as gauge studies, passive data collection and statistical process control (PCA) among others. Khuri (2000) presents a review in literature and applications of designs for estimating the variance components. He basically describes history, methods and application in genetics, statistical process control and quality improvement fields, and he also refers to Khuri and Sahai (1985) and Searle (1995) for earlier reviews on the topic. He describes the most rele-
vant findings for the one-way model as well as for the two-ways model for components of variance. Additionally nested models for when the process itself is divided in stages are described. Kim and Yum (1999) classifies models for components of variance in integrated circuits production by wafer procedures in: a) Nested Random-Effects models and b) Mixed-Effects models. Yashchin (1994), Yashchin (1995) and Woodall and Thomas (1995) among others present models and applications in the semiconductor industry of the nested random-effects models in which a particular quantity is considered to be object of nested variability between different levels of the production process, such as lot-to-lot, wafer-to-wafer-within-lot and within-wafer. Roes and Does (1995) presents a procedure to perform a mixed-effect model for a situation in which the model of nested-random effects does not adequately describe the process variation. The mixed-effect model presented includes both fixed and random effects in which the fixed amount is attributed to the levels per wafer and the random amount is due to between batch variation. Kim and Yum (1999) describes a similar application of the mixed-effects model in which the variability is not only attributed to random differences between batches but also to the fixed differences among the positions within a batch. In this application, special attention is devoted to the work performed by Roes and Does (1995). More recently, Ittzés (2001) used a nested model to monitor the variance components from the deviation of the compound-to-compound, the sample-to-sample-within-compound and the within-sample for controlling the dry matter content in butter cream in the dairy industry.

Applicability of the components of variance widely applied in the semiconductor industry for statistical process control should not be overemphasized when treating variability in optical systems. Yashchin (1995), Yashchin (1994) and Woodall and Thomas (1995) as well as Roes and Does (1995) developed models for industrial processes when a single quantity is required to be controlled at the time, namely oxide thickness or wafer thickness. Instead, the variability of the lens profile should be explored, initially, based on estimations considering Zernike polynomials (see Zernike (1934)) in which not only a single characteristic of the lens considered but several of them might be subject of study simultaneously. The Zernike polynomials\(^2\) have desired properties in statistical terms such as normalization and orthogonality which allow a proper interpretation in optical systems. A common used notation for the Zernike polynomials is given in Noll (1976) where the two-index notation is conveniently replaced by a single multi-index.

1.4 Aim and structure of Thesis

When producing a lens, the desired output shape varies from the input design shape. Understanding how the variation behaves is crucial in the early design stage of the production process. Since these variation may come from difference sources, it is relevant

\(^2\) see Appendix A for more details on the Zernike polynomials

4 A Statistical Model to Describe the Spread of Wafer Lenses.
to understand, in statistical terms, the nature of these deviations from the prescribed desired lens shape. The sources of variability should be treated not only based on process adjustments but also in identification of the different variance components, i.e. finding such sources that lead to a unfavorable lens performance.

The goal of the project is to develop a statistical model which describes the spread of the final shape of lenses made by the WaferOptics® technology. It is then required to study the spread characterized by Wafer-to-Wafer and Within-Wafer variability. Three main points are expected to obtain from the statistical model:

- Choose appropriate quantities to describe the variation in the lens shape.
- Show how to compute these quantities.
- Recommend required measurements to be made both in number and locations

At a first stage, a two-dimensional model is considered. That is, only data from the cross-section of a lens is used for the analysis. This step is followed by the development of a model for components of variance in which the input is the estimated coefficients. Secondly, based on the conclusions drawn from the two-dimensional analysis, an extension into the three-dimensional case will be performed. Data is used firstly to estimate coefficients and then used as input in the model for components of variance. It is expected that the estimation in a three-dimensional basis will entirely capture the variability of a lens shape.

The model developed is expected to help the lens/mould designer in determining the final mould shape in order to achieve the target lens, and in providing support in determining possible improvements in the production process. Such improvements can be directly related to reduce tolerances, improving lens performance and effectively achieving customer requirements.

As the project got underway we found that some of the assumptions adopted were not fulfilled under the current model formulation. It was decided to first investigate different approaches for lens profile description and its variability by taking into account such findings. The ultimate goals could not be achieved under the proposed time and the research was limited only to the cross-section of a lens rather than the surface description and its variability. Nevertheless, some remarks and recommendations are done in order to provide an sketch in how to achieve the initial proposed goals.

This report is structured as follows. Chapter 1 presents motivation and literature review as well as the problem description. The general mathematical principles to be used in later Chapters are presented in Chapter 2. A description of the collected data is presented in Chapter 3. Different curve fitting procedures are described in Chapter 4 along with linear transformation on the linear regression findings. Chapter 5 presents first a
description of the variability based on the linear regression setting, secondly variability
description based on given parameter estimations and finally the implementation of a
statistical technique is described for dimensionality and redundancy reduction. Finally,
conclusions and further research are discussed in Chapter 6. Since the initially pro-
posed goal were not entirely achieved, we provide in Chapter 7 basis for future research
regarding those goals.
2 Mathematical Principles

A lens is often modeled under either a polynomial function or a linear combination of the well-known Zernike polynomials (See Zernike (1934)). The mathematical principles in this chapter gives a general framework on the linear basis transformation needed in representing a lens and the general parameters of curve fitting procedures. For understanding how the modeling of a lens can be done we first describe the theory of finite-dimensional linear spaces on which the polynomials or the combination of Zernike polynomials take their values in Section 2.1. This description is given in an abstract setting and then it is implemented in the specific case of Zernike polynomials. Section 2.2 presents the Linear Regression theory. Estimators and its properties as well as the usage of orthogonal functions in a linear regression model are described.

2.1 Change of basis transformations

In this section, we describe the matrix transformation to transform the set of coefficients of a polynomial to another uniquely related set of coefficients. Application lies in the interest to find a set of coefficients for a particular set of polynomials given the corresponding set of coefficients in the standard monomials.

2.1.1 Change of basis

Let $V$ be a finite-dimensional linear space of dimension $k$ and let $B = \{e_1, e_2, \ldots, e_k\}$ be a basis of $V$. Then, every element of $V$, say $y$, can be expressed as linear combination of the $e_i$ as

$$y = \sum_{i=1}^{k} \beta_i e_i$$

where such representation is unique since the basis $B$ is linearly independent.

**Definition 2.1.** The $\beta_i$'s are scalars which will be called the coordinates of $y$ with respect to the basis $B$.

7 A Statistical Model to Describe the Spread of Wafer Lenses.
We will use the notation

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$$

and $\beta$ will be called the coordinate vector of $y$ with respect to the basis $B$. This means that $\beta$ is a column vector. For simplicity, $\beta$ will be written in row vector form as $\beta = [\beta_1, \beta_2, \ldots, \beta_k]^T$ with the superscript “$T$” denoting transpose.

Now, let us assume we also have another basis, say $B_1 = \{f_1, f_2, \ldots, f_k\}$, for the same space $V$. Then, in the same way as before, every element $y \in V$ can be expressed in either bases $B$ and $B_1$ as follows

$$y = \sum_{j=1}^{k} \beta_j e_j = \sum_{j=1}^{k} \alpha_j f_j. \quad (2.1)$$

Let us set $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_k]^T$ to be the coordinates of $y$ with respect to the basis $B_1$. This means that both $\beta$ and $\alpha$, are column vectors of scalars. Since, for each $j = 1, \ldots, k$, the basis vector $f_j$ has a unique representation as linear combination of the $e_i$, we set

$$f_j = \sum_{i=1}^{k} t_{ij} e_i. \quad (2.2)$$

That is, we have the following system of linear equations

\[
\begin{align*}
    f_1 &= t_{11} e_1 + t_{21} e_2 + \ldots + t_{k1} e_k \\
    f_2 &= t_{12} e_1 + t_{22} e_2 + \ldots + t_{k2} e_k \\
    \vdots &= \vdots + \vdots + \ldots + \vdots \\
    f_k &= t_{1k} e_1 + t_{2k} e_2 + \ldots + t_{kk} e_k.
\end{align*}
\]

The scalars $t_{ij}$ can be arranged in a $k \times k$ matrix which will be denoted as $T$. This matrix is referred as the basis transformation matrix from the basis $B_1$ to the basis $B$. Substituting back into (2.1) we get

$$y = \sum_{i=1}^{k} \beta_i e_i = \sum_{j=1}^{k} \alpha_j f_j = \sum_{j=1}^{k} \alpha_j \sum_{i=1}^{k} t_{ij} e_i$$

$$= \sum_{j=1}^{k} \sum_{i=1}^{k} \alpha_j t_{ij} e_i = \sum_{i=1}^{k} \sum_{j=1}^{k} t_{ij} \alpha_j e_i$$

$$= \sum_{i=1}^{k} \left( \sum_{j=1}^{k} t_{ij} \alpha_j \right) e_i.$$

8 A Statistical Model to Describe the Spread of Wafer Lenses.
By equating the coefficients of the $e_i$ we obtain
\[ \beta_i = \sum_{j=1}^{k} t_{ij} \alpha_j. \]  
(2.3)

Then, we have
\[ \begin{align*}
\beta_1 &= t_{11} \alpha_1 + t_{12} \alpha_1 + \ldots + t_{1k} \alpha_k \\
\beta_2 &= t_{21} \alpha_1 + t_{22} \alpha_2 + \ldots + t_{2k} \alpha_k \\
\vdots &= \vdots + \vdots + \ldots + \vdots \\
\beta_k &= t_{k1} \alpha_1 + t_{k2} \alpha_2 + \ldots + t_{kk} \alpha_k.
\end{align*} \]

Using the notation for the $\beta_i$'s, $\alpha_j$'s and $t_{ij}$'s we write
\[ \beta = T^t \alpha. \]  
(2.4)

So, the transpose of $T$ transforms the coordinate vector with respect to the basis $B_1$ to the coordinate vector with respect to the basis $B$. In the same way as before, for each $j = 1, \ldots, k$, the basis vector $e_j$ has also a unique representation as linear combination of the $f_i$, then, as in (2.2), we have for some $\tilde{t}_{ij}$,
\[ e_j = \sum_{i=1}^{k} \tilde{t}_{ij} f_i \]

and following the same procedure we find that the transpose of $\tilde{T}$, the $k \times k$ matrix with the scalars $\tilde{t}_{ij}$ as entries, transforms the coordinate vector with respect to the basis $B$ to the coordinate vector with respect to the basis $B_1$. That is
\[ \alpha = \tilde{T}^t \beta. \]  
(2.5)

It can be shown (see for instance Fuhrmann (1996), Corollary 2.9.1, pag. 48) that if $T$ is the basis transformation matrix from basis $B_1$ to the basis $B$, then $T^{-1}$ is the basis transformation matrix from basis $B$ to the basis $B_1$. That is, in the above description we set $\tilde{T} = T^{-1}$ and (2.5) becomes $\alpha = (T^t)^{-1} \beta$.

2.1.2 The abstract polynomial vector space $\mathcal{P}$

Our main concern, regarding the basis transformation, is to find the coordinates of $y$ with respect to a basis $B_2$ given the coordinates with respect to another basis $B_1$. We first start by giving some definitions which will be used later.

**Definition 2.2.** A polynomial is an expression of the form $p(x) = \sum_{i=0}^{n} c_i x^i$, for some $n$ where $c_i \in \mathbb{R}$ for all $i$ are called the polynomial coefficients.

9 A Statistical Model to Describe the Spread of Wafer Lenses.
Fuhrmann (1996) provides a set of additional definitions regarding Definition 2.2. Some of those are summarized as follows:

- Denote by \( \mathcal{P}(x) \) the vector space of all polynomials in \( x \) with coefficients \( b_i \in \mathbb{R} \), then we write \( \mathcal{P}(x) = \{ \sum_{i=0}^{n} b_i x^i \mid b_i \in \mathbb{R}, n = 0, 1, 2, \ldots \} \).
- Two polynomials \( p, q \in \mathcal{P}(x) \) are called equal if all their coefficients coincide.
- If \( c_n \neq 0 \) in Definition 2.2, then we say that \( n \) is the degree of the polynomial \( p \). It will be denoted by \( \deg p = n \).
- The degree of the zero polynomial, that is \( p(x) = 0 \), is defined to be \(-\infty\).
- The degree of the constant polynomial, that is \( p(x) = c \) with \( c \in \mathbb{R} \), is defined to be 0.

**Definition 2.3.** \( \mathcal{P}_n(x) \) denotes the vector space of all polynomials in \( x \) of degree at most \( n \). That is \( \mathcal{P}_n(x) = \{ p \in \mathcal{P}(x) \mid \deg p \leq n \} \).

Clearly, \( \mathcal{P}_n(x) \) is an \( n+1 \) dimensional subspace of \( \mathcal{P}(x) \). In fact, the set \( \{1, x, x^2, \ldots, x^n\} \) is a basis for \( \mathcal{P}_n(x) \). Then, every \( p \in \mathcal{P}_n(x) \) can be written down as linear combination of such basis and it has the explicit form given in Definition 2.2. We set \( k = n + 1 \) to be the dimension of \( \mathcal{P}_n(x) \) and \( \mathcal{B}_1 = \{1, x, x^2, \ldots, x^n\} \) to be one of its bases. We refer \( \mathcal{B}_1 \) as the standard basis of \( \mathcal{P}_n(x) \).

**Lemma 2.1.** Let \( p_i \) be a polynomial of exactly degree \( i \). Then any set \( \{p_i\}_{i=0}^{n} \) is a basis for \( \mathcal{P}_n(x) \).

**Proof.** Assume that the set \( \{p_i\}_{i=0}^{n} \) is not a basis. Then, there is by contradiction a polynomial \( p \in \mathcal{P}_n(x) \) that can be written as two different linear combinations of the set. That is \( p = \sum_{i=0}^{n} c_i p_i = \sum_{i=0}^{n} d_i p_i \). This implies that there exists an \( i \) such that \( c_i \neq d_i \) and thus \( c_i - d_i \neq 0 \) for some \( i \). Therefore we find \( 0 = \sum_{i=0}^{n} (c_i - d_i) p_i \), which implies the zero polynomial is a non-zero linear combination of the set \( \{p_i\}_{i=0}^{n} \).

By expanding all polynomials \( p_i \) into the standard monomials \( \{x^i\}_{i=0}^{n} \), we obtain that the zero polynomial is non-zero linear combination of monomials, which is not possible since from the Fundamental Theorem of all polynomial of degree \( j \) has exactly \( j \) complex zeros including their multiplicity. Then we have the set \( \{p_i\}_{i=0}^{n} \) is a basis for \( \mathcal{P}_n(x) \).

By Lemma 2.1, we have \( \mathcal{B}_2 = \{p_0, p_1, p_2, \ldots, p_n\} \) defines another basis for \( \mathcal{P}_n(x) \). Then, for each \( j = 0, \ldots, n \), we write, as in (2.2),

\[
p_j = \sum_{i=0}^{n} t_{ij} x^i
\]

10 A Statistical Model to Describe the Spread of Wafer Lenses.
where $t_{ij} \in \mathbb{R}$, $t_{ij} = 0$ for $i > j$ and $t_{jj} \neq 0$ since $\deg p_j = j$. Denoting $\beta$ and $\alpha$ as the coordinate vector with respect to the basis $\mathcal{B}_1$ respectively $\mathcal{B}_2$ and by arranging the $t_{ij}$ values in a $k \times k$ matrix we obtain the basis transformation matrix $T$ to transform $\alpha$ to $\beta$. In the same way, as described earlier, we can also find the basis transformation matrix $\tilde{T}$ to transform $\beta$ to $\alpha$. So we can write

$$\beta = T^t \alpha \quad \text{and} \quad \alpha = \tilde{T}^t \beta.$$  

### 2.2 The Linear Model and Linear Regression

Suppose we have a population on which each element has two characteristics, say $y$ and $x$. We are often interested in understanding how the values of $x$ affect $y$. That is, we want to model $y$, a variable response, as a functional relation of $x$, and explanatory variable. Consider then a linear model with the explanatory variable $x$. The notation for the model is as follows

$$y = \beta_1 f_1(x) + \beta_2 f_2(x) + \ldots + \beta_k f_k(x) + \varepsilon$$  

(2.6)

where

- $x = x$, is the explanatory variable,
- $x \in \mathcal{X}$ and $\mathcal{X}$ is called the experimental region, with $\mathcal{X} \subseteq \mathbb{R}^m$,
- $f_i : \mathcal{X} \to \mathbb{R}$, a continuous mapping from $\mathcal{X}$ into $\mathbb{R}$, for $i = 1, \ldots, k$ and
- $\varepsilon$ is a random variable called the error term.

The model described in (2.6) might involve non-linear functions $f_j$ but it is linear in the parameters $\beta_j$'s. For instance, consider one single explanatory variable, that is $x = x$, let $\mathcal{X} = \{x | a \leq x \leq b\}$ with $b > a$ and let $f_j(x) = x^{j-1}$ for $j = \{1, 2, 3\}$. Then the model is given by a polynomial of second order of the form

$$y = \beta_1 + \beta_2 x + \beta_3 x^2 + \varepsilon$$

We assume that $x = x$ and the experimental region $\mathcal{X} = \{x | a \leq x \leq b\}$ with $b > a$. In this setting, $x_i$ represent an element of a discrete partition of the interval $[a, b]$. Let us assume that we want to model a set of $n$ paired observations $(x, y)$ under the relation described by (2.6). Then, each paired observation $(x_i, y_i)$, corresponding to the $i^{th}$ observation of $(x, y)$, is assumed to be of the form

$$y_i = \beta_1 f_1(x_i) + \beta_2 f_2(x_i) + \ldots + \beta_k f_k(x_i) + \varepsilon_i$$  

(2.7)

11 A Statistical Model to Describe the Spread of Wafer Lenses.
This is, we have a set of \( n \) expressions that read as

\[
\begin{align*}
y_1 &= \beta_1 f_1(x_1) + \beta_2 f_2(x_1) + \ldots + \beta_k f_k(x_1) + \varepsilon_1 \\
y_2 &= \beta_1 f_1(x_2) + \beta_2 f_2(x_2) + \ldots + \beta_k f_k(x_2) + \varepsilon_2 \\
\vdots &= \vdots + \vdots + \ldots + \vdots + \vdots & + \ldots + \vdots \\
y_n &= \beta_1 f_1(x_n) + \beta_2 f_2(x_n) + \ldots + \beta_k f_k(x_n) + \varepsilon_n
\end{align*}
\]

This set of expressions can be arranged as matrix form by

\[
\begin{bmatrix}
y_1 \\
y_2 \\
y_3 \\
\vdots \\
y_n
\end{bmatrix} =
\begin{bmatrix}
f_1(x_1) & f_2(x_1) & f_3(x_1) & \cdots & f_k(x_1) \\
f_1(x_2) & f_2(x_2) & f_3(x_2) & \cdots & f_k(x_2) \\
f_1(x_3) & f_2(x_3) & f_3(x_3) & \cdots & f_k(x_3) \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
f_1(x_n) & f_2(x_n) & f_3(x_n) & \cdots & f_k(x_n)
\end{bmatrix}
\begin{bmatrix}
\beta_1 \\
\beta_2 \\
\beta_3 \\
\vdots \\
\beta_k
\end{bmatrix} +
\begin{bmatrix}
\varepsilon_1 \\
\varepsilon_2 \\
\varepsilon_3 \\
\vdots \\
\varepsilon_n
\end{bmatrix}
\quad (2.8)
\]

2.2.1 Matrix notation

Four matrices are needed to express the model in matrix notation

\( \mathbf{Y} \): The \( n \times 1 \) column vector containing the observation \( y_i \).

\( \mathbf{X} \): The \( n \times k \) matrix containing the values \( f_j(x_i) \), usually called the design matrix.

\( \beta \): The \( k \times 1 \) column vector containing the parameters to be estimated, i.e. the \( \beta_i \)'s.

\( \varepsilon \): The \( n \times 1 \) column vector of random errors.

With these definitions, using matrix notation, the arrangement of expressions given in (2.8), can be compactly expressed as

\[
\mathbf{Y} = \mathbf{X}\beta + \varepsilon. \quad (2.9)
\]

This compact description of (2.8) is a overdetermined system with \( n > k \). It is usually assumed that each of the \( \varepsilon_i \) are identically distributed with \( \mathbb{E}[\varepsilon_i] = 0 \) and \( \mathbb{V}[\varepsilon_i] = \sigma^2 \), and we easily find, under multivariate notation, that \( \mathbb{E}[\varepsilon] = 0 \) and \( \mathbb{V}[\varepsilon] = \sigma^2 \mathbb{I} \), where \( \mathbb{I} \) the identity matrix of size \( n \). From (2.7) we see that each of the \( y_i \)'s are a linear combination of \( \varepsilon_i \), then, in matrix form from (2.9), we can easily find that

\[
\begin{align*}
\mathbb{E}[\mathbf{Y}] &= \mathbb{E}[\mathbf{X}\beta] + \mathbb{E}[\varepsilon] = \mathbf{X}\beta \quad (2.10a) \\
\mathbb{V}[\mathbf{Y}] &= \mathbb{V}[\mathbf{X}\beta] + \mathbb{V}[\varepsilon] = \sigma^2 \mathbb{I} \quad (2.10b)
\end{align*}
\]
For the case of a random vector of \( n \) entries, as \( Y \), we have that \( \nabla[Y] \) is no longer a value but a matrix which contains the variances of each element of \( Y \) in the diagonal while in the off-diagonal elements are placed the corresponding covariances of the elements of \( Y \).

Before discussing the estimation procedure, we treat basic concepts which are of vital importance to the development of the estimators for \( \beta \).

Let \( \Omega \) be the linear space in which \( Y \) and \( f_j(x) \) for \( j = 1, 2, \ldots, k \), take their values. The space \( \Omega \) is usually taken to be \( \mathbb{R}^n \), the collection of all \( n \)-tuples of real numbers for a positive \( n \).

**Definition 2.4.** Let \( \mathcal{P} \) be the linear space of polynomials in the real parameter \( x \) with real coefficients. Let \( f, g \in \mathcal{P} \). Then we call

\[
(f, g) = \int_a^b f(x)g(x)dx
\]

(2.11)

the inner product of \( f \) and \( g \) w.r.t. \( [a, b] \) for \( a, b \in \mathbb{R} \) with \( b > a \).

Formula (2.11) refers to the continuous definition of inner product. We can also define a discrete inner product as

\[
(f, g) = \sum_{s=1}^{n} f(x_s)g(x_s)
\]

(2.12)

this means, we take discrete partition of \( x \) over the interval \( [a, b] \), that is \( x = \{x_1, x_2, \ldots, x_n\} \).

If we take \( f, g \) as in (2.6), evaluated as in (2.12) for the values given in (2.8) we have that

\[
(f_i, f_j) = \sum_{s=1}^{n} f_i(x_s)f_j(x_s)
\]

is the inner product of the columns of the design matrix \( X \). If we let \( \Omega \) be a linear space in which either (2.11) or (2.12) is defined, then we say \( \Omega \) is an inner product linear space.

We also have

\[
(f_i, f_i) = \sum_{e=1}^{n} f_i^2(x_e)
\]

**Definition 2.5.** We call \( \| . \| \) the norm operator and it is defined as \( \| . \| = \sqrt{\langle ., . \rangle} \).

Then, we say \( \| f \| = \sqrt{\langle f, f \rangle} \) is the norm of \( f \).

**Definition 2.6.** Let \( f, g \in \mathcal{P} \). The we say \( f, g \) are orthogonal polynomials if \( \langle f, g \rangle = 0 \).

**Definition 2.7.** Let \( f_j \in \mathcal{P} \) for \( j = 1, 2, \ldots, k \), to be evaluated in \( x = \{x_1, x_2, \ldots, x_n\} \), i.e. \( f_1, f_2, \ldots, f_p \) are \( k \) vectors in an \( n \)-dimensional space. Then the subspace spanned by \( f_j \) is the collection of all vectors \( y = b_1f_1 + b_2f_2, \ldots + b_kf_k \) for real numbers \( b_j \)’s. This subspace is denoted by \( \mathcal{L}(f_1, f_2, \ldots, f_k) \).

13 A Statistical Model to Describe the Spread of Wafer Lenses.
2.2.2 The Least Squares estimation

The principle of Least Squares uses the criterion that the estimation $\hat{\beta}$ of $\beta$ minimizes the sum of squared errors of observations $y_i$ from $\hat{y}_i$. If we write $X_i = (f_1(x_i), f_2(x_i), \ldots, f_k(x_i))$, i.e. $X_j$ denotes the $i^{th}$ row of the design matrix we have that the Least Squares criterion reads as

$$\min_{\beta} \sum_{i=1}^{n} |y_i - X_i \beta|^2 = \min_{\beta} \| Y - X \beta \|^2.$$  \hspace{1cm} (2.13)

Letting $X\beta \in \mathcal{V} = \mathcal{L} (f_1, f_2, \ldots, f_k)$, i.e. $\mathcal{V}$ is the subspace spanned by the columns of the design matrix, we find that the principle of Least Squares leads to an estimator which is the projection of $Y$ on the space spanned by the functions $f_j$, that is, $X\hat{\beta} = \hat{Y}$ (See for instance Stapleton (2009) Section 1.6).

**Theorem 2.1.** The Least Squares estimator by the Gauss-Markov theorem yields an unbiased estimator $\hat{\beta}$ for $\beta$ with minimal variance.

For more details on the Gauss-Markov theorem and for its proof the reader is referred to Stapleton (2009), pag. 87.

The normal equations

In order for $\hat{Y} = X\hat{\beta}$ to be the projection of $Y$ on $\mathcal{V} = \mathcal{L} (f_1, f_2, \ldots, f_k)$, it is required that $\langle Y, f_j \rangle = \langle \hat{Y}, f_j \rangle$ for all $j$. This leads to the so-called **normal equations** described as

$$\langle \hat{Y}, f_j \rangle = \sum_{i=1}^{k} \hat{\beta}_i \langle f_i, f_j \rangle = \langle Y, f_j \rangle \quad \text{for} \quad j = 1, \ldots, k.$$ \hspace{1cm} (2.14)

This yields a system of $k$ simultaneous equations than can be compactly written down in matrix form as

$$M\beta = U$$

where $M$ is a $k \times k$ matrix containing the inner products among the $f_j$ functions and $U$ is a $k \times 1$ column vector containing the inner products of $Y$ with the $f_j$'s. That is we have

$$M_{i,j} = \langle f_i, f_j \rangle \quad \text{and} \quad U_i = \langle f_i, Y \rangle$$

with $M_{i,j}$ being the $i, j$ element of $M$ and $U_i$ being the $i$ element of $U$, and then the compact form of the normal equations can be replaced by

$$(X^t X)\beta = X^t Y$$ \hspace{1cm} (2.15)

where $M = X^t X$ and $U = X^t Y$.

14 A Statistical Model to Describe the Spread of Wafer Lenses.
The matrix $M = X'X$ and the vector $U = X'Y$ have the explicit forms of

$$M = \begin{bmatrix} \sum f_1(x_i)f_1(x_i) & \sum f_1(x_i)f_2(x_i) & \cdots & \sum f_1(x_i)f_k(x_i) \\ \sum f_2(x_i)f_1(x_i) & \sum f_2(x_i)f_2(x_i) & \cdots & \sum f_2(x_i)f_k(x_i) \\ \vdots & \vdots & \ddots & \vdots \\ \sum f_k(x_i)f_1(x_i) & \sum f_k(x_i)f_2(x_i) & \cdots & \sum f_k(x_i)f_k(x_i) \end{bmatrix}$$

and

$$U = \begin{bmatrix} \sum f_1(x_i)y_i \\ \sum f_2(x_i)y_i \\ \vdots \\ \sum f_k(x_i)y_i \end{bmatrix}.$$

Properties of the matrix $M$ are strongly related with those of the design matrix $X$. Some of such properties include invertibility which allows to find estimators. Additionally, under the choice of functions $f_j$ with additional properties, the matrix $M$ posses a desirable structure which allows to draw important conclusions in the statistical framework as will show later.

### The Least Squares estimator

The solution of (2.13) corresponds to the solution of the normal equations (2.14). The unique solution of the normal equations, if one exists, is called The least Squares Estimator for $\beta$ and denoted as $\hat{\beta}$. It may be computed explicitly from (2.15). If the matrix $M = X'X$ is invertible, we have

$$\hat{\beta} = (X'X)^{-1}X'Y. \quad (2.16)$$

In order to check that $(X'X)^{-1}$ in fact does exist, we need first:

**Definition 2.8.** The null space of a $m \times n$ matrix $A$, denoted $N(A)$, is the collection of vectors $x \in \mathbb{R}^n$ such that $Ax = 0$.

**Definition 2.9.** The column, or range, space of a $m \times n$ matrix $A$, denoted $C(A)$, is the collection of vector $x \in \mathbb{R}^m$ such that $x = Ab$ for some $b \in \mathbb{R}^n$.

We are now able to show sufficient conditions for the existence of the Least Squares Estimator.

**Theorem 2.2.** The $k \times k$ matrix $M$ has rank $k$ if and only if the columns of $X$ are linearly independent.

**Proof.** It can be shown that $C(X'X) = C(X')$ (see Stapleton (2009) Proof of Theorem 1.6.9). From linear algebra, as recall by Stapleton (2009) pag. 36, it is known that the dimensions of the row and columns spaces of any matrix $X$ are equal. This dimension is usually called rank of $X$. Then, we conclude, that $X$, $X'$, $X'X$ and $XX'$ all have the same rank. Take in particular $M = X'X$, then it has full rank if and only if $X$ has full columns rank, i.e. $X$ has linearly independent columns. \qed

15 A Statistical Model to Describe the Spread of Wafer Lenses.
Application of Theorem 2.2 on (2.15) implies that \((X'X)^{-1}\) does exist when we consider linear independent functions in the model (2.6). In practice, (2.16) is a theoretical formula that should not be used for actual computations because of instability in finding the inverse. However, the formula is very useful to prove theoretical properties of the Least Squared estimator. Several methods in finding the inverse are proposed to compute the Least Squares estimates, among them the Cholesky decomposition and QR decomposition. Cholesky decomposition performs faster than the QR decomposition, however the latter shows better numerical properties. The standard procedure used in linear regression to compute the Least Squares estimates is by means of the QR decomposition on the matrix \(M = X'X\) which avoids the inversion in (2.16).

Observe that Theorem 2.2 not only implies existence of the Least Squares estimators but also uniqueness. Uniqueness of \(\hat{\beta}\) also follows from an independency requirement in the columns of \(X\), that is, the functions \(f_j(x)\) of (2.6) should be independent. This implies full rank in \(M\) and non-singularity in \(M\) is immediate. Solving (2.13) is an optimization problem that can be solved under any optimization procedure. The linear regression approach is the most straightforward procedure to perform Least Squares with only numerical instabilities in computing the inverse involved in \(\hat{\beta}\).

### 2.2.3 Estimation by Maximum Likelihood

As described earlier in Theorem 2.1, the Least Squares method yields the best linear unbiased estimator for the parameters in linear regression regardless the distributional form of the error term \(\varepsilon\). In order to make inferences on the estimations and model quality, it is usual to assume normality in the error term. However, when a specific form of the distribution of the error is known the method of Maximum Likelihood can be used to find estimations of the model parameters.

**Definition 2.10.** Let \(W_1, W_2, \ldots, W_n\) be \(n\) independent random variables coming from an unknown distribution function \(f(\cdot; \theta)\). The likelihood function is derived from the joint density function of the observed values \(w_1, w_2, \ldots, w_n\) and it is defined as

\[
\mathcal{L}(\theta, W) = f(w_1, w_2, \ldots, w_n; \theta) = \prod_{i=1}^{n} f(w_i; \theta) \tag{2.17}
\]

where \(\theta = [\theta_1, \theta_2, \ldots, \theta_m]\) have all unknown values.

The maximum likelihood estimates \(\theta\), say by \(\hat{\theta}\), such that (2.17) is maximized, although need not to be unique. In practice, instead of maximizing (2.17), it is the logarithmic likelihood function which is often maximized. The logarithmic likelihood function, usually
called log-likelihood function, is defined as
\[
    l(\theta, W) = \log \mathcal{L}(\theta, W) = \sum_{i=1}^{n} \log f(w_i, \theta).
\]

As the logarithm function is monotonically increasing, the maximum of (2.17) and (2.18) is attained for the same values of \( \theta \). In the case of the linear model, by considering the parameters as variables and the observations as fixed at their observed values we achieve equal estimators under the normal assumption in the error.

**Lemma 2.2.** In a linear model as in (2.6), the Maximum Likelihood estimation equals the Least Squares estimation when the error term \( \varepsilon \) is assumed to be normally and independent distributed with \( \mathbb{E}[\varepsilon] = 0 \) and \( \mathbb{V}[\varepsilon] = \sigma^2 I \).

**Proof.** Recall the linear model and write the error assumption as follows
\[
    Y = X\beta + \varepsilon \quad \text{with} \quad \varepsilon \sim \text{NID}(0, \sigma^2 I).
\]

We find then that \( Y \sim \text{NID}(X\beta, \sigma^2 I) \). Using the density function of a multivariate normal distribution with mean \( \mu = X\beta \) and covariance matrix \( \Sigma = \sigma^2 I \), it follows that the log-likelihood function (2.18) is given by
\[
    l(Y, X, \beta, \sigma^2) = -\frac{n}{2} \log \left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2} (Y - X\beta)'(Y - X\beta).
\]

The Maximum Likelihood estimations are given by the solution of
\[
    \frac{\partial l}{\partial \beta} = \frac{1}{\sigma^2} X'(Y - X\hat{\beta}) = 0 \quad \text{(2.19)}
\]
\[
    \frac{\partial l}{\partial \sigma^2} = \frac{1}{2\sigma^4} (Y - X\hat{\beta})'(Y - X\hat{\beta}) - \frac{n}{2\sigma^2} = 0. \quad \text{(2.20)}
\]

From (2.19) we have that \( X'Y = X'X\hat{\beta} \), which follows that
\[
    \hat{\beta} = (X'X)^{-1} X'Y
\]

and substituting back in (2.20) we obtain
\[
    \hat{\sigma}^2 = \frac{1}{n} (Y - X\hat{\beta})'(Y - X\hat{\beta}).
\]

\[\square\]

17 A Statistical Model to Describe the Spread of Wafer Lenses.
Thus we have that the Maximum Likelihood estimation for $\beta$ equals the corresponding Least Squares Estimator. In this setting it can be proven that $\hat{\sigma}^2$ is a biased estimator for $\sigma^2$. That is $\mathbb{E}[\hat{\sigma}^2] \neq \sigma^2$; however this estimator is asymptotically unbiased which means that the bias tends to 0 as $n$ becomes large. The Maximum Likelihood estimators have better statistical properties than the Least Squares Estimators, for instance, the estimators have minimum asymptotic variance when compared with all other unbiased estimators. A quite useful property of the Maximum Likelihood Estimators is the invariance property under parametrization of the model parameters which is stated next as Lemma.

**Lemma 2.3.** Let $\hat{\theta}$ be the Maximum Likelihood estimator of $\theta$ and $g$ a function. Then $g(\hat{\theta})$ is the Maximum Likelihood estimator of $g(\theta)$.

**Proof.** if $g$ is a one-to-one function, then we have that $g^{-1}(g(x)) = x$ for any $x$ in the domain of $g$, then, recalling Definition 2.10, we have

$$L(\theta, W) = L(g^{-1}(g(\theta, W)))$$

is equally maximized by $\hat{\theta}$, thus

$$\hat{\theta} = g^{-1}(g(\theta)) \quad \text{or} \quad g(\hat{\theta}) = g(\theta).$$

The statement remains even if $g$ is a many-to-one function. \hfill \square

The matrix $M = X'X$ is usually called the **Fisher Information matrix** because it contains a measuring of the amount of information that the design matrix carries out about the unknown parameters $\beta$ (See for instance Silvey (1975) subsection 2.9, pag. 37). The role of the Fisher information matrix in the asymptotic theory of maximum-likelihood estimation is emphasized by its use in calculating the variance-covariance matrix associated with maximum-likelihood estimates.

### 2.2.4 Properties of $\hat{\beta}$

Estimation under the Least Squares criterion yields an unbiased estimator with minimum variance (see Gauss-Markov Theorem 2.1). Let $A = (X'X)^{-1}X'$, then the explicit Formula (2.16) for the estimator of $\beta$ can be written as $\hat{\beta} = AY$, showing that the estimations of the model parameters are linear combination of the random observations $Y$ with coefficients given by the entries of $A$. This also implies that $\hat{\beta}$ is itself a random...
variable and by using properties of expectation and variance we find

$$E[\hat{\beta}] = A E[Y]$$
$$= (X'X)^{-1} X' X \beta$$  \hspace{1cm} \text{Using (2.10a)}

$$\text{Var}[\hat{\beta}] = A \text{Var}[Y] A'$$
$$= \sigma^2 I A A'$$  \hspace{1cm} \text{Using (2.10b)}
$$= \sigma^2 (X'X)^{-1} X' \left[ (X'X)^{-1} X' \right]'$$
$$= \sigma^2 (X'X)^{-1}.$$  \hspace{1cm} (2.21a)

That $\hat{\beta}$ is an unbiased estimator for $\beta$ is shown in (2.21a) and it denotes the accuracy of the estimations while (2.21b) gives a measure on the precision of the estimations. In literature, it is often denoted $C = (X'X)^{-1}$ so we can write $\text{Var}[\hat{\beta}] = \sigma^2 C$ and it is called the variance-covariance matrix. Since $M = X'X$ is a symmetric matrix, so is $C$. In this way, we can find that the variances of the estimations are given in the diagonal of the matrix $C$ in the same order as the $\beta_j$’s are placed in the vector $\beta$, i.e. $\text{Var}[\hat{\beta}_i] = \sigma^2 C_{ii}$, and the covariances are given in the off-diagonal entries of $C$, this is $\text{Var}[\hat{\beta}_i, \hat{\beta}_j] = \sigma^2 C_{ij}$ for $i \neq j$.

The variance $\sigma^2$ of the error term is usually estimated by

$$\hat{\sigma}^2 = S^2 = \frac{\|Y - X\hat{\beta}\|^2}{n - k} = \frac{\sum_{i=1}^{n} (y_i - X_i \hat{\beta})^2}{n - k}. \hspace{1cm} (2.22)$$

The numerator of (2.22) is usually called the Sum of Squared Errors (SSE), while $S^2$ is called the Mean Squared Error (MSE). The usage of the MSE value is highly important when testing, among others, model quality, significant coefficients and model assumptions. Working out the expression for the Sum of Squared Errors we have that

$$SSE = \|Y - X\hat{\beta}\|^2 = (Y - X\hat{\beta})' (Y - X\hat{\beta})$$
$$= Y'Y - 2 \hat{\beta}' X' Y + \hat{\beta}' X' X \hat{\beta}$$
$$= Y'Y - \hat{\beta}' X' Y$$

so the variance $\sigma^2$ of the error term can be also estimated with

$$\hat{\sigma}^2 = \frac{Y'Y - \hat{\beta}' X' Y}{n - k}. \hspace{1cm} (2.23)$$

The estimator of the variance is usually with the Root Mean Square which we defined next.

A Statistical Model to Describe the Spread of Wafer Lenses.
Definition 2.11. Let \( g_i, i = 1, 2, \ldots, n \) be some quantities defined over a discrete set of points, the Root Mean Square (RMS) is defined by

\[
RMS(g_i) = \sqrt{\frac{\sum_{i=1}^{n} g_i^2}{n}}.
\]

(2.24)

Definition 2.12. Let \( g \) be a function defined over a continuous region \( \Omega \). The RMS is then defined in terms of integrals as

\[
RMS(g) = \sqrt{\frac{\int_{\Omega} g(x)^2 \, da}{\int_{\Omega} da}}.
\]

(2.25)

The Root Mean Square (RMS) is an statistical measure of the magnitude of a varying quantity that when it is applied to the model residuals \( y_i - X_i \hat{\beta} \), it can be linked to the MSE of the regression model. In connection with the estimator of the variance (2.22) we can find both quantities, the RMS and the MSE, to be related by

\[
\sqrt{MSE} = \sqrt{\frac{n}{n-k}} RMS.
\]

The quantity \( \sqrt{MSE} \) is known as the standard deviation estimator and it is denoted by \( \hat{\sigma} \). Therefore, in general, \( \hat{\sigma} \geq RMS \). For data sets larger than 100, the factor in the square root above is almost 1, then we can assume \( \hat{\sigma} \approx RMS \) of the residuals.

On the other hand, apart from the covariances between the estimated coefficients, finding the correlation among them is more of useful to quantify the level of agreement (disagreement) of the estimated coefficients.

Definition 2.13. Let \( X, Y \) be two random variables, then the correlation coefficient of \( X \) and \( Y \), often denoted as \( Corr(X,Y) \), or simply \( \rho_{X,Y} \) is defined as

\[
Corr(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}
\]

(2.26)

where \( \sigma_X \) and \( \sigma_Y \) are the standard deviations of \( X \) respectively \( Y \).

We have that \( \hat{\beta} \) is a random vector with variance-covariance matrix \( \nabla[\hat{\beta}] = \hat{\sigma}^2 C \). If we define \( W \) to be a diagonal matrix with the square root of the diagonal entries of \( C \), that is \( W_{ii} = \sqrt{C_{ii}} \) for \( i = 1, 2, \ldots, k \) and \( W_{ij} = 0 \) for \( i \neq j \) we can write

\[
Corr[\hat{\beta}] = (\hat{\sigma} W)^{-1} \nabla[\hat{\beta}] (\hat{\sigma} W)^{-1} = W^{-1} C W^{-1}.
\]

(2.27)
The matrix $W$ has the explicit form

$$
W = \begin{bmatrix}
\sqrt{C_{11}} & 0 & 0 & \cdots & 0 \\
0 & \sqrt{C_{22}} & 0 & \cdots & 0 \\
0 & 0 & \sqrt{C_{33}} & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & \sqrt{C_{kk}}
\end{bmatrix}
$$

(2.28)

Formula (2.27) gives a symmetric $k \times k$ matrix which can be described by

$$
[W^{-1}CW^{-1}]_{i,j} = \begin{cases} 
1, & \text{if } i = j \\
\text{Corr}[\hat{\beta}_i, \hat{\beta}_j], & \text{if } i \neq j.
\end{cases}
$$

2.2.5 Orthogonality

There are, in practice, some inconveniences when analyzing data under a linear model. For instance, if two or more $f_j$ functions in (2.6) are highly correlated it is said that multicollinearity is present in the data. Multicollinearity gives an intuition in the near-linear dependency of the functions in the linear model. The presence of multicollinearity can seriously disturb the Least Squares estimates and reduce, in some cases, the linear model to be almost useless when the regressors in the model are nearly perfectly related. Additionally, if we would like to add a new function to the linear model (2.6), say by $\hat{\beta}_{k+1}f_{k+1}(x)$, to check whether the new function contributes in explaining the variable response, we must recompute $(X'X)^{-1}$ and the Least Squares estimates of the previous parameters will change. Important relations and consequences on the estimations will be found when considering the functions $f_j$ in model (2.6) to be orthogonal, that is, as described in Definition 2.6,

$$
\langle f_i, f_j \rangle = \int_a^b f_i(x)f_j(x)dx = 0
$$
in the continues case and

$$
\langle f_i, f_j \rangle = \sum_{e=1}^N f_i(x_e)f_j(x_e) = 0
$$
in the discrete case. Now suppose the model given in (2.6) satisfies the above described orthogonality condition (either the continuous or discrete case) with the design matrix $X$ given in (2.9) has the form

$$
X = \begin{bmatrix}
f_1(x_1) & f_2(x_1) & f_3(x_1) & \cdots & f_k(x_1) \\
f_1(x_2) & f_2(x_2) & f_3(x_2) & \cdots & f_k(x_2) \\
f_1(x_3) & f_2(x_3) & f_3(x_3) & \cdots & f_k(x_3) \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
f_1(x_n) & f_2(x_n) & f_3(x_n) & \cdots & f_k(x_n)
\end{bmatrix}.
$$
Since this matrix has orthogonal columns, the matrix \( M = X^tX \) is diagonal and looks like
\[
M = \begin{bmatrix}
\sum f_j^2(x_i) & 0 & 0 & \ldots & 0 \\
0 & \sum f_j^2(x_i) & 0 & \ldots & 0 \\
0 & 0 & \sum f_j^2(x_i) & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & 0 & \sum f_j^2(x_i)
\end{bmatrix}
\] (2.29)

and the Least Squares estimators are given by
\[
\hat{\beta}_j = \frac{\sum_{i=1}^n f_j(x_i)y_i}{\sum_{i=1}^n f_j^2(x_i)} = \frac{\langle f_j, y \rangle}{\langle f_j, f_j \rangle} \quad \text{for } j = 1, 2, \ldots, k.
\] (2.30)

Utility of having orthogonal functions lies in the fact of finding independent/uncorrelated explicit formulas for the regression coefficients \( \beta_j \). Importance of having orthogonal functions in the statistical context comes next.

**Lemma 2.4.** Let the columns of the design matrix \( X \) be orthogonal. Then the variance-covariance matrix is a diagonal matrix and then the Least Squares estimators are uncorrelated.

**Proof.** Consider the variance-covariance matrix given in (2.21b). Since \( X \) has orthogonal columns then \( M = X^tX \) is diagonal as given in (2.30). Hence its inverse \( C = (X^tX)^{-1} \) is also diagonal. Therefore we have \( C_{ij} = 0 \) for \( i \neq j \). That is \( \forall \hat{\beta}_i, \hat{\beta}_j \) = 0 for \( i \neq j \).

Then, under linear regression analysis when using orthogonal functions one can model the effect of each function on the variable response separately, and then combine these results in a model which predicts the combined effect with adding no extra information. That is, the estimation of \( \beta_s \) does not influence the estimation of \( \beta_r \) for \( r \neq s \) as shown in (2.30).

We can observe, as pointed out by Silvey (1969), that uncorrelated variables not only can be achieved by considering directly orthogonal functions in model (2.6) but also by using a suitable reparametrization over a non-orthogonal set of functions \( f_j \). Correlations on the estimations can be removed either by performing such reparametrization as a “orthogonal” basis transformation on a coordinate vector like \( \hat{\beta} \) or by performing the “orthogonal” basis transformation on the design matrix \( X \). It turns out that both procedures are equivalent.

### 2.2.6 Example Simple linear regression

To illustrate all the fitting procedure under the framework previously described we present an example. Assume, we have the data in Table 2.1. Then, suppose we want to model

22 A Statistical Model to Describe the Spread of Wafer Lenses.
these \( n = 3 \) paired observations a straight line. That is, by setting \( f_1(x) = 1 \) and \( f_2(x) = x \) we can write the model as

\[
y = \beta_0 + \beta_1 x + \epsilon
\]

Using the matrix approach, we have

\[
X = \begin{bmatrix} 1 & 4 \\ 1 & 6 \\ 1 & 8 \end{bmatrix} \quad Y = \begin{bmatrix} 15 \\ 20 \\ 28 \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}
\]

The normal equations \( (X'X)\beta = X'Y \) are

\[
\begin{bmatrix} 1 & 1 & 1 \\ 4 & 6 & 8 \end{bmatrix} \begin{bmatrix} 1 & 4 \\ 1 & 6 \\ 1 & 8 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 15 \\ 20 \\ 28 \end{bmatrix}
\]

\[
\begin{bmatrix} 3 & 18 \\ 18 & 116 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 60 \\ 380 \end{bmatrix}
\]

and \( \hat{\beta} = (X'X)^{-1} X'Y \) reads as

\[
\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \frac{1}{24} \begin{bmatrix} 116 & -18 & -18 \\ -18 & 3 & 3 \end{bmatrix} \begin{bmatrix} 60 \\ 380 \end{bmatrix} = \frac{1}{24} \begin{bmatrix} 36 \\ 78 \end{bmatrix}
\]

We compute \( \hat{\sigma}^2 = Y'Y - \hat{\beta}'X'Y \) as

\[
\hat{\sigma}^2 = \begin{bmatrix} 15 & 20 & 25 \end{bmatrix} \begin{bmatrix} 15 \\ 20 \\ 25 \end{bmatrix} - \frac{1}{24} \begin{bmatrix} 36 & 78 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 4 & 8 \end{bmatrix} \begin{bmatrix} 15 \\ 20 \\ 25 \end{bmatrix}
\]

\[
= 3/2
\]

also we have that \( C = (X'X)^{-1} \), so

\[
C = \frac{1}{24} \begin{bmatrix} 116 & -18 & -18 \\ -18 & 3 & 0 \end{bmatrix} \quad \text{and} \quad W = \frac{1}{\sqrt{24}} \begin{bmatrix} \sqrt{116} & 0 \\ 0 & \sqrt{3} \end{bmatrix}
\]
Then we compute $\hat{V}[\hat{\beta}] = \hat{\sigma}^2 \hat{C}$ and $\text{Corr}[\hat{\beta}] = W^{-1}CW^{-1}$ as

$$
\hat{V}[\hat{\beta}] = \frac{1}{24} \begin{bmatrix} 116 & -18 & -18 \\ -18 & 3 & -18 \\ -18 & -18 & 3 \end{bmatrix} ; \quad \text{Corr}[\hat{\beta}] = \begin{bmatrix} 1 & -0.96 & -0.96 \\ -0.96 & 1 & -0.96 \\ -0.96 & -0.96 & 1 \end{bmatrix}
$$

We notice then, that the estimated coefficients when a non-orthogonal basis is used are correlated. In fact, we see that $\text{Corr}[\hat{\beta}_0, \hat{\beta}_1] = -0.96$.

Now assume, that instead of using $f_1 = 1$ and $f_2 = x$ we would like to use $\tilde{f}_1 = 1$ and $\tilde{f}_2 = x - \bar{x}$. Then we can easily verify that $\tilde{f}_1$ and $\tilde{f}_2$ are orthogonal functions. To check this we use (2.12), then

$$
\langle \tilde{f}_1, \tilde{f}_2 \rangle = \sum_{s=1}^n \tilde{f}_1(x_s)\tilde{f}_2(x_s) = \sum_{s=1}^n (x_s - \bar{x}) = \sum_{s=1}^n x_s - n\bar{x} = 0
$$

and the model has the form

$$
y = \gamma_0 + \gamma_1 x + \tilde{\epsilon}
$$

When using a orthogonal basis, as explain before, we obtain uncorrelated estimators. We proceed now with the corresponding calculations. Recomputing the design matrix we have

$$
\tilde{X} = \begin{bmatrix} 1 & -2 \\ 1 & 0 \\ 1 & 2 \end{bmatrix} \quad Y = \begin{bmatrix} 15 \\ 20 \\ 28 \end{bmatrix} \quad \tilde{\epsilon} = \begin{bmatrix} \tilde{\epsilon}_1 \\ \tilde{\epsilon}_2 \\ \tilde{\epsilon}_3 \end{bmatrix} \quad \text{and} \quad y = \begin{bmatrix} \gamma_0 \\ \gamma_1 \end{bmatrix}
$$

The normal equations $(\tilde{X}^t \tilde{X}) \gamma = \tilde{X}^t Y$ are

$$
\begin{bmatrix} 1 & 1 & 1 \\ -2 & 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & -2 \\ 1 & 0 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} \gamma_0 \\ \gamma_1 \end{bmatrix} = \begin{bmatrix} 15 \\ 20 \\ 28 \end{bmatrix}
$$

and $\hat{\gamma} = (\tilde{X}^t \tilde{X})^{-1} \tilde{X}^t Y$ reads as

$$
\begin{bmatrix} \hat{\gamma}_0 \\ \hat{\gamma}_1 \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 84 \\ 26 \end{bmatrix} = \frac{1}{4} \begin{bmatrix} 84 \\ 13 \end{bmatrix}
$$

We compute $\hat{\sigma}^2 = Y^t Y - \hat{\gamma}^t \tilde{X}^t Y$ as

$$
\hat{\sigma}^2 = \begin{bmatrix} 15 & 20 & 28 \end{bmatrix} \begin{bmatrix} 15 \\ 20 \\ 28 \end{bmatrix} - \frac{1}{4} \begin{bmatrix} 84 \\ 13 \end{bmatrix} \begin{bmatrix} 15 \\ 20 \\ 28 \end{bmatrix}
$$

$$
= \frac{3}{2}
$$

24 A Statistical Model to Describe the Spread of Wafer Lenses.
also we find that 

$$\tilde{C} = (\tilde{X}^T \tilde{X})^{-1},$$

so

$$\tilde{C} = \begin{bmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{8} \end{bmatrix} \quad \text{and} \quad \tilde{W} = \begin{bmatrix} \frac{1}{\sqrt{3}} & 0 \\ 0 & \frac{1}{\sqrt{8}} \end{bmatrix}$$

Then we compute

$$\nabla[\hat{\gamma}] = \hat{\sigma}^2 \tilde{C}$$

and

$$\text{Corr}[\hat{\gamma}] = \tilde{W}^{-1} \tilde{C} \tilde{W}^{-1}$$

as

$$\nabla[\hat{\gamma}] = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{3}{16} \end{bmatrix} ; \quad \text{Corr}[\hat{\gamma}] = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Note that we could also use (2.30) to find the Least Squares estimators. Contrary with what was found previously, we notice then, that the estimated coefficients when a orthogonal basis is used are uncorrelated. In fact, we see that

$$\text{Corr}[\hat{\gamma}_0, \hat{\gamma}_1] = 0.$$  

2.3 Summary of the Chapter

We presented a procedure to perform a coefficient transformation from the standard monomials to another different set of polynomials as well as computational formulas for the estimation of coefficients along with their variance-covariance when a linear regression analysis has been performed. The change of basis transformation developed in Section 2.1 allows to describe the profile of a lens in an arbitrary set of linear combinations of polynomials. Generalities of the linear regression framework presented in Section 2.1 provides formulas to compute the estimation of coefficients in a linear regression model as well as the implications of using a orthogonal set of polynomials in the regression model. We have found that estimated coefficients when using a non-orthogonal set of polynomials are correlated and those correlations depends directly from the measurements \((x, z)\) in a lens. On the other hand, when using a orthogonal set of polynomials, the linear regression framework yields a set of non-correlated estimated coefficients. Both, the non-orthogonal and orthogonal linear regressions, contains a very accurate description of a lens shape. With the analysis developed in this chapter, we attempt to derive expressions to account the variability of a given set of estimated coefficients when a specific orthogonal basis transformation is performed under a set of estimated coefficients.
3 Data Description

This chapter briefly discusses characteristics on data collection and preprocessing in order to set data ready for model implementation. Analysis on the measuring mechanism and will give insight on how data is obtained, what preprocessing on measurements is done and finally what data is used for the analysis. In a two-dimensional setting the data basically consists in a set of $n$ paired observations $(x_1, z_1), \ldots, (x_n, z_n)$ which represents the measurements of the profile of a lens done by a “Line scan” machine. Based on this data, a profile polynomial is found under a curve fitting procedure and such polynomial is compared with the original design. The comparison operates under the basis of the RMS value of the residuals (See Definition 2.11).

Lens selection and Measuring

The first step in collecting data is a Lens selection. This is, select the lenses within a wafer to be measure. A typical wafer is depicted in Figure 3.1. A wafer is usually labeled as in a geographical system. The geographical labeling allows a tracking of a lens trough the production steps. Additionally, for ease recognition and tracking, the lenses to be measured can be numbered in advance so locating them is immediate. The total number of lenses to be measured is fixed and its selection mostly responds the research going on and to production constraints. Figure 3.2 presents some possible lens selection schemes in which some special patterns are given and also random patterns can be considered. Choice of either one or another depends on the problem under investigation. For instance, if the interest is to find radial patterns in a wafer, a proper lens selection scheme would be like the one in bottom-left of Figure 3.2.

A close look into Figure 3.1 allows to notice that each lens is located symmetrically respect to the vertical or horizontal axis of the wafer, for instance, the most left lens is located symmetrically with the most right lens respect to the vertical axis. Similarly, the lens in the bottom is located symmetrically respect to the lens in the top using as reference the horizontal axis. This symmetry property on the lenses within a wafer makes easier the selection/location of lenses for the measuring process.

Once the lenses to be measure are selected, the measuring process starts. Measure-
ments are done with a line scan machine able to measure and output values representing the profile of the lens. The machine, e.g. Talysurf, is able to measure dimension, form and texture on curved surfaces. The Talysurf maps for a fixed $x$ position its corresponding height $z$. Figure 3.3 shows how the Talysurf operates. The Talysurf is able to perform profile measurements on a radial line over the lens and then pair values on the $x$ and $z$ axis are output. The $x$-spacing if fixed from the Talysurf specification and the number of data points vary from lens to lens. The measurements represent the two-dimensional cross-section of a lens and the analysis of variability is based on these measurements.

27 A Statistical Model to Describe the Spread of Wafer Lenses.
Data preprocessing

Once a lens is measured, data can be available in the form of heights as a function of the radial position. This initial data is automatically plotted for initial inspection by the analyst\(^1\) can take a first look into the data. The initial data collected, when plotted, looks like Figure 3.4. This initial examination on the data roughly shows that the measurements present a tilt and that they are not centered in a proper coordinate system. The tilt and decenter are features of the measurements that must be corrected for the future analysis. Therefore, based on the initial data, and in order to compare the measurements with the design, the coordinate system of the measurements must be rotated and shifted. The rotation step, called “detilt”, removes the tilt of the data, and the shifting, which is called “centering”, places the top of the measurements in the origin of

\(^{1}\text{The analyst is the person in charge of making the measurements.}\)

28 A Statistical Model to Describe the Spread of Wafer Lenses.
the standard cartesian coordinate system. The entire data preprocessing is depicted in Figure 3.5.

Storing data

Each lens, as shown in Lens A of Figure 1.1, has a Clear Aperture (CA) that indicates the maximum aperture for which performance of the lens is not affected by diffraction or reflection of light beams through the lens. Once the detilting and decentering in the Talysurf data is performed, both, the total data and data from the CA is stored as .txt files which contains in the name the wafer number, the lens number and additional information that can be used for the later analysis. It is usually observed that the data at the edge of the (a)spherical surface and the horizontal plane presents a lot of variability due to the Talysurf constrains in the measuring. Nevertheless, the curve fitting procedure is based on data corresponding only to the CA and such inconvenient is no longer a source of variability in the analysis.
4 Estimation in the Cross-Section of a Lens

The design of a lens is often described in the cross-section of a lens and it is available in polynomial form in the radius ($x$ position). This description is specified given customer requirements for which, before production is started, a design has to be made. Such design is often given in polynomial form in one variable in which the height of a lens is function of the radial position. This design is in accordance with Definition 2.2 where $\deg p = d$ for some positive integer $d$. Such profile as description of a lens is assumed to be possible because a lens is considered to be symmetrical rotational, that is, when the profile depicted in Figure 4.1 is rotated around the $z$ axis, the total lens surface is produced. The symmetric assumption allows seeing the height ($z$) as a function of the radial position ($x$) only, that is, we have $z = p(x)$. In this chapter, the standard approach to curve fitting performed by Anteryon will be explained. This approach makes use of the well-known Least Squares criterion to compute coefficients of a specified model followed by a basis matrix transformation to find the desired radial Zernike polynomials\(^1\). We choose to perform the fitting of the cross-section data by using the statistical framework of linear regression analysis. Since Least Squares is deterministic, no further information regarding the quality of the model and accuracy of the estimations can be done. Therefore, by using a linear regression analysis not only the coefficients themselves are obtained but also a series of statistical tests can be performed in order to

\(^1\)See Appendix A for more details in the Zernike polynomials.

A Statistical Model to Describe the Spread of Wafer Lenses.
test the quality of the model used, the significance of the coefficients as well as find confidence bounds for those, among others possible tests.

We first start by describing the current estimation procedure adopted by Anteryon, followed by a comparison between the Anteryon procedure and a direct approach avoiding the matrix transformation. Next, some regression diagnostics is presented and finally different estimation procedures are described.

4.1 Estimation Procedure

Unless indicated otherwise, we use the following notation:

- \( x \) — the independent variable of a polynomial expression, which varies over the length of the Clear Aperture (CA).

- \( h_p(x) \) — the polynomial design which follows Definition 2.2. It is often given as

\[
    h_p(x) = \sum_{i=0}^{d} c_i x^i. \tag{4.1}
\]

- \( c \) — a column vector corresponding to the coefficients of the polynomial design. It is given by

\[
    c = [c_0, c_1, \ldots, c_d]^T.
\]

- \( d \) — the degree of the polynomial design, that is \( \text{deg} \; h_p = d \).

- \( h_f(x) \) — the polynomial used to fit the measured data per lens. It is often described by as

\[
    h_f(x) = \sum_{i=0}^{p} \beta_i x^i. \tag{4.2}
\]

- \( \beta \) — a column vector corresponding to the coefficients of polynomial used to fit the measured data. It is given by

\[
    \beta = [\beta_0, \beta_1, \ldots, \beta_p]^T.
\]

- \( p \) — the degree of the fitted polynomial, that is \( \text{deg} \; h_f = p \).

- \( h_z(x) \) — the polynomial used to represent measured data in term of radial Zernike polynomials. It is often described by as

\[
    h_z(x) = \sum_{i=0}^{p} \alpha_i R_i(x). \tag{4.3}
\]
where $i$ is a multi-index that maps the integers $i$ to $(n, m)$ tuples of the radial Zernike polynomials and $R_i(x)$ is a radial Zernike polynomial such that $\deg R_i = i$.

- $\alpha$ — a column vector corresponding to the coefficients the radial Zernike polynomials. It is given by
  $$\alpha = [\alpha_0, \alpha_1, \ldots, \alpha_p]^T.$$

### 4.1.1 Polynomial estimation

Suppose $n$ paired observations $(x_i, z_i)$ are measured in a lens. The polynomial model to represent the height (variable response) as a function of the position (explanatory variable) is often written down as

$$z_i = h_f(x_i) = \beta_0 x_i^0 + \beta_1 x_i^1 + \beta_2 x_i^2 + \ldots + \beta_p x_i^p + \varepsilon_i \tag{4.4}$$

where $h_f(x_i)$ represents the height at radial position $x_i$, the $\beta_j$'s denotes the polynomial coefficients, $p$ is the polynomial degree and $\varepsilon_i$ denotes the random error term. It is often assumed that $\varepsilon_i$ to be identical and normal distributed with mean 0 and variance $\sigma^2$ (denoted as $\varepsilon_i \sim N ID(0, \sigma^2)$). In correspondence with (2.6), we have:

- $x$ is the explanatory variable.
- $X = [a, b]$ for $a, b \in \mathbb{R}$ is the experimental region with $x \in X$.
- $f_j = x^{j-1}$ for $j = 1, 2, \ldots, k = p + 1$, i.e. $f_i$ corresponds to the monomials.
- $\varepsilon_i$ the random error term.

The matrix notation for the linear model given by (4.4) which yields a polynomial of degree $p$ reads exactly as (2.9) with

$$Y = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 & \cdots & x_1^p \\ 1 & x_2 & x_2^2 & x_2^3 & \cdots & x_2^p \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & x_n^3 & \cdots & x_n^p \end{bmatrix}, \quad \beta = [\beta_0, \beta_1, \ldots, \beta_p]^T, \quad \varepsilon = [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n]^T.$$ 

$Y$ is a column vector of size $n \times 1$, the design matrix $X$ is of size $n \times k$, the vector of unknowns $\beta$ is of size $k \times 1$ and the vector of error terms $\varepsilon$ is of size $n \times 1$. The Least Squares criterion then reads as

$$\min_\beta \sum_{i=1}^n |z_i - X_i \beta|^2 = \min_\beta \|Y - X \beta\|^2$$

32 A Statistical Model to Describe the Spread of Wafer Lenses.
with $X_i$ the $i^{th}$ row of the design matrix, that is $X_i = [1, x_i, x_i^2, \ldots, x_i^p]$. Therefore, with the given notation, the Least Squares Estimator $\hat{\beta}$ of $\beta$ is computed as in (2.16) with

$$
X'X =
\begin{bmatrix}
    n & \sum x_i & \sum x_i^2 & \cdots & \sum x_i^p \\
    \sum x_i & \sum x_i^2 & \sum x_i^3 & \cdots & \sum x_i^{p+1} \\
    \sum x_i^2 & \sum x_i^3 & \sum x_i^4 & \cdots & \sum x_i^{p+2} \\
    \vdots & \vdots & \vdots & \ddots & \vdots \\
    \sum x_i^p & \sum x_i^{p+1} & \sum x_i^{p+2} & \cdots & \sum x_i^{2p}
\end{bmatrix}
$$

(4.5)

and

$$
X'Y =
\begin{bmatrix}
    \sum z_i \\
    \sum z_ix_i \\
    \sum z_ix_i^2 \\
    \vdots \\
    \sum z_ix_i^p
\end{bmatrix}
$$

(4.6)

Uniqueness of the Least Squares estimator is given as long as the columns of the design matrix $X$ has linearly independent columns. In polynomial estimation, the design matrix $X$ has the well-known structure of a Vandermonde matrix (See Horn and Johnson (1994) Section 6.1) which guarantees linear independency of the columns of $X$ provided that there are at least $p + 1$ distinct values of $x_i$. Independency of the columns of $X$ immediately yields existence $C = (X'X)^{-1}$. By finding the values of $\hat{\beta}$ we substitute back in (4.4) to obtain

$$
\hat{h}_f(x) = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 + \ldots + \hat{\beta}_p x^p \quad \forall \ i
$$

(4.7)

which is called the estimated polynomial height function. The model predictions are then obtained by

$$
\hat{z}_i = \hat{h}_f(x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 + \ldots + \hat{\beta}_p x_i^p \quad \forall \ i
$$

and in matrix form

$$
\hat{Y} = X\hat{\beta} = X (X'X)^{-1} X'Y
$$

In literature, it is often denoted $H = X (X'X)^{-1} X'$, so the model predictions can be written down as $\hat{Y} = X\hat{\beta} = HY$. The matrix $H$ is called the hat matrix and we easily show that it symmetric and idempotent, in fact

$$
H' = \left[ X (X'X)^{-1} X' \right]' = X (X'X)^{-1} X' = H
$$

$$
H^2 = X (X'X)^{-1} X' X (X'X)^{-1} X' = X (X'X)^{-1} X' = H.
$$
4.1.2 Basis transformation

It is of interest to find the representation of the cross-section estimation performed in previous subsection in terms of the well-known radial Zernike Polynomials. For doing so, a basis transformation is performed to find the required coefficients. Letting \( P_p(x) \) be the finite-dimensional polynomial space of dimension \( k = p + 1 \) and by letting \( \hat{h} \) be in \( P_p(x) \) we have found by with \( \hat{\beta} \) the coordinates of \( \hat{h} \) with respect to the basis \( B_1 = \{ 1, x, x^2, \ldots, x^p \} \). Therefore, the element \( \hat{h} \) has a unique expansion in the basis \( B_1 \) which given by (4.7).

Consider now the well-known radial Zernike polynomials \( R^m_n(x) \) first introduced by Zernike (1934)\(^2\). The most general expression of these radial polynomials is given by

\[
R^m_n(x) = \sum_{s=0}^{\frac{n-m}{2}} (-1)^s \frac{(n-s)!}{(\frac{n+m}{2}-s)! (\frac{n-m}{2}-s)!} x^{n-2s}
\]  \hspace{1cm} (4.8)

where \( n \) and \( m \) are both integral values and satisfy \( n - m \) even. Evaluation of (4.8) for particular values of \( n \) and \( m \) yields a polynomial of degree exactly \( n \). Anteryon currently uses a specific selection of the radial Zernike polynomials in order to represent the lens profile. The radial Zernike polynomials for which \( m = \{0, 1\} \) for all required \( n \) are selected, that is, from the entire set of radial Zernike polynomials only are selected

- \( R^0_n(x) \) when \( n \) is even,
- \( R^1_n(x) \) when \( n \) is odd.

In this setting, the multi-index \( i \) used in (4.3) maps integers \( i \) to \( (n, m) \) tuples in the following way

\[ R_i(x) \rightarrow R_i^{(\text{mod } 2)}(x) \]

and the corresponding coefficients associated as

\[ \alpha_i \rightarrow \alpha_i^{(\text{mod } 2)}. \]

Define \( f_j = R_j(x) \) for \( j = 0, 1, \ldots, p \), then \( f_j \) is a polynomial of degree exactly \( j \), and by Lemma 2.1 we have that \( B_2 = \{ f_0, f_1, f_2, \ldots, f_p \} \) defines another basis for \( P_p(x) \). By plugging in the required values of \( (n, m) \), as the multi-index used in the basis \( B_2 \), in (4.8) we obtain explicit formulas for the \( R_j(x) \) as

\[
\begin{align*}
 f_0 &= R_0(x) & \rightarrow & R^0_0(x) = 1 \\
 f_1 &= R_1(x) & \rightarrow & R^1_1(x) = x \\
 f_2 &= R_2(x) & \rightarrow & R^0_2(x) = 2x^2 - 1 \\
 f_3 &= R_3(x) & \rightarrow & R^1_3(x) = 3x^3 - 2x \\
 f_4 &= R_4(x) & \rightarrow & R^0_4(x) = 6x^4 - 6x^2 + 1 \\
 \vdots & \vdots & \rightarrow & \vdots = \vdots
\end{align*}
\]

\(^2\)See Appendix A for more details on the radial Zernike polynomials.
In this way, as in (2.2), we have

\[ f_j = \sum_{i=1}^{k} t_{ij} x_i^{j-1} \]

where the \( t_{ij} \)'s are given by (4.8). Therefore, the basis transformation matrix \( T \) is

\[
T = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 & \cdots \\
0 & 1 & 0 & 0 & 0 & \cdots \\
1 & 0 & 2 & 0 & 0 & \cdots \\
0 & -2 & 0 & 3 & 0 & \cdots \\
1 & 0 & -6 & 0 & 6 & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix}
\]

Denoting \( \alpha = [\alpha_0, \alpha_1, \ldots, \alpha_p]^T \) as the coordinate vector with respect to the basis \( B_2 \) we can write

\[ \beta = T^T \alpha. \]

This is

\[
\begin{bmatrix}
\beta_0 \\
\beta_1 \\
\beta_2 \\
\beta_3 \\
\vdots \\
\beta_p
\end{bmatrix} =
\begin{bmatrix}
1 & 0 & -1 & 0 & 1 & \cdots \\
0 & 1 & 0 & -2 & 0 & \cdots \\
0 & 0 & 2 & 0 & -6 & \cdots \\
0 & 0 & 0 & 3 & 0 & \cdots \\
0 & 0 & 0 & 0 & 6 & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix}
\begin{bmatrix}
\alpha_0 \\
\alpha_1 \\
\alpha_2 \\
\alpha_3 \\
\vdots \\
\alpha_p
\end{bmatrix}
\]

The interest lies in transforming the coordinate vector \( \beta \) to the coordinate vector \( \alpha \). Then we have that \( \tilde{T} = T^{-1} \) is the corresponding basis transformation matrix to do so and we write

\[ \alpha = (T^T)^{-1} \beta \]

This is

\[
\begin{bmatrix}
\alpha_0 \\
\alpha_1 \\
\alpha_2 \\
\alpha_3 \\
\vdots \\
\alpha_p
\end{bmatrix} =
\begin{bmatrix}
1 & 0 & \frac{1}{3} & 0 & \frac{1}{3} & \cdots \\
0 & 1 & \frac{2}{3} & 0 & \frac{2}{3} & \cdots \\
0 & 0 & \frac{1}{3} & 0 & \frac{1}{3} & \cdots \\
0 & 0 & \frac{1}{3} & 0 & \frac{1}{3} & \cdots \\
0 & 0 & 0 & \frac{1}{6} & 0 & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix}
\begin{bmatrix}
\beta_0 \\
\beta_1 \\
\beta_2 \\
\beta_3 \\
\vdots \\
\beta_p
\end{bmatrix}
\]

Since we have that \( \hat{\beta} \) is the estimator of \( \beta \) it follows that \( \hat{\alpha} \) is the estimator of \( \alpha \) by the invariance principle (Lemma 2.3). We can then write

\[ \hat{h}_z(x) = \hat{\alpha}_0 R_0(x) + \hat{\alpha}_1 R_1(x) + \hat{\alpha}_2 R_2(x) + \ldots + \hat{\alpha}_p R_p(x) \]
and we can therefore conclude that given any coordinate vector with respect to the standard basis we can find the respective coordinate vector with respect to the radial Zernike basis. A complete set of functions \( f_j \) for the basis \( B_2 \) up to \( j = 25 \) is given in Table B.3. The corresponding matrix transformation for the given basis is given in Table B.4. The corresponding matrix inverse is given in Table B.5.

### 4.1.3 Structure of the basis transformation matrix \( T \)

The basis transformation matrix \( T \) presented in the previous subsection has a special structure given by the evaluation of (4.8). As shown earlier, the basis transformation matrix \( T \) allows to transform coordinate vector with respect to the radial Zernike basis to the coordinate vector with respect to the monomial basis and vice versa. The restriction \( n - m \) even, imposed in (4.8) yields a sparse matrix which can be depicted as follows:

\[
T = \begin{bmatrix}
* & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & * & 0 & 0 & 0 & 0 & \cdots \\
* & 0 & * & 0 & 0 & 0 & \cdots \\
0 & * & 0 & * & 0 & 0 & \cdots \\
* & 0 & * & 0 & * & 0 & \cdots \\
0 & * & 0 & * & 0 & * & \cdots \\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\
\end{bmatrix}
\]

The basis transformation matrix is a squared lower triangular matrix of size \( k \) with \( k = p + 1 \). We notice the matrix \( T \) is non-singular since the elements in the diagonal are all non-zero. This follows since \( R_m^n(x) \) is a polynomial of degree exactly \( n \). Because of that, we can guarantee the existence of \( T^{-1} \) and the basis transformation is invertible, that is we can compute the coordinate vector from one basis to another and vice versa. Given the sparse structure of \( T \), it is not hard to see that also \( T^{-1} \) have the same structure. This allows to choose a sub-matrix when a lower degree polynomial is being in consideration. For instance, if a \( p' \) degree polynomial with \( p' < p \) is being considered, then the corresponding basis transformation \( T' \) is the squared upper-left block of size \( k' \) with \( k' = p' + 1 \). The same follows when computing the inverse of \( T' \).

### 4.2 Basis Transformation versus Direct Estimation

In this section, we aim to describe whether it is useful to perform, instead of a matrix basis transformation to the estimated coefficients in the monomial basis, a direct estimation in the radial Zernike polynomials. Anteryon currently uses a two-step estimation procedure in which, a polynomial of certain degree is fitted and then, under a basis transformation matrix, the so-called radial Zernike polynomial coefficients are computed.
Since the radial Zernike polynomials are themselves a basis, the basis transformation can be avoided and the radial Zernike coefficients can be estimated directly. For the description, we will call Anteryon Estimation the two-step procedure used by Anteryon and Direct Estimation to the estimation directly in radial Zernike polynomials.

### 4.2.1 Anteryon Estimation

The corresponding model, similar to (2.6), reads as

\[ Y = X\beta + \epsilon_1 \]  \hspace{1cm} (4.9)

where \( Y \) is the vector of observations, the design \( X \) matrix corresponds to the monomials, and the error term is denoted with a sub-index 1 with the usual assumption \( \epsilon_1 \sim N I D(0, \sigma^2_1) \). The Least Squares estimator for \( \beta \) from this model was shown to be of the form of

\[ \hat{\beta} = (X'X)^{-1}X'Y \]

provided that \((X'X)^{-1}\) exists. Following the procedure described in Subsection 2.2.4, by letting \( A_1 = (X'X)^{-1}X' \), we see that \( \hat{\beta} \) is a linear combination of random observations \( Y \), then we can find

\[ \mathbb{E}[\hat{\beta}] = A_1\mathbb{E}[Y] \]
\[ = A_1X\beta \]
\[ = (X'X)^{-1}X'X\beta \]
\[ = \beta \]
\[ \mathbb{V}[\hat{\beta}] = A_1\mathbb{V}[Y]A_1' \]
\[ = (X'X)^{-1}X'\sigma^2_1[(X'X)^{-1}X']' \]
\[ = \sigma^2_1(X'X)^{-1}. \]

After finding the estimates of the polynomial coefficients a basis matrix transformation is performed in order to compute the radial Zernike coefficients. Following earlier procedure, denoting as \( T \) the basis matrix transformation from the radial Zernike polynomials to the standard monomials, and letting \( \alpha \) be the coordinate vector with respect to the radial Zernike polynomials, as described in Subsection 4.1.2, we have that

\[ \hat{\alpha} = L\hat{\beta} \]
where, for ease in notation, we have taken \( L = (T')^{-1} \). Given this, we can easily find that

\[
\begin{align*}
\mathbb{E}[\hat{\alpha}] &= L\mathbb{E}[\hat{\beta}] \\
&= L\beta \\
\mathbb{V}(\hat{\alpha}) &= L\mathbb{V}(\hat{\beta})L' \\
&= \sigma_1^2L(X'X)^{-1}L'
\end{align*}
\]

where \( C = (X'X)^{-1} \). We also notice that \( \hat{\alpha} \) is an unbiased estimator for \( \alpha \). We have then that \( \sigma_1^2 \) is estimated by \((n - p - 1)\hat{\sigma}_1^2 = Y'Y - \hat{\beta}'X'Y\) as from (2.23).

### 4.2.2 Direct Estimation

The direct estimation of the radial Zernike coefficients under the Least Squares framework works exactly the same by modifying the design matrix \( X \). The corresponding model as follows

\[
Y = \tilde{X}\gamma + \epsilon_2
\]

(4.10)

where \( Y \) is the vector of observations, the design matrix \( \tilde{X} \) corresponds to the radial Zernike polynomials, and the error term is denoted with a sub-index 2 with the usual assumption \( \epsilon_2 \sim \text{NID}(0, \sigma_2^2) \). Notice we have used the symbol \( \gamma \) to denote the regression radial Zernike coefficients and to make them different from those found of the basis matrix transformation. The Least Squares estimator for \( \gamma \) from this model is given by

\[
\hat{\gamma} = \left(\tilde{X}'\tilde{X}\right)^{-1}\tilde{X}'Y
\]

provided that \( \left(\tilde{X}'\tilde{X}\right)^{-1} \) does exist. In this setting, we denote \( A_2 = \left(\tilde{X}'\tilde{X}\right)^{-1}\tilde{X}' \) and see that \( \hat{\gamma} \) is also a linear combination of random observations \( Y \) and under the model given by (4.10), we can find

\[
\begin{align*}
\mathbb{E}[\hat{\gamma}] &= A_2\mathbb{E}[Y] \\
&= \left(\tilde{X}'\tilde{X}\right)^{-1}\tilde{X}'\tilde{X}\gamma \\
&= \gamma \\
\mathbb{V}[\hat{\gamma}] &= A_2\mathbb{V}[Y]A_2' \\
&= \left(\tilde{X}'\tilde{X}\right)^{-1}\tilde{X}'\sigma_2^2I\left[\left(\tilde{X}'\tilde{X}\right)^{-1}\tilde{X}'\right]' \\
&= \sigma_2^2\tilde{C}
\end{align*}
\]

where \( \tilde{C} = \left(\tilde{X}'\tilde{X}\right)^{-1} \). In this case, \( \sigma_2^2 \) is estimated by \((n - p - 1)\hat{\sigma}_2^2 = Y'Y - \hat{\gamma}'\tilde{X}'Y\).
4.2.3 Comparison

We are interested in finding whether both procedures, the given in Subsection 4.2.1 and the given in Subsection 4.2.2, are similar in estimation and variances. Not presented here, it can be shown that the design matrices from both procedures are related by

\[ \tilde{X} = XT' \]  \hspace{1cm} (4.11)

where \( T \) is as mentioned before. We also have that \( L = (T')^{-1} \). Given so, we now can write

\[ \hat{\alpha} = L\hat{\beta} \]
\[ = (T')^{-1} (X'X)^{-1} X'Y \]

and

\[ \hat{\gamma} = \left( \tilde{X}'\tilde{X} \right)^{-1} \tilde{X}'Y \]
\[ = \left[ (XT')' (XT') \right]^{-1} (XT')' Y \]
\[ = (X'XT')^{-1} (T)^{-1} TX'Y \]
\[ = (T')^{-1} (X'X)^{-1} X'Y. \]

Therefore we have \( \hat{\alpha} \) equals \( \hat{\gamma} \) meaning that both procedures lead to the same estimates. On the other hand, regarding the variances of the estimations we have

\[ \text{V}[\hat{\alpha}] = LV[\hat{\beta}]L' \]
\[ = \sigma_1^2 (TX'XT')^{-1} \]

and

\[ \text{V}[\hat{\gamma}] = \sigma_2^2 \left( \tilde{X}'\tilde{X} \right)^{-1} \]
\[ = \sigma_2^2 \left[ (XT')' (XT') \right]^{-1} \]
\[ = \sigma_2^2 (TX'XT')^{-1}. \]

Therefore we have \( \text{V}[\hat{\alpha}] \) equals \( \text{V}[\hat{\gamma}] \) if \( \sigma_1^2 = \sigma_2^2 \). Recalling that

\[ \hat{\sigma}_1 = \frac{Y'Y - \hat{\beta}'X'Y}{n - p - 1} \]
\[ \hat{\sigma}_2 = \frac{Y'Y - \hat{\gamma}'\tilde{X}'Y}{n - p - 1} \]
it follows that $\sigma_1^2 = \sigma_2^2$ if and if

$$Y^t Y - \hat{\beta}^t X^t Y = Y^t Y - \hat{\gamma}^t \tilde{X}^t Y$$

$$\hat{\beta}^t X^t Y = \hat{\gamma}^t \tilde{X}^t Y.$$ 

In the left-hand side we have

$$\hat{\beta}^t X^t Y = Y^t X (X^t X)^{-1} X^t Y$$

and in the right-hand side

$$\hat{\gamma}^t \tilde{X}^t Y = Y^t \tilde{X} (\tilde{X}^t \tilde{X})^{-1} \tilde{X}^t Y$$

$$= Y^t (XT^t) [(XT^t)^t (XT^t)]^{-1} (XT^t)^t Y$$

$$= Y^t XT^t [TX^t XT^t]^{-1} TX^t Y$$

$$= Y^t X (X^t X)^{-1} X^t Y.$$ 

Hence computing the radial zernike coefficients by means of a basis matrix transformation is, in terms of the MSE $(\hat{\sigma}^2)$ values, is equivalent to compute directly the radial Zernike coefficients when the Least Squares criterion is used. Additionally, both fitting procedures yield correlated coefficients in which the increment/decrement on the estimator variances are given by the basis matrix transformation.

### 4.2.4 Comments

As mention earlier (Subsection 4.1.2 and Subsection 4.2.2) the coordinate vector $\beta$ with respect to the monomials basis are random variables. It follows the coordinate vector $\alpha$ with respect to the radial Zernike polynomials basis are also random variables.

The expressions used for the variance-covariance matrix (2.21b) shows that the estimated coefficients are correlated unless a orthogonal basis in the linear model (2.6) is used. In polynomial regression it is known that the monomials do not conform a orthogonal basis. That is, by using Definition 2.6, we have that

$$\langle x^i, x^j \rangle \neq 0 \text{ for } i = 0 \text{ and } j = 1.$$ 

This statement can be also verified in the columns of the design matrix for polynomial regression where can be seen that

$$\langle x^i, x^j \rangle = \sum_{i=1}^{n} (x_i)^j (x_i)^i \neq 0$$

for some $i, j$ columns of $X$, when using the discrete inner product formula given in (2.12). This type of inner products can be explicitly found in the information matrix as the one
given in (4.5). Silvey (1969) showed that under a proper reparametrization we can obtain uncorrelated estimations. Such proper reparametrization transforms the design matrix in the polynomial regression into a new design matrix such that its columns are orthogonal. That is, a reparametrization that looks like (4.11) can transform a non-orthogonal set into a orthogonal one. However, when using the matrix transformation as the one described in Subsection 4.1.2 we found that such reparametrization yields inner products that looks like
\[
\langle R_m^n(x), R_{m'}^{n'}(x) \rangle = \sum_{s=1}^{n} R_m^n(x_s) R_{m'}^{n'}(x_s)
\]
for the chosen values of \(n, n', m, m'\) as in Subsection 4.1.2 in the information matrix given by \(\tilde{X}^t \tilde{X}\). Such inner products are not in accordance with the orthogonality characteristic of the radial Zernike polynomials (as presented in Appendix A) which, in a discrete setting, reads as
\[
\sum_{s=1}^{n} x_s R_m^n(x_s) R_{m'}^{n'}(x_s) = a_m^n \delta_{n,n'}
\]
where \(a_m^n\) is a normalization constant and \(\delta_{i,j}\) is the well-known Kronecker delta function. Notice the usage of the same value for \(m\) in the orthogonality requirement of the radial Zernike polynomials. Hence we conclude, that under the current basis matrix transformation used by Anteryon, no uncorrelated estimators of the radial Zernike coefficients can be obtained.

### 4.3 Diagnostics

The statistical framework of linear regression analysis allows to obtain not only the coefficients of any linear model but also allows to draw conclusions on the reliability of such estimations and to perform regression diagnostics on the model assumptions. A list of the standard regression diagnostics, as given by Montgomery et al. (2001) and Rawlings et al. (1998), is given below.

- Start with an explorative analysis of the data.
- Perform the regression calculations.
- Test the significance of the model equation.
- Test the significance of the model parameters.
- Perform a Lack-of-Fit test when replicated measurements are present.
- Check multicollinearity.
• Inspect residual plots.
• Check normality on the residuals.
• Inspect the presence of outliers.
• Check independency on the residuals
• Inspect the presence of influential points.

Fully regression diagnostics can be perform in detail when requested. As shown in Section 4.2, the estimation procedure in standard polynomials ($\beta$) is equivalent to the estimation directly in the radial Zernike polynomials ($\alpha$). Hence, the regression diagnostics can be either perform over one or another depending on the preferred choice. For practical reasons, the standard polynomial setting is chosen and the most relevant diagnostics are presented.

Before formally present the regression diagnostics, we first describe the output of the software used in the analysis and present an inspection regarding the model orden selection.

**Software Used**

There are in the market several software available to perform curve fitting procedures. We choose to use the statistical software R (see [www.r-project.org](http://www.r-project.org)). This is free, open source software that is becoming the standard for data analysis. There are thousands of add-ins (libraries) that are publicly available through the R-project web site and that are useful in several statistical areas of application. Linear models like (4.4) are usually performed in R using the standard function `lm`. This is a built-in R function which yields a linear model object, including apart the model coefficients, among other things, the RMS value and the residuals of the fit. We encounter no software limitation in the usage of the `lm` function since the data set consists usually of at most 5000 bivariate observations. The R software allows to save the results, usually refer as `lm` objects, as special `.Rdata` files that can be loaded directly into a new R session for further analysis.

An example of the output of a statistical analysis obtained by using the R software is given in Table 4.1 where a polynomial of degree 16 has been fitted. The first column gives the label of the estimated coefficient (for instance as the given in (4.4)) while the second column gives the corresponding estimate under the label `Estimate`. The standard deviation of the estimates which accounts the precision of the estimates is given in the third column under the label `Std. Error`. The t-value, corresponding to the t-student density function widely used for several statistical test is given in the fourth column as `t value`. This is the ratio of the estimate and the standard error. Generally
Table 4.1: Example of Statistical output of polynomial Fitting

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 2.787e-08 | 1.588e-06  | 0.018   | 0.98600  |
| x1             | -5.205e-05 | 5.336e-05  | -0.976  | 0.32933  |
| x2             | -8.310e-01 | 1.428e-03  | -581.928| < 2e-16 ***|
| x3             | -9.468e-03 | 2.049e-02  | -0.462  | 0.64398  |
| x14            | -3.571e+07 | 6.895e+06  | -5.180  | 2.32e-07 ***|
| x15            | -7.468e+06 | 3.234e+06  | -2.309  | 0.02097 * |
| x16            | 1.246e+08  | 2.351e+07  | 5.300   | 1.22e-07 ***|

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 1

Residual standard error: 3.155e-05 on 4382 degrees of freedom
Multiple R-squared:  1, Adjusted R-squared:  1
F-statistic: 1.029e+08 on 16 and 4382 DF,  p-value: < 2.2e-16

Speaking, t-values larger than 1.5 or 2 indicate that the parameter is statistically significant. A column labeled as Pr(>|t|) is given next to the t-values and provides a threshold value (p-value) which usually depends on the sample size. If the p-value is smaller than a standard value (0.05 is the commonly accepted value in statistics), then the corresponding parameter is declared to be significant. For convenience, significant parameters can readily be identified from this table because the last column indicates with stars which parameters are significant (the precise definition of the stars is given below the table). At the bottom of the table one can find additional information. An estimate of the square root of the variance $\sigma^2$ of (4.4) is given under the name Residual standard error. The $R^2$-values indicate how much of the variation in the data set is captured by the model and is hence an indication how well the model fits the data. The maximal value of $R^2$ is 1 (which corresponds to 100%). The last line indicates whether the model as a whole is significant using another statistical signal-to-noise ratio (the F-statistic), for which no general threshold exist and thus assessment is only possible through the given p-value.

Model Order Selection

The model order selection is often found on the basis of the RMS values of the residuals. A typical behavior of the RMS versus the order of regression of the lens data is shown in Figure 4.2.

43 A Statistical Model to Describe the Spread of Wafer Lenses.
The RMS of the residuals should decrease monotonically as the polynomial order increases. However, the RMS presents huge jumps in magnitude, for instance, when going from a polynomial of degree 1 to a polynomial of degree 2 (Figure 4.2 (a)). After 5 degree polynomials, the RMS still presents jumps, nevertheless differences are in the order of nanometers, for instance, there is a jump of 10 nm when going from a polynomial of degree 7 to a polynomial of degree 8 (Figure 4.2 (b)). We notice that after order 12 in the polynomial model the RMS presents slightly stabilization (Figure 4.2 (c)) since the improvement is of the same nanometers even though there might be more jumps after high order polynomials (Figure 4.2 (d)). This pattern seems to be similar in the regressions for all lenses. For the model order, in general, Anteryon applies a rule that reads as follows

\[
\text{Let } d \text{ be the degree of the polynomial design, then the order of the polynomial to perform the linear regression is } d + 2.
\]

With the notation described earlier, we have that if \( \deg h_p = d \) and then \( \deg h_f = d + 2 \). Under this rule, the RMS values reached per polynomial regression achieves company requirements. Therefore we found no evidence to apply another criterion for the model order selection. Therefore we decided to applied the same rule for future regressions.

A Statistical Model to Describe the Spread of Wafer Lenses.
We are now able to perform some of the regression diagnostics.

### 4.3.1 Data inspection

In Chapter 3 we have described that after the preprocessing, the data looks like Figure 4.1. A plot of the CA raw data is presented in Figure 4.3 (a) while a zoomed plot is presented in Figure 4.3 (b). The shape of the measured data gives a clear indication that may be fitted a polynomial of degree 2 or higher. For a clear visualization of a typical deviation from the measurements and polynomial design it is depicted Figure 4.4.

![Figure 4.3: Raw Data: Radius [mm] VS Height [mm]](image)

### 4.3.2 Regression Calculations

The polynomial regression is performed on the basis of the rules previously explained. That is, a polynomial $h_f(x)$ such that $\deg h_f = d + 2$ where $d$ is the degree of the polynomial design, is fitted. Results of the example presented in Table 4.1 are reproduced in Table 4.2.
4.3.3 Model significance

The most common hypothesis of interest in Multiple Linear regression, as the model given by (4.4), is to test whether there is or no useful relationship between the observed values and any of the monomials (excluding \( x^0 \)) in a polynomial of degree \( p \), that is the common test is based on \( \beta_j \), for \( j = 1, 2, \ldots, p \). In general, we can say that if at least one the \( \beta_j \)'s differs significantly from 0 the corresponding monomial is useful to explain the observations. Testing the model significance we have

\[
H_0 : \quad \beta_1 = \beta_2 = \beta_3 = \ldots = \beta_p = 0 \\
H_1 : \quad \text{At least one } \beta_j \neq 0, \quad (j = 1, 2, \ldots, p).
\]

This test is based on a statistic that has a \( F \) distribution when the \( H_0 \) is true. This test indicates whether the model as a whole is significant. The F-statistic, which is signal-to-
noise ratio, is computed as

\[ f = \frac{R^2/p}{(1-R^2)/(n-p-1)} \]

where \( R^2 \) is the coefficient of determination. It is common to use the p-value criteria to conclude about the model. The corresponding p-value is also output by the R software. From a typical output, as the one described in Table 4.1 we extract \( p\text{-value} < 2.2e-16 \) from which we can conclude, that the model under consideration is significant with an \( \alpha \)-value = 0.05.

### 4.3.4 Parameter significance

As shown earlier, the Least Squares estimator \( \hat{\beta} \) is linear combination of random observations \( z_i \)'s and so its variance is given in the diagonal of the variance-covariance matrix. Inferences related to a single estimated coefficient \( \beta_j \), namely whether the fitted parameter deviates sufficiently from 0, can be done based on a standardized variable defined as

\[ T = \frac{\hat{\beta}_j - \beta_j}{S_{\hat{\beta}_j}} \]

where the denominator is the standard deviation for a single coefficient which can be obtained from the output of the R software as in Table 4.1. The hypothesis test looks like

- \( H_0 : \beta_j = 0 \) the \( j^{th} \) parameter is not significant
- \( H_1 : \beta_j \neq 0 \) the \( j^{th} \) parameter is significant.

From the example given in Table 4.1 by using the p-value criteria we notice coefficients \( \beta_0 \) and \( \beta_1 \) are not statistically significant using a \( \alpha \)-value = 0.05.

This tests on the significance of the parameters should be assessed carefully under the presence of correlated variables. For instance, assuming that parameter \( i \) is highly correlated with parameter \( j \), the Least Squares gives equal estimates on both parameters and if one of the parameters is not significant the other is also not significant despite that the combined effect might be significant for the model under consideration.

### 4.3.5 Model Assumptions

The model assumptions involve normality and constant variance in the residuals. That is, the residuals \( \varepsilon \) are assumed to be \( NID(0, \sigma^2) \). Any violation on the normality assumption is critical for the conventional tests of hypothesis for the model as a whole and for individual parameters. Additionally, the typical construction of confidence intervals
on the estimates lose validity when the normality assumption is not fulfilled given that such intervals uses the normality assumption on the estimates. The model assumptions are usually checked using residual plots and normality tests on the residuals of the regression.

**Constant Variance**

Figure 4.5 present a plot of studentized residuals versus predicted values and versus the radial position at the top respectively bottom. It is expected the scatter plot of predicted values versus studentized residuals to be structure less and random around the drawn with horizontal reference $h = 0$. In the same way the constant variance assumption requires the residuals to be independent of the predicted values. In other words, the whole cloud of studentized residuals should lie more or less around a horizontal band. From Figure 4.5 we notice that the residuals are more or less equally spread around $h = 0$. It can be also seen that there are points that lie outside of the horizontal lines drawn at $-2.5$ and $+2.5$. This fact is a clear indication of outliers and it will be discussed later.

![Figure 4.5: Scatter Plot Residuals](image)

48 A Statistical Model to Describe the Spread of Wafer Lenses.
Normality

To investigate normality on the residuals, Figure 4.6 depicts on the left a density plot on the residuals of the regression shown in Table 4.1 while on the right the normal probability plot (Q-Q plot usually called) is given. Theoretically, when the residuals are adequately modeled by a Normal distribution, the points should fall on a straight line (Figure 4.6 right). More attention is given to points in the middle, say between the 0.25 and 0.75 percentile rather than the extreme points. The hypothesis test for normality reads as:

\[ H_0 : \text{The residuals come from a Normal distribution} \]
\[ H_1 : \text{The residuals do not come from a Normal distribution} \]

Based on the density plot (Figure 4.6 left) and in the Q-Q plot (Figure 4.6 right), we can assume that the residuals \( \varepsilon_i \) follow a Normal distribution with 0 mean and constant variance.

![Figure 4.6: Normality on the residuals](image)

Outliers

As mentioned earlier, the presence of outliers can be also investigated. In practice, when the Normal assumption is violated is due to the presence of outliers. In general, outliers should not be removed from the data, but they can be removed from large data sets or preferably they should be measured again. The presence of outliers can be also explored residuals plots, as Figure 4.5. Potential outliers can be seen as an extreme high or low point in the vertical axes of the plot. Since it is ambiguous to mention high or low values, a threshold of absolute value of the studentized residual larger than 2.5 is often used. From the residual plot (Figure 4.5), all those points can be seen above or below the horizontal dashed lines. It seems to be quite a lot, nevertheless a counting
on this outliers reveals in comparison with the total number of observations that the amount of outliers is small compared with the total data. When the linear regression is performed under data after removing outliers, it can be seen that the improvement in the RMS values and the $R^2$ is small. Additionally, an analysis in different linear regression analysis for several lenses, reveals that there is no systematic pattern in the position of the outliers.

4.3.6 Confidence Intervals

The classical confidence intervals procedures involves univariate intervals for both, the individual parameters and the expected mean for particular choices of the independent variable. Although the confidence intervals convey more information than point estimates, those do not take into account correlations among the estimators of the parameters. In order to consider the correlation, joint confidence intervals have to be computed.

![Figure 4.7: 95 % confidence bounds](image)

A confidence interval on the parameters and for the expected response is often more
informative that the point estimation because it reflects the precision of the estimates. The confidence interval on the regression parameters are usually given as a upper and lower limit indicating that the true value may fall in between. For its construction it is needed the corresponding parameter estimation and its standard deviation. The confidence interval is computed using the respective estimation at each $x$ and its standard deviation and the common way to present is by plotting. The plot convey an overall understanding on the quality for the model in representing the data in the degree on which the results can be assumed. Figure 4.7 depicts, in the top, a confidence interval for the fitted values, since no variation can be seen, in the bottom it is presented transformed confidence interval obtained by subtracting the polynomial design.

4.4 Shifting to the Origin

The current estimation procedure implemented by Anteryon involves, before the basis matrix transformation, a shifting to the origin. It means, a standard $p$ degree polynomial is fitted and then the offset coefficient is set to be 0. This step is easily seen as a shifting from the offset estimator to the origin so the top of the lens coincides with the origin of the Cartesian plane. The shifting step makes the final estimated model to be biased and may lead to $E[\hat{\beta}_0] \neq \beta_0$. Instead, the shifting step can be interpreted as regressing through the origin in which the offset $\beta_0$ is automatically set to be zero. In this section, the impact of both approaches in the MSE values is investigated.

Consider the standard matrix notation for linear model $Y = X\beta + \epsilon$ where $X$ is the design matrix and $\beta$ the regression parameters. Let $X$ be the design matrix for the polynomial regression with intercept, so the design matrix contains as the first column a vector of 1. The ordinary residuals, i.e. the differences between the observations and the model predictions (or model estimations) are given by

$$e = Y - \hat{Y}$$

where $Y$ is the vector of observations, $\hat{Y} = X\hat{\beta}$ is the vector of predictions. Then $e_i = z_i - \hat{z}_i$ is the $i^{th}$ deviation from the observed value and the model prediction. This deviation gives a measure in the degree of agreement between the data and the model used. The model prediction $\hat{z}_i$ is the $i^{th}$ predicted value of $z_i$ at $x_i$. In correspondence with (2.13) we can also define

$$e_i = z_i - X_i\hat{\beta}$$

with $X_i$ being the $i^{th}$ row of the design matrix. The Sum Squares of error can be written as

$$SSE_\beta = \sum_{i=0}^{n} e_i^2 = \sum_{i=0}^{n} |z_i - X_i\hat{\beta}|^2 = \|Y - X\hat{\beta}\|^2$$

(4.12)
The Least Squares principle makes the agreement between the data and the model in “as good as possible” sense. Notice the subindex $\beta$ in the SSE and MSE to denote the Sum of Squared errors and Mean Squared Error for the model with intercept. Another equivalent way of writing down the Sum Squares of error is given by

$$SSE_\beta = \sum_{i=1}^{n} |z_i - \hat{z}_i|^2 = \sum_{i=1}^{n} |z_i - \sum_{j=0}^{p} \hat{\beta}_j x_i^j|^2$$  \hspace{0.5cm} (4.14)$$

with $\hat{\beta}_j$ the corresponding estimations of the polynomial coefficients.

The shifting to the origin implies that the $\hat{z}_i$’s are shifted by the corresponding estimation of the offset, that is we can write a new model estimation $\hat{z}_i^S = \hat{z}_i - \hat{\beta}_0$ and then we rewrite $e_i^S = z_i - \hat{z}_i^S$. Therefore the Sum Squares of error can be redefined as

$$SSE_S = \sum_{i=1}^{n} (e_i^S)^2 = \sum_{i=1}^{n} |z_i - \hat{z}_i^S|^2$$

$$= \sum_{i=1}^{n} |z_i - \hat{z}_i + \hat{\beta}_0|^2 = \sum_{i=1}^{n} \left| z_i - \sum_{j=1}^{p} \hat{\beta}_j x_i^j \right|^2$$  \hspace{0.5cm} (4.15)$$

and the corresponding MSE is given by

$$MSE_S = \frac{SSE_S}{n - p - 1}$$  \hspace{0.5cm} (4.16)$$

with the subindex $S$ to indicate the shifted Sum Squares of errors and Shifted Mean Squared of errors. The expression for $SSE_S$ can be reduced to

$$SSE_S = \sum_{i=1}^{n} e_i^2 + 2\hat{\beta}_0 \sum_{i=1}^{n} e_i + n\hat{\beta}_0^2$$

It can be shown that $\sum_{i=1}^{n} e_i = 0$ when the offset $\beta_0$ is included in the model, hence we arrive at

$$SSE_S = SSE_\beta + n\hat{\beta}_0^2$$  \hspace{0.5cm} (4.17)$$

from we can see that $SSE_S > SSE_\beta$ always.

Consider now the linear model $Y = \tilde{X}\gamma + \epsilon$ with $\tilde{X}$ the design matrix of a polynomial regression with no intercept, $\gamma$ a regression through the origin, and $\gamma$ the regression parameters. It can shown as before that

$$SEE_\gamma = \sum_{i=0}^{n} |z_i - \tilde{X}_i\hat{\gamma}|^2 = \|Y - \tilde{X}\hat{\gamma}\|^2$$  \hspace{0.5cm} (4.18)$$

52 A Statistical Model to Describe the Spread of Wafer Lenses.
where $\tilde{X}_i$ is the $i^{th}$ row of the design matrix with no intercept. Equivalently we also have

$$SSE_\gamma = \sum_{i=1}^{n} |z_i - \sum_{j=1}^{p} \hat{\gamma}_j x_i^j|^2$$

(4.19)

and the MSE error for the model with no intercept is given by

$$MSE_\gamma = \frac{SSE_\gamma}{n - p}$$

(4.20)

Our main interest is to compare the respective SSE values for the regression with and without intercept. That is, we want to compare

1. $SSE_S$ VS $SSE_\gamma$
2. $SSE_\beta$ VS $SSE_\gamma$

Additionally of comparing the respective shifted Sum Squared of errors against the Sum Squared of errors we could test the significance of the offset $\hat{\beta}_0$ from the regression with intercept. The form of statistical test was already described in Section 4.3. This step should be perform carefully since using the polynomials through the origin may lead to a different coefficient estimators due to the non-orthogonality of the basis used for the regressions.

**Numerical Results**

The models used for the analysis are

$$h_f(x) = \sum_{i=0}^{p} \beta_i x^i ; \quad h^s_f(x) = \sum_{i=1}^{p} \beta_i x^i \quad \text{and} \quad h_\alpha(x) = \sum_{i=1}^{p} \gamma_i x^i$$

where $h^s_f(x)$ is the shifted polynomial and $h_\alpha(x)$ is the polynomial through the origin. Results presented are in terms of the RMS values instead of the corresponding MSE, since $n$, the number of observation, is quite big and we have shown in Subsection 2.2.4 that under such conditions $RMS \approx \sqrt{MSE}$. Figure 4.8 presents a model order selection when the previous regressions are performed. It can be seen, as expected by the Gauss-Markov theorem (Theorem 2.1), the RMS value achieved by the full model is minimal among all the tested polynomials, i.e., $h_f(x)$ performs always better than any other estimators.

Figure 4.9 presents a comparison of various lenses when $h_f(x)$, $h^s_f(x)$ and $h_\alpha(x)$ are performed. We can notice the changes in RMS vary from 20 nm to 80 nm for the three formulas. The changes from a model to another are very small and its behavior is...
quite similar, however, a close look in the results presented in Table 4.3 corresponding to those values of Figure 4.9 shows that, in fact, \( h^s_f(x) \) performs worst than \( h_f(x) \) and \( h_o(x) \), and that the improvement by using a model with intercept \((h_f(x))\) is of order of \( 1e-04 \) nm when compared with a model through the origin \((h_o(x))\). Not presented, the hypothesis
test on the offset $\hat{\beta}_0$ of $h_f(x)$ showed non significance for all the lenses tested. When analyzing the estimated coefficients $\hat{\beta}$ and $\hat{\gamma}$ we notice the estimated coefficients are quite similar for $h_f(x)$ and $h_o(x)$. That is, the effect of applying a regression through the origin do not severely change the estimation when the offset is included. We conclude then than it is better to use, for the polynomial regressions, polynomial model through the origin.

### 4.5 Flipping

Under the above described fitting procedures, nor the polynomial $h_f(x)$ and $h_z(x)$ are symmetric. When describing a lens, the symmetry characteristic gives a better understanding when, under the Anteryon process, a lens is manufactured. In order to obtain even polynomials, that is polynomial symmetric around the $z$-axis in Figure 4.1, Anteryon applies a “flipping” procedure which consists of the following:

1. **Left Mirroring**: Select data points in the negative $x$-axis. Mirror them in the positive $x$-axis. Estimate a polynomial under the Least Squares criterion.

2. **Right Mirroring**: Select data points in the positive $x$-axis. Mirror them in the negative $x$-axis. Estimate a polynomial under the Least Squares criterion.

3. **Averaging**: Compute a average polynomial based on the Left and Right estimations.

| L1   | 2.96138438e-05  | 2.96137061e-05  | 2.96137185e-05  |
| L2   | 3.19207130e-05  | 3.19206167e-05  | 3.19206254e-05  |
| L3   | 2.14353028e-05  | 2.14352608e-05  | 2.14352646e-05  |
| L4   | 3.89676856e-05  | 3.89676780e-05  | 3.89676787e-05  |
| L5   | 8.28694341e-05  | 8.28694276e-05  | 8.28694282e-05  |
| L6   | 9.41455176e-05  | 9.41455166e-05  | 9.41455167e-05  |
| L7   | 3.68070780e-05  | 3.68070763e-05  | 3.68070764e-05  |
| L8   | 3.61396715e-05  | 3.61396715e-05  | 3.61396715e-05  |
| L9   | 2.46089744e-05  | 2.46089085e-05  | 2.46089144e-05  |
| L10  | 2.47049065e-05  | 2.47048643e-05  | 2.47048681e-05  |
| L11  | 2.29276289e-05  | 2.29276289e-05  | 2.29276289e-05  |
| L12  | 5.05268625e-05  | 5.05268625e-05  | 5.05268625e-05  |
| L13  | 3.16153422e-05  | 3.16153422e-05  | 3.16153422e-05  |
| L14  | 2.64144494e-05  | 2.64144424e-05  | 2.64144430e-05  |

Table 4.3: RMS corresponding to Figure 4.9
In this Section, we describe in detail the Left Mirroring, Right Mirroring and we introduce two additional fitting procedures being under consideration. The fitting procedures are then

- **Left Mirroring**: Select data points in the negative $x$-axis. Mirror them in the positive $x$-axis. Estimate a polynomial under the Least Squares criterion.

- **Right Mirroring**: Select data points in the positive $x$-axis. Mirror them in the negative $x$-axis. Estimate a polynomial under the Least Squares criterion.

- **Left**: Select data points in the negative $x$-axis. Estimate a polynomial under the Least Squares criterion.

- **Right**: Select data points in the positive $x$-axis. Estimate a polynomial under the Least Squares criterion.

Numerical comparisons are made under the RMS criterion. For understanding of the flipping procedure we use the following notation.

- $n$ denotes the total points in the data set
- $L$ denotes the total negative $x_i$ points. That is $L = \sum_{i}^{n} 1_{\{x_i < 0\}}$. Then the total points for the left estimation is $n_L = 2L$.
- $R$ denotes the total positive $x_i$ points. That is $R = \sum_{i}^{n} 1_{\{x_i > 0\}}$. Then the total points for the right estimation is $n_R = 2R$.

Since it is very unlikely to have $x_i = 0$ in the data set, we have that $n = L + R$. Additionally we can have that $L > n > R$ when $L > R$; $R > n > L$ when $R > L$; and $L = n = R$ when $L = R$.

Before describing the above mentioned fitting schemes, we introduce a new Lemma which can be used later for the description.

**Lemma 4.1.** Let $p$ be a polynomial as described by Definition 2.2 and let $p \in \mathcal{P}_n(x)$. Then $p$ is an even polynomial if and only if $\mathcal{B} = \{1, x^2, x^4, \ldots, x^n\}$ with $n$ even is a basis for $\mathcal{P}_n(x)$.

**Proof.** Instead of proving Lemma 4.1 we prove that even polynomials cannot contain odd terms. Let $p$ be an even polynomial. Consider the polynomial $q$ defined by $q(x) = p(x) - p(-x)$. Then $q$ is also a polynomial which is everywhere zero. It is clear that $q$ cannot contain even powers (by construction). It cannot contain odd powers since the monomials form a basis for the vector space of polynomials. Hence the odd powers from $q$ should all be zero. Their coefficients are twice the coefficients of the odd powers of $p$ (by construction). Hence, $p$ cannot have odd powers. \qed

**56 A Statistical Model to Describe the Spread of Wafer Lenses.**
**Left Mirroring**

The data set is partitioned and arranged in a new data set in the following way

\[
\begin{align*}
\mathbf{x} &= \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_{L-1}, \bar{x}_L, \bar{x}_{L+1}, \bar{x}_{L+2}, \ldots, \bar{x}_{2R-1}, \bar{x}_{2R} \\
&= -x_n, -x_{n-1}, \ldots, -x_{L+2}, -x_{L+1}, x_{L+1}, x_{L+2}, \ldots, x_{n-1}, x_n \\
\mathbf{z} &= \bar{z}_1, \bar{z}_2, \ldots, \bar{z}_{L-1}, \bar{z}_L, \bar{z}_{L+1}, \bar{z}_{L+2}, \ldots, \bar{z}_{2R-1}, \bar{z}_{2R} \\
&= z_n, z_{n-1}, \ldots, z_{L+2}, z_{L+1}, z_{L+1}, z_{L+2}, \ldots, z_2, z_1
\end{align*}
\]

The matrix notation for the linear regression setting is \( Y = X \beta_{lm} + \varepsilon \) where the sub-index "lm" stands for **Left Mirroring** and

\[
Y = \begin{bmatrix} \bar{z}_1 \\ \bar{z}_2 \\ \vdots \\ \bar{z}_{2L} \end{bmatrix}, \quad X = \begin{bmatrix} 1 & \bar{x}_1 & \bar{x}_2 \bar{x}_3 & \ldots & \bar{x}_p \\ 1 & \bar{x}_2 & \bar{x}_3 & \bar{x}_4 & \bar{x}_5 & \ldots & \bar{x}_p \\ \vdots \\ 1 & \bar{x}_{2L} & \bar{x}_{2L} & \bar{x}_{2L} & \ldots & \bar{x}_p \end{bmatrix}, \quad \beta_{lm} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{2L} \end{bmatrix}
\]

where \( X \) is a \((2L \times p + 1)\) matrix. The Least Squares estimator is

\[
\hat{\beta}_{lm} = (X^t X)^{-1} X^t Y
\]

The information matrix \( X^t X \), as shown in (4.5), has as entries an expression of the form

\[
\sum_{k=1}^{2L} \bar{x}_k^j \quad \text{for} \quad j = 0, 1, \ldots, 2p
\]

and the \( X^t Y \) vector has elements of the form \( \sum_{k=1}^{2L} \bar{x}_k^j \bar{z}_k \) for \( j = 0, 1, \ldots, p \), for which after substituting with the original values of \( x_i \) and \( z_i \) we find

\[
\sum_{k=1}^{2L} \bar{x}_k^j = \begin{cases} 
2 \sum_{k=1}^{L} x_k^j & \text{if } j \text{ is even} \\
0 & \text{if } j \text{ is odd}
\end{cases}, \quad \sum_{k=1}^{2L} \bar{x}_k^j \bar{z}_k = \begin{cases} 
2 \sum_{k=1}^{L} x_k^j \bar{z}_k & \text{if } j \text{ is even} \\
0 & \text{if } j \text{ is odd}
\end{cases}
\]

which means the information matrix as well as its inverse are sparse matrices with the structure given by

\[
\begin{bmatrix}
\ast & 0 & \ast & 0 & \ast & 0 & \cdots \\
0 & \ast & 0 & \ast & 0 & \ast & \cdots \\
\ast & 0 & \ast & 0 & \ast & 0 & \cdots \\
0 & \ast & 0 & \ast & 0 & \ast & \cdots \\
\ast & 0 & \ast & 0 & \ast & 0 & \cdots \\
\vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots 
\end{bmatrix}
\]

57 A Statistical Model to Describe the Spread of Wafer Lenses.
and the vector $X'Y$ is also sparse with the following structure

\[
\begin{bmatrix}
* \\
0 \\
0 \\
\vdots
\end{bmatrix}
\]

which yields Least squares estimations of the form

\[
(\hat{\beta}_{lm})_j = \begin{cases} 
\hat{\beta}_j & \text{if } j \text{ is even} \\
0 & \text{if } j \text{ is odd} 
\end{cases}
\tag{4.21}
\]

**Right Mirroring**

In the same way as the Left Flipping, the data set is partitioned and arranged in a new data set in the following way

\[
\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_{2R-1}, \tilde{x}_R, \tilde{x}_{R+1}, \ldots, \tilde{x}_{2R-1}, \tilde{x}_{2R}
\]

\[
\tilde{z}_1, \tilde{z}_2, \ldots, \tilde{z}_{R-1}, \tilde{z}_R, \tilde{z}_{R+1}, \ldots, \tilde{z}_{2R-1}, \tilde{z}_{2R}
\]

The matrix notation has the same structure depicted for the Left Mirroring with a matrix $X$ of size $(2R \times p + 1)$. The Least Squares estimator is

\[
\hat{\beta}_{rm} = (X'X)^{-1} X'Y
\]

where the sub-index “$rm$” denoting **Right Mirroring**. In this setting the information matrix $X'X$, again as shown in (4.5), has as entries expressions of the form $\sum_{k=1}^{2R} \tilde{x}_k^j$ for $j = 0, 1, \ldots, 2p$ and the $X'Y$ vector has elements of the form $\sum_{k=1}^{2R} \tilde{x}_k^j \tilde{z}_k$ for $j = 0, 1, \ldots, p$, for which after substituting with the original values of $x_i$ and $z_i$ reduces to

\[
\sum_{k=1}^{2R} \tilde{x}_k^j = \begin{cases} 
2 \sum_{k=n-R+1}^{n} x_k^j & \text{if } j \text{ is even} \\
0 & \text{if } j \text{ is odd} 
\end{cases}
\]

\[
\sum_{k=1}^{2L} \tilde{x}_k^j \tilde{z}_k = \begin{cases} 
2 \sum_{k=n-R+1}^{n} x_k^j z_k & \text{if } j \text{ is even} \\
0 & \text{if } j \text{ is odd} 
\end{cases}
\]

where also the information matrix, its inverse and the $X^T Y$ vector have the sparse structures presented before. Then, the Least squares estimations are of the form

\[
(\hat{\beta}_R)_j = \begin{cases} 
\hat{\beta}_j & \text{if } j \text{ is even} \\
0 & \text{if } j \text{ is odd} 
\end{cases}
\tag{4.22}
\]

58 A Statistical Model to Describe the Spread of Wafer Lenses.
In concordance with Lemma 4.1 we encounter that the both left and right mirroring is equivalent to fit an even polynomial. That is, instead of considering a model as

\[ h_f(x) = \sum_{i=0}^{p} \beta_i x^i \]

we can consider with no cost a model of the form

\[ h^e(x) = \sum_{j=0}^{\lfloor p/2 \rfloor} \beta_{2j} x^{2j} \]

where the super-script “e” stands for even.

**Left Fitting**

The Left fitting corresponds to a linear regression by considering only the negative \( x \) points of the data set and then use them as positive \( x \) points. This process selects data of the form

\[
\bar{x} = \tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_{L-1}, \tilde{x}_L \\
= -x_L, -x_{L-1}, \ldots, -x_2, -x_1 \\
\bar{z} = \tilde{z}_1, \tilde{z}_2, \ldots, \tilde{z}_{L-1}, \tilde{z}_L \\
= z_L, z_{L-1}, \ldots, z_2, z_1
\]

The design matrix \( X \) is the size \( L \times p + 1 \). The Least Squares estimator is the usual \( \hat{\beta}_l = (X^tX)^{-1} X^tY \) with the corresponding \( Y \) vector where the sub-index “l” stands for **Left Fitting**. Different from the Left Mirroring, the estimations do not have an special form as in (4.21).

**Right Fitting**

The Left fitting corresponds to a linear regression by considering only the negative \( x \) points of the data set and then use them as positive \( x \) points. This process selects data of the form

\[
\bar{x} = \tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_{R-1}, \tilde{x}_R \\
= x_{n-R+1}, x_{n-R+2}, \ldots, x_{n-1}, x_n \\
\bar{z} = \tilde{z}_1, \tilde{z}_2, \ldots, \tilde{z}_{R-1}, \tilde{z}_R \\
= z_{n-R+1}, z_{n-R+2}, \ldots, z_{n-1}, z_n
\]

The design matrix \( X \) is the size \( R \times p + 1 \). The Least Squares estimator is the usual \( \hat{\beta}_r = (X^tX)^{-1} X^tY \) with the corresponding \( Y \) vector. Different from the Right flipping, the estimations do not have an special form as in (4.22).

59 A Statistical Model to Describe the Spread of Wafer Lenses.
4.5.1 Numerical Results

The analysis is presented based on the data for 14 measured lenses in 3 wafers. Results and analysis is done on the basis of the RMS values. The models used for the analysis are

\[ h_f(x) = \sum_{i=0}^{p} \beta_i x^i \] for all \( x \),

\[ h^l_f(x) = \sum_{i=1}^{p} \beta^l_i x^i \] for \( x \leq 0 \),

\[ h^r_f(x) = \sum_{i=1}^{p} \beta^r_i x^i \] for \( x \geq 0 \),

\[ h^{lm}_f(\tilde{x}) = \sum_{i=1}^{p} \beta^{lm}_i \tilde{x}^i \] for \( \tilde{x} \to \{x \leq 0, -x \leq 0\} \),

\[ h^{rm}_f(\tilde{x}) = \sum_{i=1}^{p} \beta^{rm}_i \tilde{x}^i \] for \( \tilde{x} \to \{-x \geq 0, x \geq 0\} \)

where the super-scripts “l”, “r”, “lm” and “rm” denotes “Left”, “Right”, “Left Mirrored” and “Right Mirrored” respectively.

Figure 4.10: RMS comparison. Flipping analysis

Figure 4.10 displays RMS values per lens over the wafers. The respective wafers are distinguished with the vertical dashed lines. As we can see, in the middle of Figure 4.10, performance of the polynomial regression is slightly worst than the other wafers. In overall, we can notice that both \( h^r_f(x) \) and \( h^{rm}_f(x) \) performs betters than the others. However,
when globally comparing the numerical results presented in Table 4.4 corresponding to Figure 4.10 we notice the improvement from one model to another is of the order or some tens of nanometers.

<table>
<thead>
<tr>
<th>wafer</th>
<th>Lens</th>
<th>Full Model</th>
<th>Left Fitting</th>
<th>Right Fitting</th>
<th>Left Mirror</th>
<th>Right Mirror</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>3.15489224e-05</td>
<td>3.54263457e-05</td>
<td>2.3317338e-05</td>
<td>3.7226608e-05</td>
<td>2.45175677e-05</td>
<td></td>
</tr>
<tr>
<td>L2</td>
<td>1.8368887e-05</td>
<td>1.7519648e-05</td>
<td>2.432911e-05</td>
<td>1.5079878e-05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L3</td>
<td>1.34120521e-05</td>
<td>1.38687507e-05</td>
<td>9.29225326e-05</td>
<td>9.68845319e-06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L4</td>
<td>1.61207055e-05</td>
<td>1.77513037e-05</td>
<td>1.13757435e-05</td>
<td>1.42135426e-05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L5</td>
<td>1.35513673e-05</td>
<td>1.91192498e-05</td>
<td>1.0805109e-05</td>
<td>1.2069725e-05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L6</td>
<td>2.74055264e-05</td>
<td>3.08698684e-05</td>
<td>2.1709661e-05</td>
<td>2.1849420e-05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L7</td>
<td>1.89032242e-05</td>
<td>1.72727272e-05</td>
<td>1.46421919e-05</td>
<td>1.65412028e-05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>W1</td>
<td>1.53715175e-05</td>
<td>1.39479935e-05</td>
<td>1.5094593e-05</td>
<td>1.56295367e-05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L9</td>
<td>1.40029862e-05</td>
<td>1.36929935e-05</td>
<td>1.0984734e-05</td>
<td>1.19313517e-05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L10</td>
<td>3.83210036e-05</td>
<td>4.56086303e-05</td>
<td>7.9951455e-05</td>
<td>9.91630599e-06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L11</td>
<td>1.42903269e-05</td>
<td>1.45010252e-05</td>
<td>9.63460414e-05</td>
<td>1.10694187e-05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L12</td>
<td>1.33499453e-05</td>
<td>1.24177360e-05</td>
<td>4.3650939e-05</td>
<td>1.40285679e-05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L13</td>
<td>1.40092862e-05</td>
<td>1.36929935e-05</td>
<td>1.0984734e-05</td>
<td>1.19313517e-05</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Additionally, when analyzing the output for the respective regression we can extract:

- Hypothesis tests on the significance of $\hat{\beta}_0$ for all the fitting procedures yields significant parameter.
- Left and Right fitting procedures yields coefficients for all the monomials.
- Both, Left and Right, mirroring produces a even polynomial, that is, only coefficients for the even monomials are computed. The coefficients for the odd mono-
mials have non-zero values as expected due to the numerical computation of the inverse involved in the corresponding Least Squares procedure.

4.6 Summary of the Chapter

We described fully how the polynomial estimation is performed under the Linear regression framework and some advantages of this setting is presented as well as the basis transformation procedure to find the required radial Zernike coefficients. We show the Linear Regression framework allows to perform additional tests on the significance of the estimations and quality of the model. We also compared the use of the transformation matrix procedure against a direct approach in finding the radial Zernike coefficients. In terms of RMS (or MSE) both procedures are similar yielding correlated coefficients. The shifting step to make the top of the lens to coincide with the origin of a cartesian plane is also investigated yielding that it is dangerous to shift the estimated model. Finally, by performing different Linear regression approaches we found there strong differences in the estimated polynomials. On the basis of what in this chapter was investigated, we found, under the linear regression analysis, always correlated coefficients and then we move forward to the analysis in finding a proper way to describe variability in terms of variances-covariances (correlations follows) in the following chapter.
5 Understanding Variation

As explained at the beginning of Section 2.2, there are many situations in which it is important to understand how two or more variables are related. The preceding chapter described different approaches of the curve fitting analysis for a lens profile description. While such an approaches provides useful curve fitting techniques, it is of more importance to describe the variability present in the lens profile description by taken as input several of those fittings, i.e. how to describe the variability of a lens when a curve fitting procedure is applied to several experimental lenses.

Describing a lens profile by a model such as (4.2) or (4.3), as currently done by Anteryon, tells that several variables have been “measured” for an experimental “unit”. In this context, the coefficients are estimated instead of measured and the experimental “units” are the lenses themselves. In the general case, the measured variables are often correlated and all of them are needed to be examined simultaneously in order to reveal patterns and key features in the data. By doing so, for instance, the variability present in the current lens profile description can be fully described and more reliable conclusions can be drawn.

As we already have seen, the current curve fitting procedure yields correlated coefficients and the basis matrix transformation do not attain uncorrelated coefficients. Hence, in this chapter, detailed analysis in the correlation from the linear regression analysis and from a given set of estimated coefficients is presented. The correlations are investigated in the sense of measuring how strong the given coefficients are related under the correlation formula (2.26). We first describe the variability from the linear regression analysis per lens in Section 5.1. We proceed in Section 5.2 with describing the variability present in a lens profile when several estimations are performed in several lenses and finally, in Section 5.3 we introduce a well-known multivariate statistical technique for describing to reveal hidden structure in a data and the better approach in describing the lens profile can be formulated.

For better understanding and ease in notation, denote the standard polynomial coefficients by $\beta_i$ following Formula (4.2) and the corresponding Zernike polynomials coefficients by $\alpha_i$ in accordance with Formula (4.3).
5.1 Variance-Covariance from Linear Regression

As mention at the end of Subsection 2.2.4, the estimates by the Least Squares from a model like (2.6) are, in general, correlated random variables. That is, the off-diagonal elements of the variance-covariance matrix (2.21b) often differ from 0. Formula (2.27) provides a computational formula to find the correlations of the estimated coefficients. Given such an expression, we aim to describe explicitly the correlations of the estimated coefficients after performing the basis matrix transformation. That is, we first find the estimations and correlations of the coefficients in standard polynomial, that is on the basis of $\beta$, and after performing the basis transformation we seek expression for the correlations of the radial Zernike coefficients, that is on the basis of $\alpha$.

In general we have a set of $p + 1$ estimated coefficients in the polynomial regression corresponding to a polynomial of degree $p$ as in (4.2). It is of interest to transform those coefficients to obtain the coefficients in the radial Zernike setting. The corresponding transformation can be set as follows:

$$
\hat{\beta}_0 = t_{00}\hat{\alpha}_0 + t_{10}\hat{\alpha}_1 + t_{20}\hat{\alpha}_2 + \ldots + t_{p0}\hat{\alpha}_p \\
\hat{\beta}_1 = t_{01}\hat{\alpha}_0 + t_{11}\hat{\alpha}_1 + t_{21}\hat{\alpha}_2 + \ldots + t_{p1}\hat{\alpha}_p \\
\hat{\beta}_2 = t_{02}\hat{\alpha}_0 + t_{12}\hat{\alpha}_1 + t_{22}\hat{\alpha}_2 + \ldots + t_{p2}\hat{\alpha}_p \\
\vdots = \vdots + \vdots + \vdots + \vdots \\
\hat{\beta}_p = t_{0p}\hat{\alpha}_0 + t_{1p}\hat{\alpha}_1 + t_{2p}\hat{\alpha}_2 + \ldots + t_{pp}\hat{\alpha}_p
$$

(5.1)

The constants $t_{ij}$ are given by any basis transformation matrix as described in earlier Chapters. We already showed that the estimated polynomial coefficients are correlated random variables and its correlations are given by (2.27) with $C = (X'X)^{-1}$ and $W$ given in (2.28). According with (2.26) we have $\rho_{\hat{\beta}_i, \hat{\beta}_j} \neq 0$ for all $i, j = 1, 2, \ldots, p$ which follows from $\text{Cov} [\hat{\beta}_i, \hat{\beta}_j] \neq 0$ according to (2.21b). We can easily find the “inverse” of the above set of equations to be of the form

$$
\hat{\alpha}_0 = u_{01}\hat{\beta}_0 + u_{10}\hat{\beta}_1 + u_{20}\hat{\beta}_2 + \ldots + u_{p0}\hat{\beta}_p \\
\hat{\alpha}_1 = u_{02}\hat{\beta}_0 + u_{12}\hat{\beta}_1 + u_{22}\hat{\beta}_2 + \ldots + u_{p2}\hat{\beta}_p \\
\hat{\alpha}_2 = u_{03}\hat{\beta}_0 + u_{13}\hat{\beta}_1 + u_{23}\hat{\beta}_2 + \ldots + u_{p3}\hat{\beta}_p \\
\vdots = \vdots + \vdots + \vdots + \vdots \\
\hat{\alpha}_p = u_{0p}\hat{\beta}_0 + u_{1p}\hat{\beta}_1 + u_{2p}\hat{\beta}_2 + \ldots + u_{pp}\hat{\beta}_p
$$

(5.2)

where the constants $u_{ij}$ are given by the inverse of the matrix transformation. Then we obtain that the desired Zernike polynomial coefficients are also random variables. We
can then investigate their variance-covariance. Thus
\[ V[\hat{\alpha}_k] = V\left[ \sum_{i=0}^{p} u_{ki} \hat{\beta}_i \right] \]
\[ = \sum_{i=0}^{p} u_{ki}^2 V[\hat{\beta}_i] + 2 \sum_{i < j} u_{ki} u_{kj} \text{Cov}[\hat{\beta}_i, \hat{\beta}_j] \]  
(5.3)
and the covariance reads as
\[ \text{Cov}[\hat{\alpha}_r, \hat{\alpha}_s] = \text{Cov}\left[ \sum_{i=0}^{p} u_{ri} \hat{\beta}_i, \sum_{j=0}^{p} u_{sj} \hat{\beta}_j \right] \]
\[ = \sum_{i=0}^{p} \sum_{j=0}^{p} u_{ki} u_{sj} \text{Cov}[\hat{\beta}_i, \hat{\beta}_j] \]  
(5.4)
With the above expressions, (5.3) and (5.4), we can easily find the corresponding correlations by using (2.26) in Definition 2.13. Thus we arrive at
\[ \rho_{\hat{\alpha}_r, \hat{\alpha}_s} = \frac{\text{Cov}[\hat{\alpha}_r, \hat{\alpha}_s]}{\sqrt{V[\hat{\alpha}_r]} \sqrt{V[\hat{\alpha}_s]}} \]  
(5.5)

**Comments**

Notice that these expressions yield correlations on the estimated coefficients in the radial Zernike setting per linear regression. That is, all variances and covariances are computed from the linear regression analysis per lens and such computations are done on the basis of measured data. Under this setting, the computed values for the correlation gives information in the within-lens variability present in the estimated coefficients and no conclusion in the lens-to-lens variability can be drawn.

In correspondence with earlier described matrix approaches, we recall the matrix formulas earlier described from the Least Squares estimation and the basis matrix transformation. Then
\[ \hat{\beta} = (X'X)^{-1} X'Y \]  
(5.6)
\[ V[\hat{\beta}] = \hat{\sigma}^2 C; \quad C = (X'X)^{-1} \]  
(5.7)
\[ \hat{\alpha} = (T'^{-1}) \hat{\beta} \]
\[ = (T'^{-1})(X'X)^{-1} X'Y \]  
(5.8)
\[ V[\hat{\alpha}] = (T'^{-1}) V[\hat{\beta}] (T')^{-1} \]
\[ = \hat{\sigma}^2 \tilde{C}; \quad \tilde{C} = (TX'XT')^{-1} \]  
(5.9)
with $X$ the design matrix in the polynomial regression, $Y$ the vector with observed values, $T$ the corresponding basis transformation matrix and $\sigma^2$ the variance of the model errors. With this formulas we have that the correlation can be also computed in matrix notation as

$$\rho_{\hat{\beta}} = W^{-1}CW^{-1}$$  \hspace{1cm} (5.10)

$$\rho_{\hat{\alpha}} = \tilde{W}^{-1}\tilde{C}\tilde{W}^{-1}$$  \hspace{1cm} (5.11)

where $W$ and $\tilde{W}$ are given by

$$W = \begin{bmatrix} \sqrt{C_{00}} & 0 & 0 & \cdots & 0 \\ 0 & \sqrt{C_{11}} & 0 & \cdots & 0 \\ 0 & 0 & \sqrt{C_{22}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sqrt{C_{pp}} \end{bmatrix}$$

$$\tilde{W} = \begin{bmatrix} \sqrt{\tilde{C}_{00}} & 0 & 0 & \cdots & 0 \\ 0 & \sqrt{\tilde{C}_{11}} & 0 & \cdots & 0 \\ 0 & 0 & \sqrt{\tilde{C}_{22}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sqrt{\tilde{C}_{pp}} \end{bmatrix}$$

Given these set of matrix formulas, we have that the above described set of simultaneous equations (5.1) and (5.2) can be easily expressed as

$$\hat{\beta} = T^t\hat{\alpha}$$  \hspace{1cm} respectively  \hspace{1cm} $$\hat{\alpha} = (T^t)^{-1}\hat{\beta}$$

where, the elements $t_{ij}$ are gathered together as $T$, the corresponding basis transformation matrix, and the elements $u_{ij}$ are the entries of $T^{-1}$ given that such inverse exists.

The variance-covariance matrices $V[\hat{\beta}]$ and $V[\hat{\alpha}]$ are symmetric matrices which the corresponding variances are placed in the diagonal entries and the covariances are placed in the off-diagonal entries. Then, following this matrix notation, we can have rewrite compactly (5.3) and (5.4) together as (5.10).

**Numerical Results**

Implementation of the above explained procedure is done for a particular Lens. Following notation introduced in Section 4.1, we have partially shown the Least Squares estimation in Table 4.1. The polynomial fitted is in the form given by (4.2), with $p = 16$. It follows that (5.7) is a squared matrix of size $17$. The variance-covariance (5.7) of example presented in Table 4.1 looks like given in Table 5.1 and the corresponding correlation matrix (5.10) is of the form given in Table 5.2 where for instance we notice that $\hat{\beta}_0$ shows high correlation with $\hat{\beta}_2$ and $\hat{\beta}_4$, (Correlation above 0.6), and that $\hat{\beta}_1$ shows strong correlation with $\hat{\beta}_3$ and $\hat{\beta}_5$, (Correlation above 0.8). We also can notice, that both Table 5.1 and Table 5.2 are symmetric matrices as expected.

With the basis transformation matrix, as explained in Subsection 4.1.2 we have that the variance-covariance matrix (5.9) and the corresponding correlation matrix (5.11) are computed. Results are omitted.
5.2 Variance-Covariance from a set of Estimated Coefficients

Section 5.1 shows how to compute the correlation present in the radial Zernike coefficients when a curve fitting procedure and a basis matrix transformation have taken place. Although this analysis is also used for finding an expected lens profile and its variability for a particular lens it does not account the influence of other lenses in a lens profile description. When considering the later, the lens profile can be taken as the result of several random variables (coefficients) measured in several lenses. The expected lens profile and its variability can be then investigated under multivariate analysis.

From the radial Zernike expression to represent a lens shape (c.f Formula (4.3)), a set of estimated coefficients \( \hat{\alpha} \) (according with our notation) is given for a finite number \( L \) of lenses. From the linear regression framework we have found the estimated coefficients are themselves random variables.

Without loss of generality, the given coefficients can be compactly gathered as in Table 5.3 where \( \hat{\alpha}_{i,j} \) represents the estimated \( \alpha_i \) coefficient of the lens \( j \), that is

\[
\hat{h}_{i}^{j}(x) = \sum_{i=0}^{p} \hat{\alpha}_{ij} R_{i}(x) \quad \text{for } j = 1, 2, \ldots, L
\]
From the linear regression analysis we have found that $h_z(x)$ is itself a random variable, then, following properties of expectation and variance we have

$$E[h_z(x)] = E \left[ \sum_{i=0}^{p} \alpha_i R_i(x) \right]$$

$$= \sum_{i=0}^{p} R_i(x) E[\alpha_i] \quad (5.12)$$

$$V[h_z(x)] = V \left[ \sum_{i=0}^{p} \alpha_i R_i(x) \right]$$

$$= \sum_{i=0}^{p} R_i^2(x) V[\alpha_i] + 2 \sum_{i \leq j} R_i(x) R_j(x) \text{Cov}[\alpha_i, \alpha_j] \quad (5.13)$$

from where it was shown already that $\text{Cov}[\alpha_i, \alpha_j] \neq 0$. The corresponding theoretical expectation, variances and covariances can be computed with the sample variances and covariances as

$$E[\alpha_i] = \mu_{\alpha_i} = \frac{1}{L} \sum_{k=1}^{L} \hat{\alpha}_{ik}$$

$$V[\alpha_i] = S_{\alpha_i, \alpha_i} = \frac{1}{L - 1} \sum_{k=1}^{L} (\hat{\alpha}_{ik} - E[\alpha_i])^2$$

$$\text{Cov}[\alpha_i, \alpha_j] = S_{\alpha_i, \alpha_j} = \frac{1}{L - 1} \sum_{k=1}^{L} (\hat{\alpha}_{ik} - E[\alpha_i]) (\hat{\alpha}_{jk} - E[\alpha_j])$$

where $\alpha_i, \alpha_j$ are random variables and $\hat{\alpha}_{ik}, \hat{\alpha}_{jk}$ for $k = 1, 2, \ldots, L$ are observations of $\alpha_i, \alpha_j$ respectively as given in Table 5.3. Notice in this section, rather than focusing on variances and covariances of the estimated coefficients for a particular lens, we describe on the basis of a finite set of estimated coefficients an expected lens profile and its variation as a function of the radial position $x$ considering the coefficients are correlated random variables.

<table>
<thead>
<tr>
<th>Lens</th>
<th>$\alpha_0$</th>
<th>$\alpha_1$</th>
<th>$\alpha_2$</th>
<th>$\cdots$</th>
<th>$\alpha_p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\hat{\alpha}_{01}$</td>
<td>$\hat{\alpha}_{11}$</td>
<td>$\hat{\alpha}_{21}$</td>
<td>$\cdots$</td>
<td>$\hat{\alpha}_{p1}$</td>
</tr>
<tr>
<td>2</td>
<td>$\hat{\alpha}_{02}$</td>
<td>$\hat{\alpha}_{12}$</td>
<td>$\hat{\alpha}_{22}$</td>
<td>$\cdots$</td>
<td>$\hat{\alpha}_{p2}$</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\ddots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>L</td>
<td>$\hat{\alpha}_{0L}$</td>
<td>$\hat{\alpha}_{1L}$</td>
<td>$\hat{\alpha}_{2L}$</td>
<td>$\cdots$</td>
<td>$\hat{\alpha}_{pL}$</td>
</tr>
</tbody>
</table>

Table 5.3: Given estimated coefficients per lens
5.3 Principal Component Analysis

The preceding sections provide expressions to describe, in the first place, the variance and covariance of the coefficients from the linear regression analysis, and secondly, an expected lens profile and its variability as a function of the coordinate denoting the radial position. With given such an analysis, we are not able to reveal possible structure in the lens shape description which allows a better understanding on how a lens profile behaves as requested by Anteryon. Hence, we introduce the Principal Components Analysis (PCA) which allows to understand better the variability present a data set concerning a lens profile description.

The Theory of PCA

Principal Component Analysis (PCA) is a multivariate technique in statistics which allows to transform a number of correlated variables into (hopefully smaller) set of uncorrelated variables. The basic goal of PCA is to describe the variation in a set of correlated variables $X_1, X_2, X_3, \ldots, X_q$ in terms of another set of variables $Y_1, Y_2, Y_3, \ldots, Y_q$ which are linear combination of the original variables and for which this new set is uncorrelated. Jackson and Wiley (1991) provides a review in history and applications of PCA. For more details, the reader is refer to Johnson and Wichern (1992) and Mardia et al. (1980). A proper development of the topic in the R software is presented in Everitt (2005) while Shlens (2005) provides an intuitive interpretation of the ultimate goal of PCA and develops the mathematical framework in the context of Linear Algebra.

Before deriving the algebraic basis of PCA according to Everitt (2005) we require the following notation.

- Let $X_1, X_2, X_3, \ldots, X_q$ be the set of correlated variables.
- Let $Y_1, Y_2, Y_3, \ldots, Y_q$ be the set of uncorrelated variables expressed as linear combination of the variables $X_i$. It is usually called $Y_k$ as the $k^{th}$ principal component.
- Let $S$ be the sample variance-covariance matrix of the variables $X_i$.
- Let $\lambda_i$ and $\nu_i$ be the $i^{th}$ eigenvalue respectively eigenvector of $S$.

We will assume $\lambda_1 > \lambda_2 > \cdots > \lambda_q$. The methodology used by PCA can be described as follows

- The first principal component, $Y_1$, is of the form

$$Y_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1q}X_q$$
such that $V[Y_1]$ is greatest among all such combinations. We must constrain the coefficients $a_{1j}$ otherwise the variance of $Y_1$ can increase with no limit. It is usually taken as constraint the sum of squares of the coefficients to be 1. That is, by letting $a_1$ be a vector with elements $a_{1j}$, we have that $a_1^T a_1 = 1$.

- The second principal component, $Y_2$, is of the form
  $$Y_2 = a_{21} X_1 + a_{22} X_2 + \ldots + a_{2q} X_q$$
  such that $V[Y_2]$ is greatest among all possible combinations under the constraints on the coefficients given by
  $$a_2^T a_2 = 1$$
  $$a_2^T a_1 = 0$$
  where $a_2$ is a vector with elements $a_{2j}$ and on which the last condition yield uncorrelated $Y_1$ and $Y_2$.

- In general we have the $k^{th}$ principal component to be
  $$Y_k = a_{k1} X_1 + a_{k2} X_2 + \ldots + a_{kq} X_q$$
  which has greatest variance subject to
  $$a_k^T a_k = 1$$
  $$a_k^T a_i = 0 \quad (i \leq k)$$

Under any optimization technique, solution to the previous optimization problem is found on the basis of the eigenvalues and eigenvector of the sample variance-covariance matrix related to the correlated variables $X_i$ (for details, Everitt (2005) refers to Morrison (1967)). Results can be gathered as follows:

- The coefficients of the first component (this is $a_{1j}$) are the elements of the eigenvector of the sample variance-covariance matrix corresponding to the largest eigenvalue ($\lambda_1$). That is $a_1 = \nu_1$

- The coefficients of the second component (this is $a_{2j}$) are the elements of the eigenvector of the sample variance-covariance matrix corresponding to the second largest eigenvalue ($\lambda_2$). That is $a_2 = \nu_2$

- The coefficients of the $k^{th}$ component (this is $a_{kj}$) are the elements of the eigenvector of the sample variance-covariance matrix corresponding to the $k^{th}$ greatest eigenvalue ($\lambda_k$). That is $a_k = \nu_k$
From the construction of the principal components, we can find

\[ V[Y_j] = \lambda_j \]

and the total variance of the principal components equals the total variance of the original variables, so

\[ \sum_{i=1}^{q} V[X_i] = \sum_{i=1}^{q} V[Y_i] \]

by taking \( S_i = V[X_i] \) as the sample variance we have

\[ \sum_{i=1}^{q} \lambda_i = \sum_{i=1}^{q} S_i = \text{trace}(S) \]

Additionally, under the assumption, \( \lambda_1 > \lambda_2 > \cdots > \lambda_q \), we have that \( V[Y_1] > V[Y_2] > \cdots > V[Y_q] \).

Previous relations are found on the basis of the sample variance-covariance matrix \( S \). However, in practice is more meaningful to use instead the sample correlation matrix, denoted usually as \( R \). There is no simple relation, as remarked by Everitt (2005), between the results by using the sample variance-covariance matrix \( S \) and the results by using the sample correlation matrix \( R \). For a proper interpretation of the principal components we have

- Covariance between \( X_i \) and \( Y_j \) is given by
  \[
  \text{Cov}[X_i, Y_j] = \lambda_j a_{ji}
  \]

- Correlation between \( X_i \) and \( Y_j \) is given by
  \[
  \text{Corr}[X_i, Y_j] = \frac{\text{Cov}[X_i, Y_j]}{\sqrt{V[X_i]}\sqrt{V[Y_j]}} = \frac{a_{ji} \sqrt{\lambda_j}}{S_i}
  \]

**The Linear Algebra approach for PCA**

The linear algebra approach proposed by Shlens (2005) provides useful matrix formulas easy to implement in any software in order to perform PCA. Let \( X \) be a \( q \times n \) matrix whose row \( i^{th} \) contains \( n \) observations of variable \( X_i \) and hence the column \( j^{th} \) of \( X \) contains a single observation of all \( q \) variables. Let \( Y \) be another \( q \times n \) associated with the linear transformation \( P \) of size \( q \times q \). That is

\[ PX = Y. \]
Matrix $P$, as we have described in earlier chapters, represents a basis matrix transformation from the system given by $X$ to another system given by $Y$. The variance-covariance matrix $S$ of $X$ is computed as

$$S_X = \frac{1}{n-1}XX^t$$

on which each variable $X_i$ is taken to have zero mean. $S_X$ is a symmetric $q \times q$ matrix on which the variances of $X_i$ are placed in the diagonal and the covariances are placed in the off-diagonal. The corresponding eigenvalues and eigenvector are computed such that $S_X v_i = \lambda_i v_i$. Then, as described earlier, the columns of the matrix $P$ are the eigenvectors of $S_X$ corresponding to the respective eigenvalues, that is

$$P = \begin{bmatrix} v_1 & v_2 & \ldots & v_q \\ \vdots & \vdots & \ddots & \vdots \end{bmatrix}$$

The new matrix $PX$ is then the principal components and the variances of the principal components can be computed from the variance-covariance matrix of $PX$. That is

$$S_Y = \frac{1}{n-1}PX X^t P^t.$$

The procedure to perform PCA can be summarized as follows:

1. Set the matrix $X$ in proper form. Each row should be a list of observations of one variable. Each column is then a single observation of all variables. Subtract the mean of each row to the matrix $X$. Set the new “scaled” matrix $X$ on which each $X_i$ variable has zero mean.

2. Compute the variance-covariance matrix $S_X = XX^t/(n-1)$

3. Find the eigenvectors of the matrix $S$. Arrange them in a matrix $P$ such that the first columns corresponds to the greatest eigenvalues and so on.

4. Compute $Y = PX$. This are the principal components.

5. Read off the variances of the principal components from the variance-covariance matrix $S_Y = PXX^tP^t/(n-1)$

**Numerical Results**

We refer the example to a data set of estimated coefficients per lens arranged similar to Table 5.3. We set our correlated variables of interest $X_i$ as the estimated coefficients $\alpha_i$.

72 A Statistical Model to Describe the Spread of Wafer Lenses.
According to the array presented in Table 5.3, each row is seen as a one sample or trial of all coefficients α’s per experimental unit (each lens), while each column represents a list of observations for a single coefficient α_j across all lenses. Then, for the PCA procedure, the matrix X is set as the transpose of Table 5.3. For the analysis, we chose the number of variables p = 16 in accordance with the polynomial fitted given by (4.3).

We first start by an explorative analysis of our data. Despite the analysis is done for all 16 variables α’s, Figure 5.1 presents a scatter plot of variables α_0 up to α_9. We can notice, for instance, that α_0 and α_2 seem to have fairly strong correlation. In the same way, α_1 and α_3 presents high correlation with outliers involved. For mostly all the remaining combination of coefficients, there are, apparently, groups which should be taken into account. The graphical analysis is not conclusive. In general, a clear message from Picture 5.1 is that some lenses should be considered as outliers. The corresponding correlation matrix of the data is displayed in Table 5.3. We can then confirm that coefficients α_0 and α_2 are strongly correlated by corr[α_0, α_2] = 0.966. Additionally we notice that α_3 and α_5 as well as α_3 and α_9 are also strong correlated variables.

Figure 5.1: Scatter plot of α_0 up to α_9

73 A Statistical Model to Describe the Spread of Wafer Lenses.
Table 5.4: Correlation matrix for first 10 estimated coefficients

<table>
<thead>
<tr>
<th></th>
<th>α₀</th>
<th>α₁</th>
<th>α₂</th>
<th>α₃</th>
<th>α₄</th>
<th>α₅</th>
<th>α₆</th>
<th>α₇</th>
<th>α₈</th>
<th>α₉</th>
<th>α₁₀</th>
<th>α₁₁</th>
<th>α₁₂</th>
<th>α₁₃</th>
<th>α₁₄</th>
<th>α₁₅</th>
<th>α₁₆</th>
</tr>
</thead>
<tbody>
<tr>
<td>α₀</td>
<td>1.000</td>
<td>0.178</td>
<td>0.966</td>
<td>0.539</td>
<td>0.506</td>
<td>-0.271</td>
<td>0.520</td>
<td>0.011</td>
<td>0.352</td>
<td>-0.343</td>
<td>-0.293</td>
<td>-0.022</td>
<td>0.377</td>
<td>-0.379</td>
<td>-0.261</td>
<td>0.368</td>
<td>0.304</td>
</tr>
<tr>
<td>α₁</td>
<td>0.178</td>
<td>1.000</td>
<td>0.173</td>
<td>0.048</td>
<td>0.008</td>
<td>-0.043</td>
<td>0.064</td>
<td>-0.084</td>
<td>0.149</td>
<td>-0.019</td>
<td>-0.159</td>
<td>-0.133</td>
<td>0.028</td>
<td>0.075</td>
<td>0.002</td>
<td>0.154</td>
<td>0.097</td>
</tr>
<tr>
<td>α₂</td>
<td>0.966</td>
<td>0.173</td>
<td>1.000</td>
<td>0.498</td>
<td>0.623</td>
<td>-0.200</td>
<td>0.438</td>
<td>-0.027</td>
<td>0.476</td>
<td>-0.339</td>
<td>-0.409</td>
<td>-0.040</td>
<td>0.443</td>
<td>-0.421</td>
<td>-0.338</td>
<td>0.373</td>
<td>0.419</td>
</tr>
<tr>
<td>α₃</td>
<td>0.539</td>
<td>0.048</td>
<td>0.498</td>
<td>1.000</td>
<td>0.114</td>
<td>-0.798</td>
<td>0.460</td>
<td>0.459</td>
<td>0.108</td>
<td>-0.740</td>
<td>-0.252</td>
<td>0.119</td>
<td>0.340</td>
<td>-0.132</td>
<td>-0.163</td>
<td>0.299</td>
<td>0.353</td>
</tr>
<tr>
<td>α₄</td>
<td>0.506</td>
<td>0.008</td>
<td>0.623</td>
<td>0.114</td>
<td>1.000</td>
<td>0.212</td>
<td>0.500</td>
<td>-0.295</td>
<td>0.295</td>
<td>-0.048</td>
<td>-0.275</td>
<td>0.067</td>
<td>0.276</td>
<td>-0.259</td>
<td>-0.257</td>
<td>0.271</td>
<td>0.252</td>
</tr>
<tr>
<td>α₅</td>
<td>-0.271</td>
<td>-0.043</td>
<td>-0.200</td>
<td>-0.798</td>
<td>0.212</td>
<td>1.000</td>
<td>-0.241</td>
<td>-0.614</td>
<td>-0.056</td>
<td>0.644</td>
<td>0.103</td>
<td>-0.002</td>
<td>-0.065</td>
<td>0.028</td>
<td>0.082</td>
<td>-0.146</td>
<td>-0.235</td>
</tr>
<tr>
<td>α₆</td>
<td>0.520</td>
<td>0.094</td>
<td>0.458</td>
<td>0.460</td>
<td>0.500</td>
<td>-0.241</td>
<td>1.000</td>
<td>0.003</td>
<td>-0.094</td>
<td>-0.329</td>
<td>-0.010</td>
<td>0.240</td>
<td>0.238</td>
<td>-0.163</td>
<td>-0.211</td>
<td>0.238</td>
<td>0.155</td>
</tr>
<tr>
<td>α₇</td>
<td>0.011</td>
<td>-0.084</td>
<td>-0.027</td>
<td>0.458</td>
<td>-0.295</td>
<td>-0.614</td>
<td>0.003</td>
<td>1.000</td>
<td>0.100</td>
<td>-0.488</td>
<td>-0.053</td>
<td>0.053</td>
<td>0.089</td>
<td>-0.165</td>
<td>0.070</td>
<td>0.117</td>
<td>0.188</td>
</tr>
<tr>
<td>α₈</td>
<td>0.352</td>
<td>0.149</td>
<td>0.476</td>
<td>0.108</td>
<td>0.295</td>
<td>-0.056</td>
<td>-0.094</td>
<td>0.100</td>
<td>1.000</td>
<td>-0.208</td>
<td>-0.130</td>
<td>-0.181</td>
<td>0.017</td>
<td>-0.359</td>
<td>-0.449</td>
<td>0.025</td>
<td>0.243</td>
</tr>
<tr>
<td>α₉</td>
<td>-0.343</td>
<td>-0.019</td>
<td>-0.339</td>
<td>-0.740</td>
<td>-0.048</td>
<td>0.644</td>
<td>-0.329</td>
<td>-0.488</td>
<td>-0.208</td>
<td>1.000</td>
<td>0.122</td>
<td>-0.065</td>
<td>-0.238</td>
<td>0.130</td>
<td>0.233</td>
<td>-0.178</td>
<td>-0.429</td>
</tr>
<tr>
<td>α₁₀</td>
<td>-0.293</td>
<td>-0.159</td>
<td>-0.409</td>
<td>-0.252</td>
<td>-0.275</td>
<td>0.103</td>
<td>-0.010</td>
<td>-0.053</td>
<td>-0.130</td>
<td>0.122</td>
<td>1.000</td>
<td>-0.249</td>
<td>-0.306</td>
<td>0.267</td>
<td>-0.118</td>
<td>-0.112</td>
<td>-0.374</td>
</tr>
<tr>
<td>α₁₁</td>
<td>-0.022</td>
<td>-0.133</td>
<td>-0.040</td>
<td>0.119</td>
<td>0.067</td>
<td>-0.002</td>
<td>0.240</td>
<td>0.053</td>
<td>-0.181</td>
<td>-0.065</td>
<td>-0.249</td>
<td>1.000</td>
<td>0.034</td>
<td>0.027</td>
<td>0.173</td>
<td>-0.056</td>
<td>-0.050</td>
</tr>
<tr>
<td>α₁₂</td>
<td>0.377</td>
<td>0.028</td>
<td>0.443</td>
<td>0.340</td>
<td>0.276</td>
<td>-0.065</td>
<td>0.238</td>
<td>-0.089</td>
<td>0.017</td>
<td>-0.238</td>
<td>-0.306</td>
<td>0.034</td>
<td>1.000</td>
<td>-0.194</td>
<td>-0.098</td>
<td>0.373</td>
<td>0.242</td>
</tr>
<tr>
<td>α₁₃</td>
<td>-0.379</td>
<td>0.075</td>
<td>-0.421</td>
<td>-0.132</td>
<td>-0.259</td>
<td>0.028</td>
<td>-0.163</td>
<td>-0.165</td>
<td>-0.359</td>
<td>0.130</td>
<td>0.267</td>
<td>0.027</td>
<td>-0.194</td>
<td>1.000</td>
<td>0.186</td>
<td>-0.056</td>
<td>-0.177</td>
</tr>
<tr>
<td>α₁₄</td>
<td>-0.261</td>
<td>0.002</td>
<td>-0.338</td>
<td>-0.163</td>
<td>-0.257</td>
<td>0.082</td>
<td>-0.211</td>
<td>0.070</td>
<td>-0.449</td>
<td>0.233</td>
<td>-0.118</td>
<td>0.173</td>
<td>-0.098</td>
<td>0.186</td>
<td>1.000</td>
<td>-0.096</td>
<td>-0.192</td>
</tr>
<tr>
<td>α₁₅</td>
<td>0.368</td>
<td>0.154</td>
<td>0.373</td>
<td>0.299</td>
<td>0.271</td>
<td>-0.146</td>
<td>0.238</td>
<td>0.117</td>
<td>0.025</td>
<td>-0.178</td>
<td>-0.112</td>
<td>-0.056</td>
<td>0.373</td>
<td>-0.056</td>
<td>-0.096</td>
<td>1.000</td>
<td>0.126</td>
</tr>
<tr>
<td>α₁₆</td>
<td>0.304</td>
<td>0.097</td>
<td>0.419</td>
<td>0.353</td>
<td>0.252</td>
<td>-0.235</td>
<td>0.155</td>
<td>0.188</td>
<td>0.243</td>
<td>-0.429</td>
<td>-0.374</td>
<td>-0.050</td>
<td>0.242</td>
<td>-0.177</td>
<td>-0.192</td>
<td>0.126</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Figure 5.2: Scatter Plot of the first 5 estimated coefficients

74 A Statistical Model to Describe the Spread of Wafer Lenses.
For analysis the behavior of the individual estimated coefficients, a time series plot for some of the coefficients is presented in Figure 5.2. The vertical dashed lines separates estimated coefficients for different wafers. There is a clear variation from wafer-to-wafer for all the coefficients. We see there is a clear difference of the estimations of each coefficient. Almost half of the data presents a different behavior when compared with the other half. Intuitively, this fact gives an idea of the grouping characteristic present in the data.

Now, taken into account the correlation of coefficients as described before, we aim to look for a better description of the data in terms of uncorrelated variables. In other words, the lens shape variation which is expressed in terms of (correlated) coefficients is sought to be expressed in terms of new uncorrelated variables which are in fact a combination of the original coefficients.

As pointed out before, it seems to be outliers in the data, however, an explorative analysis based in all data is suggested to be done. The principal components of the data is presented is Table 5.5. In here, we can see, that the very first principal component already accounts for almost 97% of the variance of the original coefficients. This also can be seen in Figure 5.3 where the variances of the principal components is depicted. The top panel shows a bar plot with the variances of the principal components while the bottom panel depicts the scaled standard deviation with the total standard deviations of all principal components. From this, we can select, for instance, the scores of the first 5 principal components to summarize the information in all the estimated coefficients without none significant loss of information. That is, we can reduce the dimensionality in describing the lens shape to only the very first principal components. In the loadings from Table 5.5 all the coefficients for each principal component are scaled in such a way that the sum of their squares add up to 1 and the “−” symbols (blanks) indicate near-zero values.

<table>
<thead>
<tr>
<th>Importance of components</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comp.1</td>
</tr>
<tr>
<td>Standard deviation</td>
</tr>
<tr>
<td>Proportion of Variance</td>
</tr>
<tr>
<td>Cumulative Proportion</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Loadings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comp.1</td>
</tr>
<tr>
<td>Coefficient α₀</td>
</tr>
<tr>
<td>Coefficient α₁</td>
</tr>
<tr>
<td>Coefficient α₂</td>
</tr>
<tr>
<td>Coefficient α₃</td>
</tr>
<tr>
<td>Coefficient α₄</td>
</tr>
<tr>
<td>Coefficient α₅</td>
</tr>
<tr>
<td>Coefficient α₆</td>
</tr>
<tr>
<td>Coefficient α₇</td>
</tr>
<tr>
<td>Coefficient α₈</td>
</tr>
<tr>
<td>Coefficient α₉</td>
</tr>
</tbody>
</table>
| ... | ... | ... | ... | ... | ...

Table 5.5: Principal Components of the data

75 A Statistical Model to Describe the Spread of Wafer Lenses.
Figure 5.3: Variance of the Principal Components

We can, for instance, use the first five principal components, which account for almost 99.8% of the total variance of the Zernike coefficients, to summarize the lens profile description with little loss of the information in the entire lens. Since we have concluded that the Zernike coefficients are correlated the post-analysis of the lens performance should not be done in terms of individual analysis of the Zernike coefficients but in terms of the principal components.

Figure 5.4: Scatter plot of Comp.1 VS Comp.2

A Statistical Model to Describe the Spread of Wafer Lenses.
Scatter plots like Figure 5.4 are useful to determine additional behavior of the data set. Analyzing these we see clearly, one more time, the presence of groups as well as the presence of outliers. We finish the numerical analysis by pointing out the interest in removing the outliers. As mentioned by Everitt (2005), the simplest approach is removing the outliers from the data set and perform all the analysis again. Instead, the minimum volume ellipsoid approach can be used which it is found an estimator of the covariance matrix with minimum determinant. With this new covariance matrix, the principal component analysis can be performed. Additionally we could try to overcome what was originally observed about the grouping. An intuitive splitting of the data on the basis of Figure 5.2 can be done. Not presented in here, both, the minimum volume ellipsoid approach, and the splitting was performed showing similar results. No significant improvement was found.
6 Conclusions and Recommendations

6.1 Summary of the Work

In this report different approaches related to the properties of curve fitting are presented and used to describe the variability in the lens shape and to gain insight in the quality of fitting mechanism. Additionally, a multivariate technique is used to reveal hidden structure in the representation of a lens profile. The principles for basis transformation and linear regression analysis are described in Chapter 2. These principles covered the representation of a lens profile in different basis and how to switch from a basis to another. Least Squares Estimators and their properties, as well as implications of orthogonal functions in linear regression analysis were discussed. These principles were extensively used in Chapter 4 where different approaches to represent the profile in the cross-section of a lens are investigated. Such approaches covered the standard linear regression for bivariate data when fitting a polynomial in the cross-section of a lens along with its corresponding transformation into the well-known radial Zernike polynomials. Main results of the curve fitting were used in Chapter 5 to treat, in more detail, the description of variability present in the current lens profile description in its cross-section. Variability was described in terms of variance-covariance and/or correlation of the estimated coefficients from the linear regression analysis and from a set of estimated coefficients for different lenses. A Principal Components Analysis is elaborated by considering the lens profile description as a multivariate data analysis.

From the analysis performed in Chapter 4 we conclude that either in the standard monomials or the radial Zernike polynomials, the description of a lens profile under the linear regression analysis yields correlated coefficients. Linear regression can be directly implemented in radial Zernike polynomials avoiding the basis matrix transformation since results are similar. A regression through the origin is preferred over the shifting process since the improvement in MSE when using a model with intercept is of the order of $10^{-4}$ nm and the changes in the estimated coefficients are quite small. The proposed linear regression approaches showed that the behavior is similar in terms of RMS values. The mirroring procedures yield a even polynomial which performs better in terms of RMS when compared with the partitioned data. The models for the partitioned data yields an odd polynomial. It was shown that the polynomial model for the mirroring procedure can be substituted by a polynomial model when only even powers of $x$ are

78 A Statistical Model to Describe the Spread of Wafer Lenses.
considered.
In Chapter 5, correlations among coefficients are obtained from the variance-covariance matrix per lens. However, this yields to correlations among the estimated coefficients within a lens. When considering a set of estimated coefficients from several lenses, the expected lens profile and its variation can be computed by using the sample means and the sample variances-covariances of the corresponding coefficients. The implementation of the PCA showed that in fact the coefficients in the current description of the lens profile are correlated. It was shown that most of the variability of the lens profile can be attributed to the coefficients corresponding to those radial Zernike polynomials of lower order. Given so, under the results of the PCA, the description of a lens profile can be made by using a lower number of uncorrelated variables which are linear correlation of the given coefficients. However, the uncorrelated variables may not have a physical interpretation.

We summarize the findings as follows:

- The Least Squares estimation when fitting a polynomial or radial Zernike polynomials yields correlated variables. This follows from the non-orthogonality in the standard basis for the polynomial space $P(x)$. It happens that the selection $m = 0$ for $n$ even and $m = 1$ for $n$ odd of the radial Zernike polynomials to describe the cross-section of a lens is also a non-orthogonal basis for the polynomial space $P(x)$.

- The basis matrix transformation to find the radial Zernike polynomial can be avoided and the Least Squares can be applied directly in terms of radial Zernike polynomials.

- Since the standard monomials and the radial Zernike polynomials are both bases for the polynomial space $P(x)$, the respective estimated models when considering either one or another bases have the same RMS. This follows since the estimated model that corresponds to the projection of the measurements in the polynomial space $P(x)$ is unique.

- The current order selection rule applied by Anteryon seems to be adequate. According to the analysis of the model order selection, the RMS does not reach significant improvement after a 12th degree polynomial indicating that a lower order polynomial could be used for the linear regression analysis. This is however not feasible since the order of the polynomial design is higher.

- The shifting step in the current fitting procedure implemented by Anteryon can be improved. The analysis revealed that a polynomial model through the origin is desirable for the curve fitting procedure. This follows from comparison between the respective RMS values and from comparison in the estimated coefficients.
When using a polynomial model to fit mirrored data, even polynomials are estimated. When partitioning the data odd polynomials are obtained. In terms of RMS, the obtained even polynomials perform better.

Describing the lens profile should take into account correlations of the coefficients since by the current estimation procedure in the cross-section of lens yields correlated variables. When a set of estimated coefficients from several lenses is given, the expected lens profile and its variation can be computed from the sample coefficients means and the sample coefficients variances.

Most of the variability present in the current description of a lens profile comes from the coefficients of the radial Zernike polynomials of lower order. This implies that the description of a lens can be given in terms of a lower dimensional set of uncorrelated variables. The new set of uncorrelated variables are found under the PCA implementation and it contains most of the information provided by the estimated coefficients.

6.2 Recommendations

Although the initially proposed goals were not successfully achieved under the prescribed time we have managed to provide useful methodologies for the curve fitting procedures to be implemented by Anteryon and a practical methodology for the variability present in the description of a lens profile. That is, the results provided in this work can be used to describe the lens profile under the presence of correlated coefficients.

While the existing models at Anteryon for the estimations of the coefficients in a polynomial model are purely deterministic, there is still room for further improvements regarding implementation of curve fitting procedures under the framework of linear regression analysis.

Therefore we propose the following recommendations:

- Use the linear regression analysis for the coefficient estimation. This, apart from the estimations themselves, also the accuracy of such estimations are obtained and further analysis regarding regression diagnostics can be implemented (See Section 4.3 for further details).

- Use the linear regression analysis for the coefficient estimation directly in radial Zernike polynomials. As shown, the basis matrix transformation can be avoided and the linear model for the linear regression analysis can be set directly in term of radial Zernike polynomials.
When fitting a polynomial model, avoid the shifting step in the estimation procedure since it could yield unbiased estimators. Instead, use a polynomial model through the origin which is optimal with minimum variance.

When required, instead of partitioning and mirroring the data, use directly a even polynomial model over all data available.

Invest in exploring and learning open source software as the one introduced in Section 4.3. Using this software not only the estimated coefficients but extra information on the quality of the fitting can be obtained. Additionally, open source software allows flexibility in setting up the corresponding requirements for any investigation.

Although obtaining data is mostly time consuming and accuracy in the measurements is always a challenge, data is preferred over the surface of a lens rather than in the cross-section.

If data for the surface of a lens is available, the analysis can be directly implemented in terms of Zernike polynomials. There are indeed some advantages in using a 3D setting over a 2D setting. As already explained, the obtained coefficients are uncorrelated. Additionally, the 3D setting could reveal extra information on the symmetry of the lens around its radial axis.

We conclude the recommendations by mentioning that although the goals proposed at the beginning of the project were not achieved, we have managed to provide basis for future analysis regarding those goals. We refer to Chapter 7 for more details.
7 Future Research

As described at the end of the Introduction chapter (Chapter 1), as the project was going on, we encountered, that under the proposed time, the project goals were not feasible to be accomplished. We then moved our interest in describing, under the current procedures implemented in Anteryon, reliable measures of the variability present in a lens profile description. However, we have managed to set up the initial basis for the implementation of procedures in order to achieve the initial goals. We first start by describing the general approach of curve fitting over the lens surface in Section 7.1. Once the fitting has been performed, the estimates can be used as input in models for components of variance as explained in Section 7.2. Finally, Section 7.3 presents details about optimal designs in which we treat the optimal selection of lenses within a wafer both in number and locations. By proceeding in this order, we tackle one by one the initially proposed goals.

7.1 Estimation on the Surface of a Lens

In Chapter 4 we presented a entire set of curve fitting procedures on the cross-section of a lens. We encounter that, under the current fitting procedure implemented by Anteryon, either the polynomial coefficients or the radial Zernike coefficients are correlated random variables. Under the presence of correlation, the implementation of models for components of variance requires to consider jointly more than one coefficient at the time.

Data for the surface fitting consists of \( n \) 3-tuples \((x_1, y_1, z_1), (x_2, y_2, z_2), \ldots, (x_n, y_n, z_n)\) obtained under any measuring method implemented by Anteryon. Although data obtained from those methods have to be preprocessed as described in Section 3, we assume data is available ready for implementation of the surface fitting procedure. Data is then modeled by a polynomial in the radial and angular position. The \( n \) 3-tuples \((x_i, y_i, z_i)\) are given in the standard Cartesian coordinates, which can be transformed into Polar coordinates resulting in \( n \) 3-tuples \((\rho_i, \theta_i, z_i)\). The linear models, as introduced in (2.6), are of the form

\[
z_i = z(\rho_i, \theta_i) = \beta_1 Z_1(\rho_i, \theta_i) + \beta_2 Z_2(\rho_i, \theta_i) + \cdots + \beta_p Z_p(\rho_i, \theta_i) + \varepsilon_i
\]  \hspace{1cm} (7.1)

A Statistical Model to Describe the Spread of Wafer Lenses.
for \( i = 1, 2, \ldots, n \), the \( \beta_j \)'s are the model coefficients, the \( Z_j \)'s are the Zernike polynomials which are indexed by a single label and, as usual, it is assumed \( \varepsilon_i \sim N ID(0, \sigma^2) \). The single index notation for the Zernike polynomials follows a mapping from a multi-index \( j \) into \((n, m)\) tuples corresponding to the standard two-index notation of the Zernike polynomials (see Appendix A for details). The single indexing usually used was introduced by Noll (1976) which is given by

\[
Z^n_m(\rho, \theta) \rightarrow Z_j(\rho, \theta) \quad \text{where} \quad j = \frac{n(n + 2) + m}{2}.
\]

The index \( n \) of the Zernike polynomials should not be confused with \( n \), the number of observations. The curve fitting approach to compute the coefficients of a model such as (7.1) is by means of the Least Squares criterion already introduced in Chapter 2. We reproduce the statistical framework of linear regression to compute the respective coefficients. In matrix notation, the linear model, as (2.9), is given by

\[
Y = X\beta + \varepsilon
\]

with the corresponding vector \( Y \) of observations of size \( n \times 1 \), \( X \) the design matrix of size \( n \times p \), the vector \( \beta \) of size \( p \times 1 \) and the vector of errors assumed to be \( \varepsilon \sim N ID(0, \sigma^2 I) \). Then the Least Squares estimator \( \hat{\beta} \), as (2.16), is given

\[
\hat{\beta} = (X'X)^{-1}X'Y.
\]

Recalling Definition 2.4, we have that the \( i, j \) entry of the information matrix \( X'X \) corresponds to the discrete inner product of \( Z_i \) and \( Z_j \) given in Formula (2.12), that is

\[
[X'X]_{ij} = \langle Z_i, Z_j \rangle
\]

Since the Zernike polynomials are by definition orthogonal in the unit disc (see Appendix A) we have that the columns of the design matrix \( X \) are orthogonal and by Lemma 2.4 we conclude that the coefficients estimates are uncorrelated.

For further details regarding the orthogonality of the Zernike polynomials on a discrete set over the unit disc, we refer to Pap and Schipp (2005) and more recently Soumelidis et al. (2010).

### 7.2 Models for Components of Variance

Once the surface fitting has been performed, we might be able to proceed with the well-known models for components of variance described in the literature review (see Chapter 1). Implementation of these models follows from uncorrelated coefficients which can be accomplished given the orthogonality of the Zernike polynomials. We first give a general introduction to the Linear Mixed Models (LMM) for components of variance (see McCulloch and Neuhaus (2001), McCulloch and Searle (2001), West et al. (2007) and Wu (2010)) and then propose some models in order to quantify the wafer-to-wafer variability and the within-wafer variability.
7.2.1 Linear Mixed Models

A linear Mixed Model (LMM) is a parametric linear model that quantifies the possible relationships of a variable response and various predictor variables. These models are described by a linear relationship in the parameters and that the covariates, or independent variables that may involve a mix of fixed and/or random effects. LMM models are widely used to treat clustered, longitudinal or repeated-measured data. For instance, data from lenses produced under wafer techniques can be seen as clustered data in which each lens within a wafer is considered as the unit of analysis, the wafers are considered clusters of lenses and, possibly, wafers can be grouped into lots.

7.2.2 Proposed Models

More specific models were already introduced in literature review (see Chapter 1). The proposed models can be used to find out the different sources of variability that influences the total variation of the Zernike coefficients $\beta_j$’s. It is assumed that the variability of each Zernike coefficient comes from the differences of the wafers and of the estimations of each lens within wafers. Then we are often interested in controlling the mean, the wafer-to-wafer and within-wafer component of variability for each of the Zernike coefficients.

The input used in models for components of variance will be the estimations obtained in the Zernike fitting described in the preceding chapter. From each wafer, a sample of $L$ lenses is selected and $p$ Zernike coefficients are estimated. Since the estimated coefficients are uncorrelated random variables, each coefficients can be isolated and analyzed individually.

Model 1

We first try to identify where the estimated coefficients follows a radial and/or angular pattern within a wafer. In this model, each Zernike coefficient within a wafer is said to follow a linear combination of the distance from the lens to the centre of the wafer and the angle between the lens and the horizontal axis. So we have

$$Z_{ij} = \beta_{0j} + \beta_{1j} \gamma_i + \beta_{2j} \sin (\phi_i) + \varepsilon_{ij} \quad (7.2)$$

where $Z_{ij}$ is the $j^{th}$ Zernike coefficient of the $i^{th}$ lens, $\gamma_i$ is the radial distance from the lens $i^{th}$ to the center of the wafer, the angle $\phi_i$ is measured counter clockwise from the local $+x$ axis of the wafer, and $\varepsilon_{ij}$ is the error from the $j^{th}$ Zernike coefficient of the $i^{th}$ lens. Model 7.2 is nothing else but a linear regression model that might be analyzed under multiple regression techniques. In total, Model 7.2 accounts for $p$ linear relations.
(for $j = 1, 2, \ldots, p$). One may think in a possible influence of an interaction between the radial distance $\gamma_i$ and the angle $\varphi_i$. Therefore such model will look like

$$Z_{ij} = \beta_{0j} + \beta_{1j} \gamma_i + \beta_{2j} \sin(\varphi_i) + \beta_{3j} (\gamma \sin \varphi)_i + \varepsilon_i$$

We also can try to test a model where the input is the standard deviation of the coefficients given in the Zernike estimation performed in Section 7.1. Therefore, such model is in the form

$$\ln \sigma_{ij} = \beta_{0j} + \beta_{1j} \gamma_i + \beta_{2j} \sin(\varphi_i) + \varepsilon_i$$

where a usual choice for the variable response would be the natural logarithm of the estimated standard deviation. This model with interaction will look like similar as the one for the Zernike coefficient, see Equation 7.2. So

$$\ln \sigma_{ij} = \beta_{0j} + \beta_{1j} \gamma_i + \beta_{2j} \sin(\varphi_i) + \beta_{3j} (\gamma \sin \varphi)_i + \varepsilon_i$$

In the models for the Zernike coefficients, it is assumed $\varepsilon_i \sim N(0, \sigma^2_Z)$ meaning that the errors are independent and normally distributed with mean 0 and variance $\sigma^2_Z$ while in the models for the standard deviation of the Zernike coefficients the natural logarithm is required in order to make the assumption $\varepsilon_i \sim N(0, \sigma^2_{SD})$.

**Model 2**

Each Zernike coefficient may be seen as a “measured” characteristic of a lens in a certain population of lenses. We can also assume that such population has a mean $\mu$ and standard deviation $\sigma_i$. We can therefore decompose such variance by

$$\sigma^2_i = \sigma^2_w + \sigma^2$$

where $\sigma^2_i$ stands for the total variability, $\sigma^2_w$ represents the Wafer-to-Wafer component of total variance and $\sigma^2$ is the variance of the Zernike coefficient under consideration. Thus, we can say that each coefficient follows a model like

$$Z_{ij} = \mu + W_j + \varepsilon_{ij} \quad (7.3)$$

where $Z_{ij}$ is the Zernike coefficient of the $i^{th}$ lens of the $j^{th}$ wafer, $\mu$ is the Zernike coefficient mean, $W_j$ is the random effect of the $j^{th}$ wafer, and $\varepsilon_{ij}$ is the random noise representing the effect of the Zernike coefficient estimation of the $i^{th}$ lens from the $j^{th}$ wafer. For this model, it is assumed that $W_j$ and $\varepsilon_{ij}$ are independent and normally distributed with mean 0 and variance $\sigma^2_w$ and $\sigma^2$ respectively.

**Model 3**

As an explorative analysis, we could try a combination of the previous described models. Such model would look like

$$Z_{ij} = \mu + W_j + \left[ \beta_{1j} \gamma_i + \beta_{2j} \sin \varphi_i + \varepsilon_{ij} \right] \quad (7.4)$$

85 A Statistical Model to Describe the Spread of Wafer Lenses.
Description of each parameter on the previous model, is already given in the above described models.

Model 4

Because in practice there are situations in which a model of nested random effects presented in Model 7.3 does not adequately fit the process variation (see Roes and Does (1995)), a mixed model including one fixed effect in addition to the nested random effects can be considered. Such model can be represented by

$$Z_{ij} = \mu + L_i + W_j + \varepsilon_{ij}$$ (7.5)

where $Z_{ij}$ is the Zernike coefficient of the $i^{th}$ lens of the $j^{th}$ wafer, $\mu + L_i$ represents the fixed mean level of the lens $i^{th}$ ($\sum_{i=1}^{n} L_i = 0$ by assumption), $W_j$ models the random component between wafers and $\varepsilon_{ij}$ gives the random noise within a wafer. For this model, it is also assumed that $W_j$ and $\varepsilon_{ij}$ are independent and normally distributed with mean 0 and variance $\sigma_w^2$ and $\sigma^2$ respectively.

With the above described models might be able to quantity the differente components of variability regarding the industrial batch process of manufacturing lenses.

7.3 Optimal Designs

When investigating the relation of a variable response as a function of several explanatory variables it is usual to implement a regression model. When we have the possibility to choose the values of the explanatory variables we can investigate the quality of the experiment. Optimal designs are well-known when we can manipulate the experiments in order to achieve prescribe goals e.g. minimum variance of a particular estimator of the model under consideration or the total variance of the estimator. Regarding the number of points needed to be selected in order to guarantee orthogonality we refer to Pap and Schipp (2005). On the other hand, Soumelidis et al. (2010) proposed a mesh over the unit disc under applications of simulated corneal data. A full optimal design was investigated by Dette et al. (2007) where two classical criterions for the design selection was used. We describe the selection of lenses within a wafer in the particular case of a linear model involving Zernike polynomials under the general framework of Optimal Designs Theory presented by Silvey (1980). It is needed, in number and locations, a design that fulfills prescribed interests in a linear model model involving Zernike polynomials. We first provide assumptions for the implementation of the model and then describe how to implement different experiments.

86 A Statistical Model to Describe the Spread of Wafer Lenses.
Assumptions

The model in consideration is in the form of

\[ z = \beta_1 Z_1(x, y) + \beta_2 Z_2(x, y) + \cdots + \beta_p Z_p(x, y) + \epsilon \]  

(7.6)

Note the similitude with the model used for the estimation of a lens surface given by (7.1) in Section 7.1. Zernike polynomials allows to describe a lens surface given the symmetry properties in the unit disc. Since a wafer is similarly a disc, which under normalization can be taken as a unity, we can investigate different choices of \( x, y \) in (7.6) in such a way that we can evaluate the performance of each possible selection of lenses. In optimal design theory a model like (7.6) has the following characteristics:

- \( x, y \) the explanatory variables.
- \( x, y \in X \), the experimental region.
- \( Z_j \) is a continuous map from \( X \) into \( \mathbb{R} \), for \( j = 1, 2, \ldots, p \).
- \( \epsilon \) is the error term.

Note the similarities with those characteristics given for the linear model (2.6). This model allows continuous values for the explanatory variables within the unit disc. Since we are interested in selecting lenses, only a discrete selection over a wafer will be allowed. Each lens is specifically located within a wafer and the “coordinates” of its position can be given as in a Cartesian coordinate system. Hence the experimental region is described by \( X = \{(x, y) \mid x^2 + y^2 \leq r^2\} \) where \( r \) denotes the radius of the wafer.

Design

Since we aim to find an optimal design for the lenses to be selected and measured. A particular selection of lenses, which are determined by their positions \((x, y)\) within a wafer, determines the quality of the design. Consider we can take \( N \) lenses to be measured. In relation with model (7.6), it means we can select in advance values of \( x, y \) where we can evaluate each function \( Z_j \) and study the given relation. Such positions are labeled as paired positions \((x_i, y_i)\) for \( i = 1, 2, \ldots, N \). Then the model (7.6) can be analyzed under the linear regression framework. We set then the design matrix to be of the form

\[
X = \begin{bmatrix}
Z_1(x_1, y_1) & Z_2(x_1, y_1) & Z_3(x_1, y_1) & \cdots & Z_p(x_1, y_1) \\
Z_1(x_2, y_2) & Z_2(x_2, y_2) & Z_3(x_2, y_2) & \cdots & Z_p(x_2, y_2) \\
Z_1(x_3, y_3) & Z_2(x_3, y_3) & Z_3(x_3, y_3) & \cdots & Z_p(x_3, y_3) \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
Z_1(x_L, y_L) & Z_2(x_L, y_L) & Z_3(x_L, y_L) & \cdots & Z_p(x_L, y_L)
\end{bmatrix}
\]

87 A Statistical Model to Describe the Spread of Wafer Lenses.
and with the standard Least Squares approach we find parameter estimation of the form

$$\hat{\beta} = (X'X)^{-1} X'Y.$$  \hspace{1cm} (7.7)

with the standard convention for $Y$. The variance-covariance of the estimations are

$$\text{V}[\hat{\beta}] = \sigma^2 (X'X)^{-1}$$  \hspace{1cm} (7.8)

We are interested in controlling the experiment in such a way that we obtain satisfactory values for the variability of the estimators. Since (7.8) only depends on the positions $(x_i, y_i)$ and the model variance $\sigma^2$ we may choose the positions of $x$ and $y$ in such a way that (7.8) is minimal under any prescribed criterion. The problem then reduces to the information matrix of the design $X'X$, which is given by

$$X'X = \begin{bmatrix}
\sum Z_1(x_i, y_i)Z_1(x_i, y_i) & \sum Z_1(x_i, y_i)Z_2(x_i, y_i) & \cdots & \sum Z_1(x_i, y_i)Z_p(x_i, y_i) \\
\sum Z_2(x_i, y_i)Z_1(x_i, y_i) & \sum Z_2(x_i, y_i)Z_2(x_i, y_i) & \cdots & \sum Z_2(x_i, y_i)Z_p(x_i, y_i) \\
\vdots & \vdots & \ddots & \vdots \\
\sum Z_p(x_i, y_i)Z_1(x_i, y_i) & \sum Z_p(x_i, y_i)Z_2(x_i, y_i) & \cdots & \sum Z_p(x_i, y_i)Z_p(x_i, y_i)
\end{bmatrix}$$

Under this setting, following notation given by Silvey (1980), the information matrix is denoted by $M(x, y)$ and it is defined as

$$M(x, y) = \sum_{i=1}^{L} Z(x_i, y_i)Z(x_i, y_i)'$$

where $Z(x_i, y_i)$ is a column vector defined by

$$Z(x_i, y_i) = [Z_1(x_i, y_i), Z_2(x_i, y_i), \ldots, Z_p(x_i, y_i)]'$$

and the objective is to select values of $(x, y)$ to make $M(x, y)$ large as possible in some sense\(^1\). We next summarize the most common criterions for the quality of a design.

- **D-optimality**: This criterion selects a design in such a way that the determinant of the information matrix is large as possible.

- **G-optimality**: This criterion chooses a design in $(x, y)$ such that minimizes

$$\max_{c \in \mathcal{X}} c' [M(x, y)]^{-1} c$$

for a fixed $c = (c_x, c_y)$

\(^1\)As can be seen, the information matrix is fully given by inner products of the Zernike polynomials. However, the information matrix is not diagonal since only few points (lens locations), over the unit disc (wafer) are allowed to be selected. In this matter, we refer to Pap and Schipp (2005) where it is shown that orthogonality is not achieved when we choose $L$ points in a linear model like (7.6) such that $p > L$. 

88 A Statistical Model to Describe the Spread of Wafer Lenses.
• **E-optimality**: This criterion selects a design such that the minimum eigenvalue of $M(x, y)$ is maximized.

• **A-optimality**: This criterion selects a design for which the total variance of the estimator is minimized. That is, $tr\left[M(x, y)\right]^{-1}$ is minimal.

For more details on the above criterions we refer to Silvey (1980) where details and more criterions are described. An Optimal Design implies a search over a feasible region to find the optimal value of a prescribed objective function. Silvey (1980) remarks however that in the very strong sense optimality for an Optimal Design is only accomplished in exceptional circumstances.

Then, it remains to propose different schemes, in number and locations, as those given in Figure 3.2, and under any of the above described criterions select the scheme with the “best” performance over the others.

### 7.4 Summary

In this chapter we provide the basis achieve, one by one, the initially proposed goals. The estimation of the surface of a lens is described in Section 7.1. Several models for components of variance are proposed in Section 7.2 which after implementation gives notion in the variability present in the manufacturing process. Finally, under application of the optimal designs theory explained in Section 7.3 we can provide a satisfactory scheme for the number of lenses to be measured and their position in a wafer.
A Background Literature

In the section, the set of Zernike Polynomials, widely used to describe the corneal eye aberrations, are presented along with its properties, notation and a description of ZE-MAX software mostly used in the optics field will be introduced.

A.1 Zernike Polynomials

Zernike polynomials, firstly introduced by Zernike (1934), are a set of orthogonal polynomials defined on the unit disc with important characteristics which make them very useful in optical systems. Those polynomials are defined in Cartesian coordinates $x$ and $y$ which after a coordinate transformation from cartesian to polar coordinates can be expressed as a product of a polynomial in the radial coordinate $\rho$ and an harmonic function in the tangential coordinate $\theta$. The most general expression of the Zernike circle polynomials in polar coordinates is given by

$$Z^m_n(\rho, \theta) = R^m_n(\rho)e^{im\theta}$$  \hspace{1cm} (A.1)

where $n \in \mathbb{N}$ is called the radial order and $m \in \mathbb{N}$ is such that $-n \leq m \leq n$ is called the angular or azimuthal order. Born et al. (1970), Appendix VI, describes the following properties on the radial polynomial $R^m_n(\rho)$:

- $R^m_n(\rho)$ is a polynomial in $\rho$ of degree $n$ and contains no power of $\rho$ less than $m$.
- $R^m_n(\rho)$ is an even or an odd polynomial according as $m$ is even or odd.

The set of Zernike circle polynomials is distinguished from all other orthogonal sets in the unit circle by the property that it contains a polynomial for each $n, m$ such that $n - |m|$ even and $n > |m|$. Additionally we have that $0 \leq \rho \leq 1$ and $0 \leq \theta \leq 2\pi$. Given that only permissible values of $n$ (degree) and $m$ (angular dependence) can be expressed as $e^{im\theta}$ in powers of $\sin \theta$ and $\cos \theta$, (A.1) can be redefined as

$$Z^m_n(\rho, \theta) = R^m_n(\rho)\cos(m\theta)$$  \hspace{1cm} (A.2)

for $m \leq 0$ and

$$Z^m_n(\rho, \theta) = R^m_n(\rho)\sin(m\theta)$$  \hspace{1cm} (A.3)
Formulas (A.2) and (A.3) are usually referred to the even and odd Zernike polynomials, respectively. The angular functions are found to be the basis functions for a two-dimensional rotation group and the radial functions can be developed from the well-known Jacobi polynomials (for details see Born et al. (1970), Appendix VI). The expression of the radial part \( R_n^m(\rho) \) of the Zernike polynomials is given by

\[
R_n^m(\rho) = \sum_{s=0}^{\frac{n-m}{2}} (-1)^s \frac{(n-s)!}{(s)!(\frac{n+m}{2}-s)!(\frac{n-m}{2}-s)!} \rho^{n-2s}.
\]  

(A.4)

Because the Zernike circle polynomials are defined in cartesian coordinates \( x \) and \( y \) not all possible combinations of \( n, m \) in the polar coordinates \( \rho \) and \( \theta \) yield a polynomial in \( x, y \). For instance, consider \( n = 1 \) and \( m = 0 \) which would yield a polynomial in \( \rho \) of first order, besides consider a constant function in \( \theta \), so (A.1) becomes

\[
Z_1^0(\rho, \theta) = \rho
\]

which in cartesian coordinates is

\[
Z_1^0(x, y) = \sqrt{x^2 + y^2}
\]

which evidently is not a polynomial in \( x, y \). In practice, since not all combinations of \( n, m \) are allowed, it can be defined a radial Zernike polynomial to be identically 0 for \( n - m \). The set of the Zernike circle polynomials contains \( \frac{1}{2} (n + 1)(n + 2) \) linearly independent polynomials of degree \( \leq n \). Hence every monomial \( x^i y^j \) (\( i \geq 0, j \geq 0 \)) and every polynomial in \( x, y \) may be expressed as a linear combination of a finite number of the Zernike circle polynomials. A list of the first radial Zernike polynomials is given in Table A.1 while a extended list of the radial Zernike polynomials is presented in Table B.2.

<table>
<thead>
<tr>
<th>( n )</th>
<th>( m )</th>
<th>( R_n^m(\rho) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>( \rho )</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>( 2\rho^2 - 1 )</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>( \rho^2 )</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>( 3\rho^3 - 2\rho )</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>( \rho^3 )</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>( 6\rho^4 - 6\rho^2 + 1 )</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>( 4\rho^4 - 3\rho^2 )</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>( \rho^4 )</td>
</tr>
</tbody>
</table>

Table A.1: Radial Zernike polynomials up to degree \( n = 4 \)

As mentioned earlier, the set of Zernike polynomials are orthogonal in the unit disc. The orthogonality properties can be summarized as

91 A Statistical Model to Describe the Spread of Wafer Lenses.
• Orthogonality in the radial part
\[ \int_{0}^{1} R_{n}^{m}(\rho)R_{n'}^{m}(\rho)\rho d\rho = a_{n}^{m}\delta_{n,n'} \]
where \( n - m \) and \( n' - m \) are both even and \( a_{n}^{m} \) is a normalization constant.

• Orthogonality in the angular part
\[ \int_{0}^{2\pi} \cos(m\theta)\cos(m'\theta)d\theta = \epsilon_{m}\pi\delta_{m,m'} \]
\[ \int_{0}^{2\pi} \sin(m\theta)\sin(m'\theta)d\theta = (-1)^{m+m'}\pi\delta_{m,m'} \quad m \neq 0 \]
\[ \int_{0}^{2\pi} \cos(m\theta)\sin(m'\theta)d\theta = 0 \]

• Total orthogonality
\[ \int Z_{n}^{m}(\rho,\theta)R_{n'}^{m'}(\rho,\theta)\rho d\rho d\theta = \epsilon_{m}\pi \frac{\delta_{n,n'}\delta_{m,m'}}{2n+2} \]
where \( n - m \) and \( n' - m' \) are both even and where and \( \epsilon_{m} \) is defined as 2 if \( m = 0 \) and 1 if \( m \neq 0 \).

A.2 Notation and Zemax

Unfortunately there are various conventions regarding the Zernike polynomials, both with respect to normalization and ordering. Several authors, among them Malacara (2007), Noll (1976) and Born et al. (1970) uses a different notation which in most cases lead to misunderstanding in interpretation. The ordering and normalization, most commonly used for statistical purposes corresponds to the Noll’s ordering and ANSI normalization. Additionally, ZEMAX Optical Design Program is widely used for lens design, illumination, laser beam propagation, stray light, freeform optical design and many other applications uses as “Zemax standard Zernike” the ANSI normalization and Noll’s ordering.

Notation

Noll (1976) introduced a slightly different notation than the usual for the set of Zernike polynomials. The difference is that the normalization chosen turns out to be more convenient for statistical analysis. The typical way of reporting Zernike polynomials uses both

\[ \epsilon_{m} \]
usually is called the Neumann factor. It often appears in in conjunction with Bessel functions
indexes $n$ and $m$, which unambiguously defines a typical polynomial of the set; however the two index notation may lead to miscommunication in reporting the eye aberrations. Because of that, Noll (1976) also introduce a single index notation in the form

$$Z_m^n(\rho, \theta) \rightarrow Z_j(\rho, \theta) \quad \text{where} \quad j = \frac{n(n+2) + m}{2}.$$  

Given this single index notation, the Zernike polynomials are then defined as

$$Z_{\text{even} \ j} = \sqrt{n+1} R_n^m(\rho) \sqrt{2} \cos(m\theta)$$

$$Z_{\text{odd} \ j} = \sqrt{n+1} R_n^m(\rho) \sqrt{2} \sin(m\theta) \quad m \neq 0$$

and so this notation appeals to be more convenient because it gives a logical ordering in the Zernike polynomials and the orthogonality relations can be easily be written in a single index as

$$\int W(\rho) Z_j Z_{j'} d^2\rho = \delta_{jj'}$$

where $d^2\rho = \rho d\rho d\theta$ and $W(\rho) = 1/\pi$ for $\rho \leq 1$ and defined to be identically 0 for $\rho > 1$.

Zemax

ZEMAX is a well-known Optical Design Program to evaluate a measured data in spherical surfaces. It is widely used by optical engineers and designers for choosing a lens design, illumination, laser beam propagation, stray light, fiber optics, and many other optical technology applications. A description of the ZEMAX software can be found in the literature from ZEMAX Corporation (see www.zemax.com):

The ZEMAX Optical Design Program is a comprehensive software tool for optical design. ZEMAX integrates all the features required to conceptualize, design, optimize, analyze, tolerance, and document virtually any optical system. All of these powerful features are integrated into an intuitive user interface. ZEMAX offers power, speed, flexibility, ease of use, and value in one comprehensive program.

The expression for the Zernike polynomials is:

$$R_n^m(\rho) e^{im\theta} = \begin{cases} R_n^m \cos m\theta \\ R_n^m \sin m\theta \end{cases}$$

where $n$ denotes the radial order, $m$ denotes the azimuthal index and, $0 < m < n, n + m$ even. For this definition, the angle $\theta$ is measured in radians and counter clockwise from

93 A Statistical Model to Describe the Spread of Wafer Lenses.
the local $+x$ axis, and $\rho$ is the dimensionless normalized radial coordinate. The ZEMAX software is able to compute orthonormal Zernike coefficients on which the indexing used is given by Noll (1976) as previously mentioned. Table B.1 presents the first 28 Zernike polynomials under Noll's notation also given in Zemax (2007) under the title of “Zernike Standard Polynomials".
B Tables and Lists

Table B.1 Zernike Polynomials under the Noll’s notation
Table B.2 Extended list of radial Zernike polynomials
Table B.3 Extended list of radial Zernike polynomials for selected $m$ values
Table B.4 Basis Matrix transformation up to $j = 25$
Table B.5 Inverse of Basis Matrix transformation up to $j = 25$
<table>
<thead>
<tr>
<th>Term</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z1</td>
<td>1</td>
</tr>
<tr>
<td>Z2</td>
<td>$\sqrt[4]{4}\rho \cos(\theta)$</td>
</tr>
<tr>
<td>Z3</td>
<td>$\sqrt[4]{4}\rho \sin(\theta)$</td>
</tr>
<tr>
<td>Z4</td>
<td>$\sqrt[3]{(\rho^2 - 1)}$</td>
</tr>
<tr>
<td>Z5</td>
<td>$\sqrt[6]{\rho^2 \sin(2\theta)}$</td>
</tr>
<tr>
<td>Z6</td>
<td>$\sqrt[6]{\rho^2 \cos(2\theta)}$</td>
</tr>
<tr>
<td>Z7</td>
<td>$\sqrt[8]{(3\rho^3 - 2\rho) \sin(\theta)}$</td>
</tr>
<tr>
<td>Z8</td>
<td>$\sqrt[8]{(3\rho^3 - 2\rho) \cos(\theta)}$</td>
</tr>
<tr>
<td>Z9</td>
<td>$\sqrt[8]{\rho^3 \sin(3\theta)}$</td>
</tr>
<tr>
<td>Z10</td>
<td>$\sqrt[8]{\rho^3 \cos(3\theta)}$</td>
</tr>
<tr>
<td>Z11</td>
<td>$\sqrt[5]{(6\rho^4 - 6\rho^2 + 1)}$</td>
</tr>
<tr>
<td>Z12</td>
<td>$\sqrt[10]{(4\rho^4 - 3\rho^2) \cos(2\theta)}$</td>
</tr>
<tr>
<td>Z13</td>
<td>$\sqrt[10]{(4\rho^4 - 3\rho^2) \sin(2\theta)}$</td>
</tr>
<tr>
<td>Z14</td>
<td>$\sqrt[10]{\rho^4 \cos(4\theta)}$</td>
</tr>
<tr>
<td>Z15</td>
<td>$\sqrt[10]{\rho^4 \sin(4\theta)}$</td>
</tr>
<tr>
<td>Z16</td>
<td>$\sqrt[12]{(10\rho^5 - 12\rho^3 + 3\rho) \cos(\theta)}$</td>
</tr>
<tr>
<td>Z17</td>
<td>$\sqrt[12]{(10\rho^5 - 12\rho^3 + 3\rho) \sin(\theta)}$</td>
</tr>
<tr>
<td>Z18</td>
<td>$\sqrt[12]{(5\rho^5 - 4\rho^3) \cos(3\theta)}$</td>
</tr>
<tr>
<td>Z19</td>
<td>$\sqrt[12]{(5\rho^5 - 4\rho^3) \sin(3\theta)}$</td>
</tr>
<tr>
<td>Z20</td>
<td>$\sqrt[12]{\rho^5 \cos(5\theta)}$</td>
</tr>
<tr>
<td>Z21</td>
<td>$\sqrt[12]{\rho^5 \sin(5\theta)}$</td>
</tr>
<tr>
<td>Z22</td>
<td>$\sqrt[7]{(20\rho^6 - 30\rho^4 + 12\rho^2 - 1)}$</td>
</tr>
<tr>
<td>Z23</td>
<td>$\sqrt[14]{(15\rho^6 - 20\rho^4 + 12\rho^2) \sin(2\theta)}$</td>
</tr>
<tr>
<td>Z24</td>
<td>$\sqrt[14]{(15\rho^6 - 20\rho^4 + 12\rho^2) \cos(2\theta)}$</td>
</tr>
<tr>
<td>Z25</td>
<td>$\sqrt[14]{(6\rho^6 - 5\rho^4) \sin(4\theta)}$</td>
</tr>
<tr>
<td>Z26</td>
<td>$\sqrt[14]{(6\rho^6 - 5\rho^4) \cos(4\theta)}$</td>
</tr>
<tr>
<td>Z27</td>
<td>$\sqrt[14]{\rho^6 \sin(6\theta)}$</td>
</tr>
<tr>
<td>Z28</td>
<td>$\sqrt[14]{\rho^6 \cos(6\theta)}$</td>
</tr>
</tbody>
</table>

Table B.1: Zernike Polynomials under Noll’s notation used in ZEMAX
<table>
<thead>
<tr>
<th>$n$</th>
<th>$m$</th>
<th>$R_n^m(\rho)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>$\rho$</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>$2\rho^2 - 1$</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>$3\rho^3 - 2\rho$</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>$6\rho^4 - 6\rho^2 + 1$</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>$\rho^3$</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>$6\rho^4 - 6\rho^2 + 1$</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>$4\rho^4 - 3\rho^2$</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>$\rho^4$</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>$10\rho^5 - 12\rho^3 + 3\rho$</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>$5\rho^5 - 4\rho^3$</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>$\rho^5$</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>$20\rho^6 - 30\rho^4 + 12\rho^2 - 1$</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>$15\rho^6 - 20\rho^4 + 6\rho^2$</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>$6\rho^6 - 5\rho^4$</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>$\rho^6$</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>$35\rho^7 - 60\rho^5 + 30\rho^3 - 4\rho$</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>$21\rho^7 - 30\rho^5 + 10\rho^3$</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>$7\rho^7 - 6\rho^5$</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>$\rho^7$</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>$70\rho^8 - 140\rho^6 + 90\rho^4 - 20\rho^2 + 1$</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>$56\rho^8 - 105\rho^6 + 60\rho^4 - 10\rho^2$</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>$28\rho^8 - 42\rho^6 + 15\rho^4$</td>
</tr>
<tr>
<td>8</td>
<td>6</td>
<td>$8\rho^8 - 7\rho^6$</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>$\rho^8$</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>$126\rho^9 - 280\rho^7 + 210\rho^5 - 60\rho^3 + 5\rho$</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>$84\rho^9 - 168\rho^7 + 105\rho^5 - 20\rho^3$</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>$36\rho^9 - 56\rho^7 + 21\rho^5$</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
<td>$9\rho^9 - 8\rho^7$</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>$\rho^9$</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>$252\rho^{10} - 630\rho^8 + 560\rho^6 - 210\rho^4 + 30\rho^2 - 1$</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>$210\rho^{10} - 504\rho^8 + 420\rho^6 - 140\rho^4 + 15\rho^2$</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>$120\rho^{10} - 252\rho^8 + 168\rho^6 - 35\rho^4$</td>
</tr>
<tr>
<td>10</td>
<td>6</td>
<td>$45\rho^{10} - 72\rho^8 + 28\rho^6$</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
<td>$10\rho^{10} - 9\rho^8$</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>$\rho^{10}$</td>
</tr>
</tbody>
</table>

Table B.2: Extended table of Radial Zernike Polynomials

97 A Statistical Model to Describe the Spread of Wafer Lenses.
A Statistical Model to Describe the Spread of Wafer Lenses.

\[ R_\alpha^m(\rho) \]

Table B.3: Extended list of Radial Zernike Polynomials for \( m = 0, 1 \)
Table B.4: Basis Matrix Transformation T
Table B.5: Inverse of the Basis Matrix Transformation $T$
Bibliography


101 A Statistical Model to Describe the Spread of Wafer Lenses.


102 A Statistical Model to Describe the Spread of Wafer Lenses.


103 A Statistical Model to Describe the Spread of Wafer Lenses.