Approximation of multi-variable signals and systems : a tensor decomposition approach

Approximation of multi-variable signals and systems: a tensor decomposition approach

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van
de rector magnificus, prof.dr.ir. C.J. van Duijn, voor
een commissie aangewezen door het College voor
Promoties in het openbaar te verdedigen op
maandag 6 juni 2011 om 16.00 uur

door

Femke van Belzen

geboren te Stuttgart, Duitsland

Dit proefschrift is goedgekeurd door de promotoren:

prof.dr. S. Weiland
en
prof.dr.ir. A.C.P.M. Backx

Copromotor:
prof.dr.ir. P.P.J. van den Bosch

This work is supported by the Dutch Technologiestichting STW under project number EMR.7851.

A catalogue record is available from the Eindhoven University of Technology Library
ISBN: 978-90-386-2489-1

Approximation of multi-variable signals and systems: a tensor decomposition approach

Samenstelling promotiecommissie:

prof.dr. S. Weiland
prof.dr.ir. A.C.P.M. Backx
prof.dr.ir. P.P.J. van den Bosch
prof.dr.ir. J. de Graaf
prof.dr.ir. P.M.J. van den Hof
dr. L. De Lathauwer
prof.dr.ir. J.M.A. Scherpen
dr.ir. O.M.G.C. Op den Camp

Contents

Chapter 1

Introduction

1.1 Introduction

In most works of fiction, the reader only discovers the true meaning of a novel's title *after* the last page of the book has been read. What you are about to read is not a work of fiction, but a PhD thesis and the title that has been chosen reflects the aim of this work. Therefore, contrary to a novel, this first chapter will be devoted to explaining the aim, and therefore the title, of this work. This aim will serve as a guide throughout this thesis, as the chapters that follow will reveal increasing amounts of detail.
The approach that will be followed in this chapter is comparable to that of the system of Google Earth satellites. They can be used to view the entire world and then gradually zoom in, until they reveal your home and backyard, or your holiday destination. In a similar way, we will start this chapter by introducing some global issues that are the driving forces behind the development of new technology of which this work is a small part and by continually zooming in further, we will end up explaining the aim of this work and stating the contributions of this thesis.

1.2 Towards a brighter future

Two very interesting phenomena are occurring in our society. On the one hand, there is a desire to eradicate extreme poverty and hunger, while on the other hand, there is an increased concern for our planet and the need to secure its magnificence for future generations. These two phenomena are seemingly conflicting. Indeed, increased living standards for the poor can be realized by providing access to the technology that increased the living standards in the West. Providing this access implies an even

greater strain on the environment and an increase in our use of natural resources. Therefore, to realize a brighter future for our planet and its inhabitants we should increase the living standards of the poor and at the same time decrease our environmental footprint. It is this goal that both politicians and scientists set out to achieve. Politicians set targets, scientists look for ways to achieve these targets. This thesis is about science, not politics, and I hope that it is part of this road towards a brighter future.

The first of the United Nation's eight millennium goals is to 'Eradicate Extreme Poverty and Hunger' [58]. According to World Bank figures of 2008, 13.6% of the world population live on less than 1 US dollar a day and 79.7% of the world population, excluding industrialized nations, live on less than 10 US dollar a day, which is the poverty line for industrialized nations [19]. Needless to say, it is imperative to bridge this poverty gap. At the close of a United Nations (UN) summit in September 2010 a global action plan was launched to achieve this and seven other millennium goals by 2015. Looking back into history, Western prosperity levels were raised by the increase of productivity thanks to industralization. Therefore, the way to eradicate extreme hunger and poverty globally requires technology that will allow for another such increase of productivity.

In parallel to these efforts, due to economic growth in Asian countries like China and India, but also in Brazil and Russia, the number of people who attain Western living standards increases steadily, from currenly 600 million people, to 2.5 billion people in the near future. This increase in prosperity has two important implications. Firstly, increased prosperity implies that less people will be willing to work in health-threatening production environments. Cheap human labor in countries like China will be history soon. This will create a problem in the future, since our industry partially relies on this cheap human labor. Secondly, our planet does not provide sufficient traditional sources of energy and natural resources to supply 2.5 billion people in the same way it is currently supplying 600 million people. This implies that a new generation of sustainable technology is needed. This is reflected in Millennium goal number seven, which is to 'Ensure Environmental Sustainability' [58].

Meanwhile, in the industrialized world conflicting trends are visible in a similar way. On the one hand, there is a steady demand from the public for technological innovation, while on the other hand, environmental legislation ànd awareness are also putting demands on the development of future technology. In the Spring of 2010 US corporation Apple sold 1 million of its iPad devices within the first 28 days of its release [65]. This is just an example of the continuous pressure on industry to keep evolving and keep putting innovative products on the market. At the same time governments are setting clear targets via environmental legislation. These targets have to be met by industry somehow. As an example, in a European Union agreement,

the Dutch government has pledged that by 2020 14% of the energy consumed in the Netherlands will be produced by renewable resources [79]. These two developments put seemingly conflicting demands on industry: to preserve its market share it has to keep innovating, while parts of traditional technology can no longer be re-used due to environmental legislation.

What is needed to build a brighter future is a new generation of technology that is able to achieve three goals. First is to increase the living standards of people in developing nations. Second, to ensure that the living standards of industrialized nations are preserved. Third, to preserve our planet and its resources for future generations. This new technology should have no footprint. This means that natural resources should be fully reused and energy should be supplied from renewable resources. This footprint-free technology implies that production processes should be run highly automated, to ensure that they are as efficient as possible and do not rely on manual labor. Development of a new generation of technology requires a significant effort in terms of research and innovation. This thesis forms a small step towards the development of this new generation technology. The focus of this work has been on efficient operation of industrial production processes.

Industrial production is largely responsible for the use of our planet's resources. A large part of the available energy and natural resources go to manufacturing plants, ranging from power plants, to manufacturing of electronic devices, to production of pharmaceutical products etc. As an example, in the Netherlands 55% of the total energy used in 2009 went to industrial manufacturing, including energy production facilities. In contrast, only 13% of the total energy was consumed by Dutch households [78]. Due to the expected shortage of natural resources and the desire to produce footprint-free, environmental legislation already pressures the industry into decreasing its use of energy and natural resources. In the future, this pressure will only increase. Therefore, it is imperative to develop technology that will allow existing processes to be operated more efficiently and enable the development of future generation footprint-free production processes. This work contributes to the development of this technology. In the next section, we will take a closer look at industrial production and the challenges it faces.

1.3 Industrial production

As mentioned above, industrial production is responsible for a large part of the use of energy and natural resources across the globe. In this section we will explain what we mean exactly by industrial production and how efficient operation of industrial production processes can contribute to the issues of sustainability, eradication of poverty

and technological innovation.
Industrial production is the process of taking a supply, performing some work on it such that it becomes a product that is either used as supply for another production process or as an end-product by the consumer. Most products require processing by different branches of industry before they are finalized. Examples of industrial production are numerous, one can consider the production of electrical energy in a nuclear power plant, manufacturing of Integrated Circuits such as processors that are then built into PCs in other facilities, etc.
Development of industrial production has changed western life drastically since the industrial revolution. Through the use of steam power, petrochemical processing and modern computing power, productivity has risen drastically causing a major increase in living standards. As mentioned before another new wave of technological innovation is needed to increase prosperity outside of the Western countries as well.
The major challenge that industrial production currently faces is to increase its sustainability. This means two things. Firstly, further automation of existing production processes. This automation will further decrease the use of scarce natural resources. Secondly, development of a new generation of production processes that is able to achieve the global increase in prosperity and is footprint free.
To make the discussion more specific, we divide industrial production processes into five categories [16]

1. Project: this is a one-of-a-kind product that will likely only be produced once, such as a large building.

2. Job process: these processes are designed for flexibility. The equipment can be used for a range of products and the people working in such a process are usually highly skilled.

3. Batch or Intermittent processing: the equipment in these facilities tends to be more specialized, but still there is a range of products. Facilities tend to produce a large number of one product before changing their setup to allow production of another product. Many products are produced in this way, such as clothing, pharmaceutical products, glues etc.

4. Repetitive processing: these processes are used to produce a very large volume of a very limited variety of products. The equipment used is highly specialized, requiring very little, usually unskilled, manual labor. One can think of robotic assembly and the production of most consumer electronics.

5. Continuous processing denotes the situation where production is smooth and uninterrupted in time, although production rates may vary over time. There is

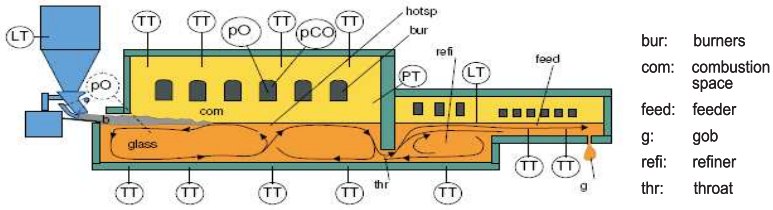


Figure 1.1: Cross section of a glass furnace

a constant influx of supplies and a constant output of end result. Examples include (petro-)chemical processing, industrial distillation, glass production, steel production and production of electrical energy using steam generators. Continuous processing is generally used to produce large quantities of product per year.

Some processes may have characteristics from all categories, but nevertheless this categorization will be useful in our exposition. The remainder of this work considers efficient operation of *continuous processes* only. We will investigate techniques that allow continuous processes to be operated more efficiently and can at the same time be used to develop next-generation continuous processes. An example of a continuous process is now introduced in more detail.

Example 1.3.1 (Glass manufacturing)**.** *An example of a continuous production process is the manufacturing of glass. The process of making glass roughly consists of melting raw materials at high temperature and letting the molten glass circulate and mix for a while to give it the opportunity to attain certain properties. The glass is then gradually cooled to proper temperatures at which it can be formed into (intermediate) products. Figure 1.1 shows a cross section of a glass furnace. There is a continuous influx of raw materials at the left and a continuous supply of energy through the burners above the furnace. A glass furnace can be compared to a swimming pool, filled with molten glass that circulates at approximately* 1500° *C. The right part of the furnace is called the working end and the feeder, in this part the cooling of the glass takes place. Production of glass is very energy-intensive, approximately* 40% *of production costs are energy costs. Natural gas is burned to produce the heat that is necessary to melt the raw materials.*

The circulations indicated in Fig. 1.1 are generated by temperature differences. The quality of the end product is highly dependent on the temperature and velocity profiles present in the furnace.

1.4 Operation of continuous processes

Operation of continuous industrial processes in a systematic manner requires process knowledge. Usually this knowledge is available from two sources. The first carrier of process knowledge is the experience of the process operators. These specialists by experience instinctively know what will work and what will not work, which situations should be avoided and how this can be achieved. Mathematical models form the second source of process knowledge. These models describe the evolution of physical variables such as temperature, concentration, flow etc. The interaction of physical principles and chemical reactions, heat transfer and material flows are captured by these models. The models are usually derived from the laws of physics and are based on the assumption that energy, mass and momentum are conserved quantities. In all cases the model describes the evolution of the physical variables over time and for some processes the evolution of physical variables also varies in space. In the glass furnace, Example 1.3.1, the temperature and flow profiles vary over time, but also as a function of the different locations in the furnace. A model of a glass furnace therefore describes the evolution of temperature over both space and time.
Most of the time, process models are formulated in terms of mathematical expressions. In case of evolution over time only, the mathematical description is often a collection of ordinary differential equations, describing the change in the physical quantities as a function of change of time. In case of evolution over both space and time, the mathematical equations describe the change in the physical quantities as both time and space change. This description is therefore usually in terms of Partial Differential Equations.
Process models are usually generic, yet they are used in different contexts, namely Research and Development, Production, and Maintenance. They are used in simulation to design and improve understanding of what is happening inside a plant. Process models are also used for monitoring purposes, to check whether the process is on the right track. Thirdly, a process model can be used to evaluate different strategies for operation. To make this clearer, recall the glass furnace example. Usually, a furnace is used to produce a range of products. Say that we are operating a container glass furnace which is producing brown-colored beer bottles. The next order that needs to be fulfilled consists of clear marmalade containers. To go from the brown to the clear-colored glass, furnace operation has to be adjusted. There are different ways

to achieve this change in set point and a process model can be used to predict which strategy will be fastest, or use the least energy.
From a management perspective, a reliable mathematical process model is a useful tool. Although the practical knowledge of process operators is often indispensable, it may take years to build that experience and there is no systematic way to transfer this knowledge from one person to the next. While it may only take a couple of weeks to get familiar with a mathematical model and to learn how to use and maintain it. Business may benefit from relying not only on the knowledge of operators but also using the knowledge available in process models during day-to-day operation. Furthermore, mathematical models may give additional insight into production processes, since a wider variety of scenarios can usually be tested, compared to those that can be experimentally verified. Finally, a mathematical model is an indispensable tool in process analysis and optimization.
Let's take a closer look at what one would want to use a process model for from an operation and control perspective. One would like to use a process model to determine how to go from the current average operating situation to a desired improved operating range, realizing a certain set of wishes, such as the shortest time, or the highest energy efficiency. While such a path is pursued, unexpected occurrences have to be taken into account and the model should tell us how to achieve our goal given the new situation. This is exactly the way in which a car navigation system works.

Example 1.4.1 (Car Navigation)**.** *Consider a car navigation system. The navigation system uses a GPS signal to determine the car's current location, say we are in Hochstetterstrasse 23, Hemmingen, Germany. The driver enters a desired destination, say Den Dolech 2, Eindhoven, The Netherlands and a wish. In this case our driver would like to get to Eindhoven as fast as possible. Let's say the driver invested in an expensive navigation system and the maps of all of Europe are available as well as up-to-date information on traffic jams and construction sites. Given the current location, the destination and the information on traffic jams, the navigation system determines which maps to use and what would be the fastest route. It uses the maps of Germany and the Netherlands as its model. When a serious traffic jam occurs near Venlo, the system re-calculates the route to go via Aachen and Maastricht instead of Venlo.*

In case of the glass furnace, the operator would like to use the process models in a way similar to that of Example 1.4.1. The process model, which in case of the glass furnace is a set of coupled partial differential equations, is equivalent to the maps of Europe in the navigation example. The navigation system was able to determine that it only

needed a small part of the maps of Germany and the Netherlands to plan a route. To go from brown-colored to clear glass, one should be able to extract from the process models only the trajectories that are relevant to such a color change. Furthermore, if a disturbance occurs, for example the composition of the raw materials is different from what was expected, there should be a selection mechanism available to re-calculate the relevant trajectories. Just in the same way that the navigation system was able to re-route in the case of a traffic jam. And finally, similar to the driver's wish of getting to Eindhoven as fast as possible, the operator of the glass furnace would like to be able to decide which trajectory will give him the fastest transition from brown-colored to clear glass.

To summarize, there are a couple of elements that we would like to extract from a process model in a control context. Given the current state, a desired operating range and a target, e.g. minimizing energy, we want to select from the model the trajectories that are relevant, the one that minimizes the cost involved in realizing our target and we want to be able to do this repeatedly, so that we are able to deal with disturbances and changing desires. Naturally, in the context of other model purposes, such as process design or process modeling, the specific elements we wish to extract from the model may be different. However, the techniques discussed in this work will remain relevant in these contexts.

Assuming a process model is available, it is not always straightforward to extract the information that is needed from this model. When dealing with processes where the physical variables evolve over space and time especially, there are not many tools available to automatically extract the relevant information within reasonable time. The difficulty is the following. The mathematical expressions that constitute a process model describe all possible trajectories of physical variables such as temperature and concentration over space and time, for all possible disturbances. We say that the model describes the global behavior of the process. There exist no mathematical techniques to extract from this generic process model the information that is relevant to the current operating condition or current trajectory. Ideally, one would like to obtain from the generic process model a description of those trajectories that are relevant to the current situation. Unfortunately, the mathematical techniques that are currently available do not allow this.

1.5 Extracting information from process models

Extracting information from mathematical process models is not straightforward, as mentioned in the previous section. To explain why this is the case, we explain how numerical mathematical techniques deal with process models. Given a process model

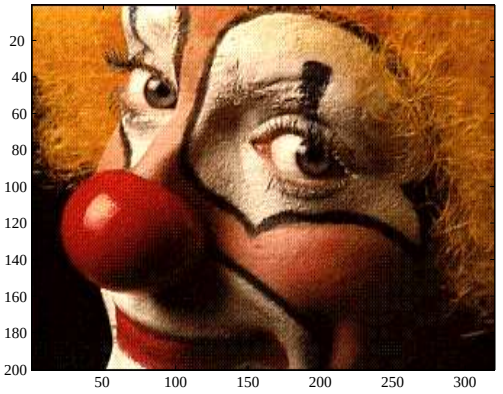
that describes the change in the physical variables as a function of the change in space and time. Information about individual trajectories is obtained by dividing the global problem into a large number of small problems. These small problems are then pieced together like a mosaic to form a trajectory of the full system. Unfortunately, this division into small problems introduces a lot of computational overhead, since one has to make sure that the individual pieces in the mosaic represent a whole when combined. All this bookkeeping translates to high computation times. Therefore, although numerical mathematical techniques are a useful tool to extract information from process models, in most cases it takes too much time to extract the information. If the information comes too late it may no longer be relevant. In the car navigation example the re-route via Aachen and Maastricht is only useful for the driver if this alternative route is determined *before* the driver is stuck in the traffic near Venlo. If he is already in the middle of this traffic jam, providing an alternative route is useless.

The devil is in the details, as they say, and this is most certainly the case for process models and the accompanying numerical techniques. The process models contain all details of all possible trajectories of the process. This is precisely the reason why most numerical techniques become so computationally complex. The issue is that we are hardly interested in all of these details. For the glass furnace and the change from brown-colored to clear glass, the process operators want to use a process model to determine the *direction* in which they should steer the furnace and the details will be taken care of in another way. The details that are available in the process models are not of interest in this case. Unfortunately, no method exists yet, that will extract information from a process model at a certain level of detail.

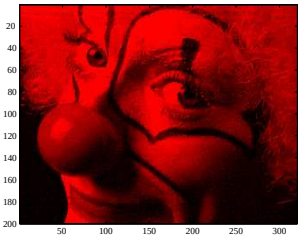
To explain what we mean by level of detail consider Figure 1.2. In this figure, an image of a clown is visible. This image was coded in RGB colors and thus consists of tree layers. It is possible to infer from each layer individually that a clown is pictured. However, if we want to see what color eyes this particular clown has, the combination of the three layers is needed to see that the clown indeed has brown eyes.

In this work, we are looking for automated procedures that are able to decide what information is relevant at a certain operating point or trajectory. This information should be extracted from the generic process model into a compact model. This compact model should be accurate for the purpose at hand. The techniques we are looking for should provide an alternative to existing numerical methods.

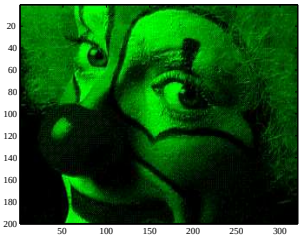
If we are able to extract this information from generic process models, industrial production processes can be further automated. Based on the information extracted from compact models, systematic decisions about process operation in the most efficient manner can be made. This decreases the footprint of currently existing production processes. Process models describe the interaction of physical mechanisms, if we are able to extract information about this interaction on different levels of detail, this in-



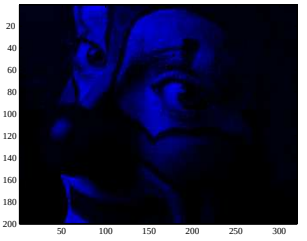
(a) Original Image



(b) Red Layer



(c) Green Layer



(d) Blue Layer

Figure 1.2: The clown image consists of three different layers. Each individual layer also carries information, independent of the others.

formation can be used to develop new production processes. This way, information extracted from process models can contribute towards developing sustainable technology for the future. Therefore, the techniques developed in this work contribute both towards sustainable operation of industrial production processes and towards the development of next generation production processes.

1.6 Aim and contributions of this work

The premise of this chapter was to explain the title and thus the aim of this work. Specifically, we will state what is meant by *Approximation of multi-variable signals and systems*. Given a mathematical process model, approximation is the process of extracting certain system trajectories from this model. In this case, the main interest is in those system trajectories that are relevant to the control objective. Approximation is the step that neglects the unnecessary detail to obtain only the essence of the process one is interested in. *Multi-variable signals and systems* refers to the class of process models that is under consideration. Namely, we consider process models that describe the evolution of multiple physical quantities, over both space and time. Rephrased, the aim of this work is to derive mathematical techniques that allow approximation of multi-variable systems.
The method that has been developed and that is described in this thesis allows for such approximations. It links traditional approximation techniques to concepts from multilinear algebra. This way, the approximation method is able to explicitly deal with multiple physical variables and a spatial-temporal domain. The advantage that is thus gained is that it gives extra flexibility in choosing approximation levels in each spatial variable. Furthermore, the scaling of the physical quantities can be tuned separately.
To make the aim and contributions of this work more specific, some additional concepts have to be introduced. This will be done in the next chapter, which again ends with a problem statement and contributions of this work.

Chapter 2

Problem statement

2.1 Introduction

An important class of distributed systems models the evolution of signals that evolve both in space as well as in time. Examples of such systems can be found in virtually all engineering disciplines including fluid dynamics, aerodynamics, seismology, etc. Usually, first-principle models of these systems involve coupled sets of Partial Differential Equations (PDEs) that are inferred from physical conservation laws.

Today, many commercial and dedicated packages exist that allow an efficient simulation of such models. These numerical tools operate via discretization of the spatial and temporal domain of the signals via finite elements or finite volumetric elements. The accuracy of these methods largely depends on the density of the mesh, where fine meshes need to be generated at spatial locations or temporal instants where large signal variations occur. By doing so, the system dynamics, represented by the PDE, are typically located in each and every element in the grid by copying the physical laws in every element, and describing the interconnections of individual elements and their neighbors. In this way, the original global problem as described by the PDE is translated into a large number of local problems and their interconnections via Finite Element (FE) or Finite Volume (FV) methods, see Fig. 2.1. Depending on the specific application, the number of finite elements or finite volumetric elements may be substantial and easily lead to large-scale models that require the solution of up to $10^6 - 10^8$ equations at every time step.

The large number of equations that have to be solved for each time-step in FE or FV methods leads to a number of problems. Most naturally, the sheer number of equations already makes simulation of FE (FV) models computationally demanding. This

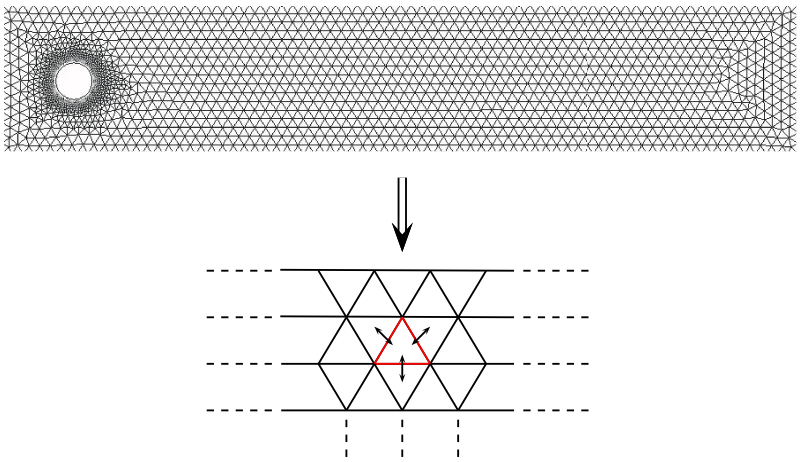


Figure 2.1: This figure shows the mesh that was generated for the FE model of a flow around an obstacle [21]. A global problem is replaced by multiple local problems plus their interconnections, this introduces computational overhead.

may prevent the models from being used for process monitoring if the simulations are slower than real-time. Furthermore, the number of equations also implies that these models are not suitable for model-based control design, since the complexity of the model is usually a lower-bound for the complexity of the controller. Finally, the large size of FE (FV) models implies that they are less suitable for analysis of system properties such as stability, reachability, passivity, etc.

To reduce computation time and to enable the use of model-based analysis and design tools, it then becomes necessary to construct simplified models that consist of considerably smaller number of equations. These substitute models should be of lower complexity, yet retain the information that is relevant for simulation, control design, analysis, etc. The central question then becomes, given the FE model and a desired objective, how to extract relevant information from the FE model so as to substantially reduce its complexity. Here, substantial will mean to reduce the number of equations from $10^6 - 10^8$ to less than 10^2.

As we shall describe in more detail below, approximation of multi-variable systems involves both a signal and a system approximation step. Signals that are a function of multiple variables or indices occur in all fields of science and engineering. Consider for example measurements of the distribution of temperature across the globe during a certain period of time. This temperature varies as a function of location and time.

Since locations are referenced by three coordinates, the temperature measurements are indexed by four indices. Three of these indices refer to the location, the fourth to time. One can come up with scores of examples of signals that are a function of multiple indices or multiple independent variables.
Generally, it is not the signals themselves one is interested in, but the information they carry. The global temperature distribution is generally used to derive temperature gradients or find out the temperature at specific locations across the globe. This and other types of information can be extracted from multi-variable signals using analysis tools. Signal approximation is one of these signal analysis tools.
Formally, signal approximation may be viewed as a low-rank approximation problem. Whenever the signal under consideration is a function of multiple variables, i.e. a multi-dimensional signal, low-rank approximations can be obtained via multi-linear functionals, tensors. We will show that approximation of multi-variable *systems* also boils down to low-rank approximation problems. And, for multi-dimensional systems, the solution of the system approximation problem involves the use of tensors.
The work presented in this thesis builds on previous work in this area, [2, 41, 80]. Here the aim was to derive approximate models of multi-variable systems with the specific purpose of deriving models that are suitable for control of large scale processes such as the glass furnace introduced in Example 1.3.1. As is explained in the references mentioned, the model of a glass furnace involves a combination of Navier-Stokes equations [12] and heat transfer in a three-dimensional spatial domain. The premise of this work is to examine the implications of a Cartesian structure in the spatial domain on the approximation process.
It is the aim of this chapter to explain the role of multi-linear functionals, tensors, in low-rank approximation of multi-dimensional signals and systems. We will introduce the signal and system approximation problems, give a formal problem statement and provide an overview of the main body of this thesis.

2.2 Signal approximation

Signals that evolve over multi-dimensional domains are the focus of this section. In general, these signals themselves are not of immediate interest. One is interested in the information contained in these signals, where the context defines what information is of importance. Extraction of information from signals then becomes a two-step process. The first step is to decide which part or property of the signal contains the information. The second step is to actually extract this specific part or property.
As an example consider the one-dimensional periodic signal in Example 2.2.1. Suppose that the information of interest is the low-frequent component of the signal. This

information can be extracted by making a signal decomposition in which spectral components represent the harmonic content of the signal. This is done with classical Fourier analysis. An approximation w_r of the original signal w is obtained by projection of w on a subspace spanned by a set of harmonic functions. This projection is the *operation* that extracts the information from w. In this case, we obtain the low-frequent component of w which is represented by w_r.

Example 2.2.1. *Consider the approximation of one-dimensional periodic signals. Let* $w : \mathbb{X}_1 \to \mathbb{R}$ *be of period* 2π *and continuous, i.e.* $w(x_1 + 2\pi) = w(x_1)$ *for all* $x_1 \in \mathbb{X}_1$. *Then,* w *may be approximated by a truncated Fourier series as follows. By definition, the Fourier series of* w *of order* r *is the trigonometric series*

$$w_r(x_1) := a_0 + \sum_{k=1}^{r} a_k \cos(kx_1) + b_k \sin(kx_1) \tag{2.1}$$

with coefficients a_k *and* b_k *given by the Euler formulas [51]*

$$a_0 = \frac{1}{2\pi} \int_0^{2\pi} w(x_1) dx_1$$

$$a_k = \frac{1}{2\pi} \int_0^{2\pi} w(x_1) \cos(kx_1) dx_1, \quad k = 1, 2, \ldots$$

$$b_k = \frac{1}{2\pi} \int_0^{2\pi} w(x_1) \sin(kx_1) dx_1, \quad k = 1, 2, \ldots.$$

With this approximation, we have convergence in the sense that

$$\lim_{r \to \infty} \|w - w_r\| = 0$$

where $\|w - w_r\|^2 = \int_0^{2\pi} \mid w(x_1) - w_r(x_1) \mid^2 dx_1$, *i.e. convergence in the* L_2 *norm on* $[0, 2\pi]$. *This particular approximation* w_r *of* w *can be viewed as a projection as follows. Let* $\mathcal{X} = L_2[0, 2\pi]$ *and let* $\mathcal{X}_r = \text{span}\{1, \cos(kx_1), \sin(kx_1) \mid k = 1, \ldots, r\}$. *Then we have that* $w_r = \Pi_{\mathcal{X}_r} w$, *where* $\Pi_{\mathcal{X}_r}$ *represents the orthogonal projection of* $w \in \mathcal{X}$ *onto* $\mathcal{X}_r$.

Figure 2.2 shows the approximation of a block signal w, w_r, *for different values of* r. *Here, complexity of* w_r *is measured by the number of independent harmonic functions in* $\mathcal{X}_r$. *The projection of* w *on* $\mathcal{X}_r$ *extracts the low-frequent content of* w.

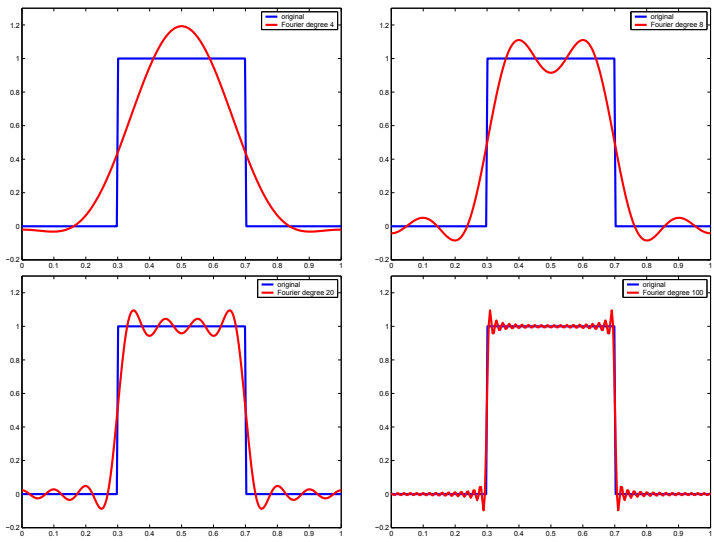


Figure 2.2: Approximation of a block signal by Fourier Expansions of different orders. From top-left in clockwise order the approximation degrees are 4, 8, 20 and 100.

In this work, we focus on extracting information via projection of the original signal on low-dimensional subspaces. This projected signal generally is an approximation of the original signal. The subspaces that are considered for projection depend entirely on the context. In Example 2.2.1 we considered harmonic functions for frequency analysis, but one may also consider subspaces spanned by polynomials etc. In this thesis, the focus is on low-dimensional subspaces spanned by *empirical* basis functions. By this we mean that projection spaces are inferred from measured or simulated data, acquired from the process.

Just like the information contained in a signal depends on the context, the complexity of a signal depends on the choice of basis functions used to represent the signal. The complexity may be described by the bandwidth, or rather, more generally speaking, by the dimension of the span of basis functions that are used to represent the signal. Such basis functions can be harmonic functions in case of Fourier analysis, but may also be polynomials, etc. The number of basis functions used to represent the signal is the rank of the signal in terms of these basis functions. Low-rank approximation of signals then means representation of the signal with respect to a smaller number of basis functions. The projected signal is then said to be of lower complexity that the

original.
In general, *signal projections* are defined as follows. Consider a signal $w : \mathbb{D} \to \mathbb{R}^n$ where the domain $\mathbb{D}$ consists of a finite interval $[0, L] \subset \mathbb{R}$. Throughout, we consider the discretized signals only, i.e. we assume that the domain $\mathbb{D}$ has been sampled into a finite set of points, i.e. $\mathbb{X} = \{p_x^{(1)}, \ldots, p_x^{(L_X)}\} \subset \mathbb{D}$. We assume this discretization step has already been carried out and that the error introduced by this discretization is sufficiently small.
Let $\mathcal{X}$ be the space of functions $g : \mathbb{R}^{L_X} \to \mathbb{R}^n$ with associated inner product

$$\langle \xi_1, \xi_2 \rangle = \sum_{k=1}^{L_X} \langle \xi_1(x_k), \xi_2(x_k) \rangle_n \tag{2.2}$$

where $\langle \cdot, \cdot \rangle_n$ denotes the Euclidean inner product in $\mathbb{R}^n$. Furthermore, let $\{f^{(k)}\}$ be an orthonormal basis for $\mathcal{X}$. Then, any $\underline{w} \in \mathcal{X}$ can be expressed as a spectral expansion

$$\underline{w} = \sum_k \underline{w}_k f^{(k)}$$

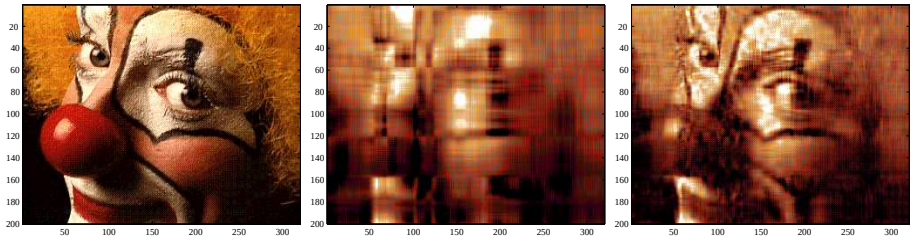
where $\underline{w}_k$ are called the coefficients of $\underline{w}$ with respect to the basis $\{f^{(k)}\}$ of $\mathcal{X}$. Since we consider an orthonormal basis $\{f^{(k)}\}$ of $\mathcal{X}$, the coefficients $\underline{w}_k$ satisfy

$$\underline{w}_k = \langle \underline{w}, f^{(k)} \rangle, \quad k = 1, 2, \ldots$$

and are uniquely determined by $\underline{w} \in \mathcal{X}$ and the basis $\{f^{(k)}\}$. Let $\mathcal{X}_r \subseteq \mathcal{X}$ be the r-dimensional subspace $\mathcal{X}_r = \operatorname{span}\{f^{(1)}, \ldots, f^{(r)}\}$ and let $\Pi_{\mathcal{X}_r}$ denote the orthogonal projection of elements in $\mathcal{X}$ onto $\mathcal{X}_r$. Then the approximation $\underline{w}_r$ of $\underline{w}$ is defined as

$$\underline{w}_r = \Pi_{\mathcal{X}_r} \underline{w} = \sum_{k=1}^{r} \underline{w}_k f^{(k)}. \tag{2.3}$$

In this work, we consider subspaces $\mathcal{X}_r$ that are spanned by empirical basis functions. By replacing the original signal by one of lower complexity, signal approximation allows information to be stored in a more compact manner. Furthermore, the low-complexity replacement signal usually allows for faster computations, for example in post-processing of signals. Finally, signal approximation may give information about the phenomena underlying the signal. For example, it may give information about the system that generated the signal. These underlying phenomena may be of lower complexity than the dimension of the original data may imply. Referring to Example 2.2.1, a question relevant here is how to find the best approximation of a certain degree of complexity, given the original signal, where 'best' is measured in the norm associated with $L_2[0, 2\pi]$.



(a) Original Image (b) Rank-5 Approximation (c) Rank-15 Approximation

Figure 2.3: Optimal rank approximation of the clown image. The original image is of size 200×300 and the matrix describing it has rank 200. In the middle and on the right two rank approximations of the original image are shown.

We will now consider approximation of two-dimensional signals on finite, discrete domains. Let $\underline{w} : \mathbb{X}_1 \times \mathbb{X}_2 \to \mathbb{R}$ be a two-dimensional signal and $\mathbb{X}_k = \{p_k^{(1)}, \ldots, p_k^{(L_k)}\}$ for $k = 1, 2$. Then, the signal values can be represented by matrices $\underline{w} \in \mathbb{R}^{L_1 \times L_2}$. Consider the problem of finding decompositions of $\underline{w}$ in terms of rank-one matrices as follows

$$\underline{w} = \sum_k w_k \underline{f}^{(k)},$$

where $\underline{f}^{(k)} \in \mathbb{R}^{L_1 \times L_2}$ is of rank one. Approximations of $\underline{w}$ may be obtained via truncation of this decomposition in terms of rank-one matrices. The degree of complexity of the approximation is in this case given by the rank of the approximating matrix. The solution to the question of finding the best rank-k approximation of a matrix, was presented in [27]. As is discussed in detail in Appendix A.3, the solution to the problem of optimal rank approximation of matrices can be found via the Singular Value Decomposition (SVD). An imaging example of optimal rank approximation to matrices is shown in Fig. 2.3. This shows an image which can be represented by a 200×300 matrix of rank 200. The middle and right of Fig. 2.3 show rank approximations of the image. These low-rank approximations do not capture all detail of the original, yet it is clear that one is looking at a (distorted) image of a clown. In the case of $N > 2$, i.e. when considering higher order tensors, the question of optimal rank approximations becomes much more involved, as we shall see later on.

The signal approximation problem we will consider is the following. We consider signals that evolve over a multi-dimensional domain of independent variables. To be more specific, we consider signals $\underline{w}(p_1, \ldots, p_N) \in \mathbb{R}^n$ with $(p_1, \ldots, p_N) \in \mathbb{X} = \mathbb{X}_1 \times \cdots \times \mathbb{X}_N$ where $\mathbb{X}_k$ is a set of finite cardinality L_k, i.e. $\mathbb{X}_k = \{p_k^{(1)}, \ldots, p_k^{(L_k)}\}$.

We want to obtain an approximation $\underline{w}_r$ to $\underline{w}$ such that the error $\|\underline{w}-\underline{w}_r\|$ is minimized in some norm while the structure of the independent variables is kept intact. The reason this problem is more difficult to solve than the one- or two-dimensional case, is that there are different rank concepts for multi-dimensional signals. Each of these rank concepts lead to different low-rank approximations. Hence, a method to compute optimal low-rank approximations to multi-dimensional signals does not yet exist.
The solution strategy to the problem of low-rank approximations to multi-dimensional signals is the following. We will associate a tensor W with $\underline{w}$ and determine low (modal-)rank approximations to this tensor. These approximations define projection spaces $\mathcal{X}_k^{(r_k)} \subseteq \mathcal{X}_k$ for $k = 1, \ldots, N$, where $\mathcal{X}_k^{(r_k)} = \operatorname{span}\{\varphi_k^{(1)}, \ldots, \varphi_k^{(r_k)}\}$. These projection spaces are then used to define low-rank approximations to the signal $\underline{w}$. This way, an approximation $\underline{w}_r$ of $\underline{w}$ is defined via an approximation W_r of the tensor W associated with $\underline{w}$. Since there may be different generalizations of the rank concept to tensors, there are also different notions of low-rank approximations W_r of W. The computation of W_r in a systematic manner is one of the problems considered in this work.

2.3 System approximation

In mathematical terms, the complexity of a model is usually defined in terms of the number of coupled first-order differential or difference equations [1]. The accuracy of a model is usually defined by comparing measured and simulated process data where the same excitation is used for the model and the real process. Naturally, one would like the accuracy of a process model to be as high as possible. This usually increases both the cost of obtaining the model and the model complexity.
One would like the complexity of a process model to be as low as possible. This, because the implications of a high model complexity on model-based control design and simulation of the model are manifold. The simulation time of high-complexity models is large. Therefore, it may not be possible to use such a model for simulation purposes. Furthermore, large simulation times also impede on-line model-based optimization. In most model-based control strategies the complexity of the controller is equal to or exceeds the complexity of the process model. This may lead to problems in the derivation and implementation of such a controller.
There is a clear trade-off between complexity and accuracy. Namely, increased accuracy implies increased complexity and vice versa. However, the situation is not as black-and-white as it appears to be. This can be seen when one takes into account what the process model will actually be used for. Generally, not all aspects of the process behavior are relevant to the purpose for which the model will be used. For

example, if one is interested in steady-state process behavior, a model that accurately describes not only the steady-state behavior but also the transients basically contains a lot of details that are not relevant. Another model of significantly lower complexity that accurately predicts steady-state behavior but is less accurate for the transients may be a good substitute for its purpose, with lower complexity.
The low-complexity replacement model should satisfy a number of demands:

- The error between the original model and its replacement should satisfy some upper bound.
- If the replacement model is to be used in a predictive control setting, the computation time of the model should be sufficiently small.
- Qualitative properties of the original model should be preserved as much as possible in the replacement model. Examples of such properties include symmetry, dissipativity, stability, conservation of energy, etc.

Several model approximation techniques exist. Techniques that are suitable for linear, lumped systems include balancing, Hankel norm reduction a and Krylov methods. Balancing is also applicable to nonlinear lumped systems. The method of Proper Orthogonal Decompositions (POD) is one of the few model reduction methods that is suitable for distributed systems. POD is also known under the names of Principal Component Analysis (PCA) [44] and Karhunen-Loève-Decomposition [56]. The essence of all of these methods is the same. Namely, they seek to obtain model approximations via projection of the original (state space) equations onto some lower-dimensional subspace. This concept of approximation via projection will be introduced for lumped systems in the next subsection first. Then we will extend this framework to discrete-domain distributed systems.

2.3.1 System approximation via projections

We will consider a special case of approximation through projection, namely that of a lumped system in input/state/output-form. This case is generally known and can be found for instance in [1]. Consider the following system

$$\Sigma := \begin{cases} x(t+1) & = f\left(x(t), u(t)\right) \\ y(t) & = g\left(x(t), u(t)\right). \end{cases} \tag{2.4}$$

Here, $x(t) \in \mathbb{R}^n$ is the state vector, $u(t)$ is the input and $y(t)$ is the output. We will derive a Petrov-Galerkin projection for this system. Consider the following projection space

$$\mathcal{X}_r := \{x : \mathbb{Z} \to \mathbb{R}^n \mid x(t) \in \mathcal{V}\}$$

here $\mathcal{V} \subset \mathbb{R}^n$ is a subspace of dimension r. Furthermore, consider a subspace $\mathcal{W} \subset \mathbb{R}^n$ of dimension r and suppose that $\mathcal{V} = \operatorname{Im} V$ and $\mathcal{W} = \operatorname{Im} W$, where V and W are matrices of dimension $n \times r$.
State approximation will be considered first.Decompose $x(t)$ according to

$$x(t) = \underbrace{\hat{x}(t)}_{\in \mathcal{V}} + \underbrace{\tilde{x}(t)}_{\in \mathcal{V}^\perp}. \tag{2.5}$$

Here, $\hat{x}(t) = \Pi_{\mathcal{X}_r} x(t)$ is the projection of the state on the lower-dimensional subspace $\mathcal{V}$ and $\tilde{x}(t) = (I - \Pi_{\mathcal{X}_r})x(t) = x(t) - \hat{x}(t)$. $\hat{x}(t)$ is defined by

$$\begin{aligned} \hat{x}(t) &= \Pi_{\mathcal{X}_r} x(t) && (2.6)\\ &= V \underbrace{(V^\top V)^{-1} V^\top x(t)}_{x_r(t)} && (2.7)\\ &= V x_r(t). && (2.8) \end{aligned}$$

where $x_r(t) \in \mathbb{R}^r$ will be the new state vector.
The residual equation that will be projected is $x(t+1) - f(x(t), u(t)) = 0$. We define the following approximate residual projection

$$\langle x(t+1) - f(x, u), \xi \rangle = 0, \quad \forall \xi \in \mathcal{W}. \tag{2.9}$$

Combined with the approximation $\hat{x}(t)$ of $x(t)$ this gives the following reduced order model

$$\begin{aligned} \langle \hat{x}(t+1) - f(\hat{x}, u), \xi \rangle &= 0, \quad \forall \xi \in \mathcal{W} \\ \langle V x_r(t+1) - f(V x_r, u), \xi \rangle &= 0, \quad \forall \xi \in \mathcal{W} \\ \langle V x_r(t+1) - f(V x_r, u), W\varphi \rangle &= 0, \quad \forall \varphi \in \mathbb{R}^r \\ \langle W^\top V x_r(t+1) - W^\top f(V x_r, u), \varphi \rangle &= 0, \quad \forall \varphi \in \mathbb{R}^r. \end{aligned}$$

The last line implies that

$$W^\top V x_r(t+1) = W^\top f(V x_r(t), u(t)) \tag{2.10}$$

which defines the approximate model of state dimension $\dim(\mathcal{X}_r) = r$. A Galerkin projection assumes that $W = V$, which leads to the new system $\hat{\Sigma}$

$$\hat{\Sigma} = \begin{cases} x_r(t+1) & = (V^\top V)^{-1} V^\top f(V x_r, u) \\ y(t) & = g(V x_r, u). \end{cases} \tag{2.11}$$

The crucial question behind this system approximation is the following. Given the system Σ, how to choose the projection spaces $\mathcal{W}$ and $\mathcal{X}_r$ such that the error between Σ and $\hat{\Sigma}$ is sufficiently small in a suitable norm. All existing model reduction methods provide a solution strategy to precisely this question and it also this question that will be considered in this work. The class of systems under consideration in this work is that of distributed systems on discrete domains, defined in Appendix A.2.
Although the state vector of the system $\hat{\Sigma}$ is of lower dimension than the original state vector, function evaluations $(V^\top V)^{-1}V^\top f(Vx_r, u)$ in (2.11) still require the computation of f on (Vx_r, u) which are elements in the original state and input space. If f is nonlinear, this implies that the computational efficiency of $\hat{\Sigma}$ is hardly improved with respect to that of Σ. This problem is addressed in more detail in [3, 17] and in Chapter 5 of this work.

2.3.2 Approximation of multi-variable systems

The concept of system approximation via projection, as introduced for lumped systems in the previous section can be formulated for distributed systems as follows.
Consider an arbitrary linear distributed system described by the following Partial Difference Equation

$$D(\varsigma_1, \varsigma_1^{-1}, \ldots, \varsigma_N, \varsigma_N^{-1})\underline{w} = 0. \tag{2.12}$$

Here $D \in \mathbb{R}^{\ell\times n}[\xi_1, \ldots, \xi_N, \eta_1, \ldots, \eta_N]$ is a real matrix-valued polynomial in $2N$ indeterminates and ς_k (ς_k^{-1}) is the forward (backward) shift operator acting on the spatial discretization in the kth mode according to Definition A.2.1. The domain of the signal $\underline{w}$, will be denoted by $\mathbb{X}$. Solutions $\underline{w}$ to this PDE assume the form $\underline{w} : \mathbb{X} \to \mathbb{R}^n$ where $\mathbb{X}$ is a set of finite cardinality, say L_X. This type of system is formally introduced in Section A.2.
Let Ξ be a set of functions $f : \mathbb{X} \to \mathbb{R}^\ell$ equipped with the following bi-linear form

$$\langle \xi_1, \xi_2 \rangle_\Xi = \sum_{k=1}^{L_X} \langle \xi_1(x_k), \xi_2(x_k) \rangle_{N+1} \tag{2.13}$$

where $\langle \cdot, \cdot \rangle_{N+1}$ denotes the standard Euclidean inner product in $\mathbb{R}^\ell$. We will first define a *residual projection* of (2.12) on Ξ as

$$\langle D(\varsigma_1, \varsigma_1^{-1}, \ldots, \varsigma_N, \varsigma_N^{-1})\underline{w}, \xi \rangle_\Xi = 0, \qquad \forall \xi \in \Xi. \tag{2.14}$$

Equation (2.14) will be viewed as a new, and weaker, constraint on the variable $\underline{w}$. Indeed, any solution $\underline{w}$ of (2.12) satisfies (2.14), but the converse is obviously not true. We call (2.14) a residual projection of $D(\varsigma_1, \varsigma_1^{-1}, \ldots, \varsigma_N, \varsigma_N^{-1})\underline{w} = 0$ onto Ξ. We will

be especially interested in finite dimensional subspaces Ξ_r of mappings $\mathbb{X} \to \mathbb{R}^\ell$, say of dimension r, i.e. $\Xi_r = \text{span}\{\xi^{(1)}, \ldots, \xi^{(r)}\}$ with $\{\xi^{(\ell)}\}$ a basis of Ξ_r. The expression (2.14) then becomes

$$\langle D(\varsigma_1, \varsigma_1^{-1}, \ldots, \varsigma_N, \varsigma_N^{-1})\underline{w}, \xi\rangle_\Xi = 0, \qquad \forall \xi \in \Xi_r. \tag{2.15}$$

The system associated with (2.15) is interpreted as the solution set $\underline{w}$ that satisfies (2.15).
The residual projection can also be combined with the signal approximation concepts as given in (2.3). That is, we now consider a projection space $\mathcal{X}_r$ for $\underline{w}$ and consider again the PDE (2.12). We will use the approximate residual projection, (2.15) and substitute into this equation a projected signal $\underline{w}_r$. This gives

$$\langle D(\varsigma_1, \varsigma_1^{-1}, \ldots, \varsigma_N, \varsigma_N^{-1})\underline{w}_r, \xi\rangle_\Xi = 0, \qquad \forall \xi \in \Xi_r,\ \underline{w}_r = \Pi_{\mathcal{X}_r}\underline{w}. \tag{2.16}$$

This projection method is called a *Petrov-Galerkin* projection. Whenever the two projection spaces are equal, i.e. $\mathcal{X}_r = \Xi_r$, (2.16) is called a *Galerkin* projection.
The question this work deals with is the following. Consider the case that the domain $\mathbb{X}$ of (2.12) has Cartesian structure, i.e. $\mathbb{X} = \mathbb{X}_1 \times \cdots \times \cdots \mathbb{X}_N$. Furthermore we are interested in the case when $\mathcal{X}_r = \Xi_r$ and $\mathcal{X}_r$ is obtained from data. That is, we consider the problem of computing empirical projection spaces in case the domain $\mathbb{X}$ has Cartesian structure. The computation of these projection spaces is again a low-rank approximation problem for a signal (our data) on a Cartesian grid. The solution of this problem again involves low-rank approximations to tensors, as we shall see in Chapter 4.

2.4 Problem statement

The aim of this work is to develop numerical techniques for the approximation of large-scale multi-variable systems, i.e. systems where the evolution of the state is over both space and time. The techniques that will be developed allow the construction of low-complexity replacement models from large-scale Finite Element models. These replacement models can then be used for on-line process monitoring, model-based control design for example.
This aim translates into signal and system approximation problems, for signals and systems defined on multi-dimensional domains. The main body of this thesis consists of three parts. The common factor between these chapters is that tensors are associated with multi-dimensional signals on discrete Cartesian grids. These tensors are then used to solve the original approximation problems.
The following problem statements can be formulated for each chapter:

1. Chapter 3 considers the problem of finding low-rank approximations to tensors. Specifically, we consider the following problems:

   (a) Given a tensor $W : \mathcal{X}_1 \times \cdots \times \mathcal{X}_N \to \mathbb{R}$ of modal rank $R = (R_1, \ldots, R_N)$, find a tensor $W_r : \mathcal{X}_1 \times \cdots \times \mathcal{X}_N \to \mathbb{R}$ of modal rank $r = (r_1, \ldots, r_N)$ where $r_k \leq R_k$, $k = 1, \ldots, N$ such that the error $W - W_r$ is minimized in Frobenius and/or operator norm.

   (b) If such a low-rank approximation method is found, what are its properties and can the error $\|W - W_r\|_F$ or $\|W - W_r\|$ be characterized?

   (c) Derive a method for numerical computation of W_r.

   (d) Demonstrate the low-rank approximation method in a numerical example and compare its performance with existing methods.

2. Chapter 4 considers the problem of finding approximations to systems that evolve over a multi-dimensional domain. Specifically, we consider a distributed dynamical system Σ on a discrete domain $\mathbb{X}$, as in Def. A.2.3. Furthermore, we assume that the domain $\mathbb{X}$ has Cartesian structure, i.e. $\mathbb{X} = \mathbb{X}_1 \times \cdots \times \mathbb{X}_N$. The question that will be addressed in Chapter 4 is that of finding a replacement model $\hat{\Sigma}$ to Σ via the Galerkin projection method (2.16). The projection spaces are required to be empirical and have a Cartesian structure.

3. Chapter 5 considers the problem of reconstruction of multi-dimensional signals that have been sampled on a non-uniform Cartesian grid. Specifically, we consider signals $w : \mathbb{X} \to \mathbb{R}$, where $\mathbb{X} = \mathbb{X}_1 \times \cdots \times \mathbb{X}_N$. The questions that will be addressed are the following:

   (a) Given a subset of sample points $\mathbb{X}_0 \subseteq \mathbb{X}$ and the restriction $\tilde{w} := w|_{\mathbb{X}_0}$, under which conditions is it possible to exactly recover w from $\tilde{w}$ via a reconstruction map $R : \mathbb{X}_0 \to \mathbb{X}$ such that $\|w - R(\tilde{w})\| = 0$.

   (b) In case exact reconstruction is not possible, can we characterize the error $\|w - R(\tilde{w})\|$?

2.5 Reading guide per chapter

This section provides an overview of the contents of the main body of this work

Chapter 3

Chapter 3 considers the problem of finding low-rank approximations to tensors. For order-2 tensors, matrices, this problem is well understood, see Appendix A.3. Generalization of these results to higher-order tensors, however, is not straightforward. Finding tensor decompositions that allow suitable approximations after truncation is an active area of research [47], to which this chapter contributes in the following way. The problem of low-rank approximations to tensors is ill-posed, see [26] for a thorough discussion and overview of this issue. Therefore, we consider a different rank concepts, referred to as multi-linear or modal rank, and define a method to obtain such tensor decompositions. This method will be referred to as *TSVD*, which is short for Tensor SVD. The naming of this method is for convenience only, there are many other SVD-type tensor decomposition methods, of which the HOSVD [24] is the most known. In Chapter 3 we derive properties of the TSVD and in certain cases we give error bounds when the method is used for low-rank approximations to tensors. In Sec. 3.7 we propose an adaptation of the TSVD method that may give better approximation results when not all modal directions are approximated. This adaptation will be referred to as *dedicated TSVD*. In Sec. 3.8 we propose a numerical algorithm for the computation of the (dedicated) TSVD. With a small adaptation, this algorithm can also be used to compute successive rank-one approximation to tensors. Finally, in Sec. 3.9, we include a simulation example which demonstrates the methods proposed in this work and compares them to a well-known existing method.

Especially in the signal processing community, tensors are commonly viewed as multi-dimensional arrays. Since changes of coordinate systems are among the most elementary operations, we believe that it is particularly important to understand tensors as general multi-linear functionals. This is reflected in the way this chapter has been written. The results in this chapter have been published in [81, 7].

Chapter 4

Chapter 4 considers the problem of finding system approximations. As discussed in Sec. 2.3, methods for finding system approximations all rely on projection of the state vector and the system dynamics on low-dimensional subsystems. The method of Proper Orthogonal Decompositions (POD) is such a model reduction method. POD is suitable for systems that evolve over multi-dimensional domains. As we will explain in our review of POD in Chapter 4, the projection spaces that are considered in POD are empirical projection spaces, derived from measured or simulated data.

Whenever the system domain $\mathbb{X}$ is a Cartesian domain, i.e. $\mathbb{X} = \mathbb{X}_1 \times \cdots \times \mathbb{X}_N$, tensors can be used to compute the empirical projection spaces. Specifically, measured

(simulated) data on a Cartesian domain defines a tensor and decompositions of the kind discussed in Chapter 3 can be used to compute projection spaces that are used to derive the system approximations.
We first introduce the POD method as it is found in literature. Then we incorporate tensors in the case of a Cartesian structure of the domain. The results of this chapter have been published in [7, 8].

Chapter 5

Chapter 5 considers the problem of reconstruction and approximation of multi-dimensional signals, but now in the case that these signals are sampled with non-uniformly distributed sensors. The motivation for this problem statement stems from the Missing Point Estimation (MPE) method derived in [3]. The MPE technique aims to decrease the computational cost of the reduced models derived using POD by considering system dynamics on a selection of grid-points only. The MPE method was developed for one-dimensional signals.
Here, we consider multi-dimensional signals on a Cartesian domain $\mathbb{X} = \mathbb{X}_1 \times \cdots \times \mathbb{X}_N$. Furthermore, we define the restriction of the signal $w : \mathbb{X} \to \mathbb{R}$ to a subset $\mathbb{X}_0 \subseteq \mathbb{X}$, as a *sampling* of w.
The central question of this chapter is that of finding a reconstruction $\hat{w}$ of w from the sampled signal $\tilde{w}$. We consider a reconstruction map R and present conditions for exact reconstruction of w from $\tilde{w}$. In case that exact reconstruction is not possible, we derive an expression for the reconstruction error. The results of this chapter have been published in [5].

Chapter 3

Tensor Decompositions

3.1 Introduction

This chapter proposes techniques to analyze and approximate tensors by low-rank approximations. For the matrix case (that is, tensors of order 2), the problem of finding low-rank approximations is well understood. The solution consists of truncating a dyadic expansion (i.e., a finite sum of orthonormal rank one matrices) of the matrix, that is directly inferred from its singular value decomposition [27]. For higher-order tensors, this problem has been studied by many authors, such as [18, 26, 46, 48, 55, 24]. With the approximation error defined by the Frobenius norm, and with a suitable notion of tensor rank, it was found that the optimal lower rank tensor approximation problem is ill-posed in the sense that optimal low rank approximations may fail to exist or may not be unique. More specifically, the space of rank r tensors is non-compact and the non-existence of low-rank approximations occurs for many different ranks and orders, regardless of the norm, see [26] for an overview of these issues. The existence, uniqueness and computability of optimal lower rank approximations of higher order tensors has therefore been recognized as a major problem in numerical multi-linear algebra.

Within the existing literature, one can distinguish two main classes of tensor decompositions. The first one is known as a Tucker decomposition [75] and represents an order N tensor T as the product of a core tensor of the same size as the original one together with N nonsingular matrices whose columns span the domain of each of the arguments of T. A special case of this decomposition is the higher order singular value decomposition (HOSVD) that has been proposed in [24]. The second class of decompositions amounts to representing T as a linear combination of normalized

rank-1 tensors (outer-products of norm 1). The latter is usually referred to as a *CP decomposition* [15, 37]. Both classes of tensor decompositions have been used for lower rank tensor approximation. However, neither of these classes provide optimal low rank approximations as in the matrix case. One can therefore only reach the conclusion that the algebraic and geometric properties of matrices and tensors of order $N > 2$ are highly dissimilar.

The purpose of this chapter is to develop a notion of singular decompositions for tensors (TSVD's) and to study its implication for the problem of finding (optimal) low rank approximations of tensors. We will do this by introducing a decomposition that combines a choice of orthonormal bases in the domain of the tensor with a suitable truncation of its expansion. In addition, we aim to develop suitable computational algorithms for the calculation of such decompositions and prove their stability and convergence properties.

The focus on the topic of singular value decompositions for optimal rank approximation problems is most natural for a number of reasons. Firstly, the SVD provides a useful way to numerically implement the algebraic concept of rank of matrices. It is doing this by quantifying near rank deficiencies or distances to lower rank approximations [33]. Secondly, singular vectors define orthonormal bases of both the domain and codomain of a linear map in such a way that the matrix representation of this mapping is maximally sparse with respect to these bases. Thirdly, singular values provide relevant information to analyse invertibility and the numerical conditioning of matrices and matrix operations. Fourthly, the SVD is well defined by performing successive rank-one approximations of a matrix.

A widely used generalization of the singular value decomposition to tensors was first introduced in [24] and is referred to as the higher order singular value decomposition (HOSVD). This decomposition involves the classical singular value decomposition of all possible matrix unfoldings of a tensor. In [24, 25] the authors propose an algorithm to construct the HOSVD and derive lower rank approximations by restricting the domain of the tensor to subspaces spanned by the first few left singular vectors of all possible matrix unfoldings. This procedure is easy to compute and implement, but the resulting low order tensors do not optimally approximate T. An upper bound on the approximation error is derived in [24]. Although the basic idea behind tensor unfoldings is interesting, at a more fundamental level it involves replacing the multi-linear structure of a tensor by multiple bi-linear structures and, therefore, hides the intrinsic multi-linear and algebraic properties of a tensor.

Especially in the signal processing community, tensors are commonly viewed as multi-linear arrays and tensor operations are carried out with regular matrix manipulations. Although useful for many applications in signal processing such as [22, 76, 77], this point of view has serious shortcomings when studying tensors at a more funda-

mental algebraic level. Since changes of coordinate systems are among the most elementary algebraic operations, we believe that it is particularly important to understand tensors as general multi-linear functionals in a coordinate-free algebraic context. Therefore, a discussion on coordinate-free concepts such as inner products, orthogonality, contractions, modal ranks and norms of tensors precedes the definition of a singular value decomposition and aims to provide insight in the true and more subtle nature of tensors as operators.
This chapter is organized as follows. In Section 3.2 tensors are formally introduced. We discuss tensor norms and inner products. Section 3.3 introduces the field of tensor decompositions and gives a short overview of the current state of the field. Section 3.4 then formally defines tensor rank and several decompositions related to these rank concepts. It also gives a formal problem statement for the problem of low-rank approximation. Section 3.5 defines a new modal rank decomposition method. This decomposition method will be referred to as TSVD and properties of the decomposition are derived. Section 3.6 gives the low-rank approximation results that can be achieved with this method. Section 3.7 proposes an adaptation of the TSVD that may yield better approximation results when not all modal directions are approximated. Section 3.8 presents a numerical algorithm that can be used to compute the TSVD. This chapter is concluded with a numerical example in Section 3.9 and a number of conclusions.

3.2 Tensors

An *order-N tensor* is a multi-linear functional

$$W : \mathcal{X}_1 \times \ldots \times \mathcal{X}_N \to \mathbb{R}$$

defined on vector spaces $\mathcal{X}_1, \ldots, \mathcal{X}_N$ that are assumed to be finite-dimensional. That is, W is a linear functional in each of its N arguments. Elements of W are specified by real numbers $w_{\ell_1 \ldots \ell_N}$ where ℓ_k ranges from 1 till the dimension L_k of $\mathcal{X}_k$, and k ranges from 1 till N. Elements of W are commonly encoded in the N-way array $[[w_{\ell_1 \ldots \ell_N}]] \in \mathbb{R}^{L_1 \times \cdots \times L_N}$ which, especially in signal processing, is taken as a (coordinate-dependent) definition of a tensor [24],[47]. We will refer to the k-th argument of W as the *k-th mode* of the tensor and to L_k as the k-th mode dimension. The elements $w_{\ell_1 \ldots \ell_N}$ represent W with respect to a specific collection of bases

$$\{f_1^{\ell_1},\ \ell_1 = 1, \ldots, L_1\}, \ldots,\ \{f_N^{\ell_N},\ \ell_N = 1, \ldots, L_N\} \tag{3.1}$$

of $\mathcal{X}_1, \ldots, \mathcal{X}_N$, respectively, in the sense that $w_{\ell_1 \ldots \ell_N} = \frac{W(f_1^{\ell_1}, \ldots, f_N^{\ell_N})}{\|f_1^{(\ell_1)}\|^2 \cdots \|f_N^{(\ell_N)}\|^2}$.

Throughout, the set of all order-N tensors on $\mathcal{X}_1 \times \ldots \times \mathcal{X}_N$ is denoted by $\mathcal{T}_N$ which becomes a vector space over the field $\mathbb{R}$ when equipped with the standard definitions of addition and scalar multiplication. Precisely, given tensors $V, W \in \mathcal{T}_N$ and a scalar $\alpha \in \mathbb{R}$, we have the following definitions.

1. Addition: $U := V + W$ is the order-N tensor

$$U(x_1, \ldots, x_N) := V(x_1, \ldots, x_N) + W(x_1, \ldots, x_N) \tag{3.2}$$

   for all $x_k \in \mathcal{X}_k$. If V and W are represented with respect to the same sets of basis functions, with coefficients $v_{\ell_1 \ldots \ell_N}$ and $w_{\ell_1 \ldots \ell_N}$, the coefficients of U are given by $u_{\ell_1 \ldots \ell_N} := v_{\ell_1 \ldots \ell_N} + w_{\ell_1 \ldots \ell_N}$.

2. Scalar multiplication: For any $\alpha \in \mathbb{R}$, $U := \alpha W$ is the tensor

$$U(x_1, \ldots, x_N) = \alpha W(x_1, \ldots, x_N) \tag{3.3}$$

   with $x_k \in \mathcal{X}_k$, $k = 1, \ldots, N$. If the coefficients of W are $w_{\ell_1 \ldots \ell_N}$, the coefficients of U are given by $u_{\ell_1 \ldots \ell_N} := \alpha w_{\ell_1 \ldots \ell_N}$.

To define approximations to tensors we will need a *norm* on the space $\mathcal{T}_N$. For this let $\|\cdot\|_k$ denote the induced norm corresponding to the inner product $\langle \cdot, \cdot \rangle_k$ of $\mathcal{X}_k$, i.e. $\|x\|_k = \sqrt{\langle x, x \rangle_k}$. We assume this structure for $k = 1, \ldots, N$. The *inner product* of two tensors $S, T \in \mathcal{T}_N$ with elements $s_{k_1 \ldots k_N}$ and $t_{\ell_1 \ldots \ell_N}$, both defined with respect to the bases (3.1), is given by

$$\langle S, T \rangle := \sum_{k_1} \cdots \sum_{k_N} \sum_{\ell_1} \cdots \sum_{\ell_N} s_{k_1 \ldots k_N} t_{\ell_1 \ldots \ell_N} \langle f_1^{k_1}, f_1^{\ell_1} \rangle \cdots \langle f_N^{k_N}, f_N^{\ell_N} \rangle.$$

It is immediate that the right-hand side of this expression is invariant under unitary basis transformations (i.e., transformations $Q_k : \mathcal{X}_k \to \mathcal{X}_k$ for which $\|Q_k x\|_k = \|x\|_k$ for all $x \in \mathcal{X}_k$) and so $\mathcal{T}_N$ becomes a well defined inner product space. The Frobenius norm of a tensor $W \in \mathcal{T}_N$ is then defined as

$$\|W\|_F := \sqrt{\langle W, W \rangle}. \tag{3.4}$$

It is easily seen that if W is represented by $[[w_{\ell_1 \ldots \ell_N}]]$ with (3.1) orthonormal bases, then $\|W\|_F^2 = \sum_{\ell_1, \ldots, \ell_N} w_{\ell_1 \ldots \ell_N}^2$.
One may also consider the *operator norm* of $W \in \mathcal{T}_N$ defined by

$$\|W\| := \max_{\substack{x_k \in \mathcal{X}_k,\ \|x_k\|_k = 1 \\ k = 1, \ldots, N}} |W(x_1, \ldots, x_N)|.$$

That is, $\|W\|$ reflects the maximal amplitude that a tensor can assume when ranging over the Cartesian product of all unit spheres in $\mathcal{X}_k$, $k = 1, \ldots, N$. This norm satisfies the properties $\|W\| \geq 0$, $\|W\| = 0$ only if $W = 0$, $\|\alpha W\| = |\alpha|\|W\|$ for any scalar $\alpha \in \mathbb{R}$ and $\|W + S\| \leq \|W\| + \|S\|$ for any $S, W \in \mathcal{T}_N$. Therefore, $\mathcal{T}_N$ becomes a normed linear space when equipped with the operator norm $\|\cdot\|$.
For fixed elements $u_k \in \mathcal{X}_k$, $k = 1, \ldots, N$, the functional

$$U(x_1, \ldots, x_N) := \langle u_1, x_1 \rangle_1 \cdots \langle u_N, x_N \rangle_N = \Pi_{n=1}^N \langle u_n, x_n \rangle_n$$

defines an order-N tensor which will be denoted by $U = u_1 \otimes \cdots \otimes u_N$. Whenever non-zero, such a tensor will be referred to as a *rank-1 tensor*. With respect to the bases (3.1), the elements of U are $u_{\ell_1 \ldots \ell_N} = u_1^{\ell_1} \cdots u_N^{\ell_N}$ where $u_k^{\ell_k} = \langle u_k, f_k^{\ell_k} \rangle_k$ is the coefficient of u_k with respect to the basis vector $f_k^{\ell_k}$. We have that $\|U\| = \Pi_{k=1}^N \|u_k\|_k$. Every tensor can be represented as a weighted sum of rank-one tensors as follows

$$W = \sum_{\ell_1=1}^{\dim(\mathcal{X}_1)} \cdots \sum_{\ell_N=1}^{\dim(\mathcal{X}_N)} w_{\ell_1 \ldots \ell_N} f_1^{(\ell_1)} \otimes \cdots \otimes f_N^{(\ell_N)} \tag{3.5}$$

We distinguish between different types of orthogonality (cf. [46, 54]) regarding tensors. These distinct orthogonality concepts lead to different types of tensor decompositions, as will be shown later in this chapter.

Definition 3.2.1. *Let $U = u_1 \otimes \cdots \otimes u_N$ and $V = v_1 \otimes \cdots \otimes v_N$ be two rank-1 tensors.*

1. *U and V are said to be* orthogonal, *denoted $U \perp V$, if $\langle U, V \rangle = \prod_{k=1}^N \langle u_k, v_k \rangle_k = 0$.*

2. *They are said to be* completely orthogonal, *denoted $U \perp_c V$, if $\langle u_k, v_k \rangle_k = 0$ for all $k = 1, \ldots, N$.*

3.2.1 Some additional tensor concepts

A linear mapping $G : \mathcal{T}_N \to \mathcal{T}_M$ is defined as $B := G(A)$ where $B \in \mathcal{T}_M$ is the tensor

$$B = \sum_{m_1} \ldots \sum_{m_M} b_{m_1, \ldots, m_M} \hat{e}_1^{m_1} \otimes \ldots \otimes \hat{e}_M^{m_M}$$

obtained from the coefficients $a_{\ell_1, \ldots, \ell_N}$ of A by

$$b_{m_1, \ldots, m_M} = \sum_{\ell_1} \ldots \sum_{\ell_N} g_{\ell_1, \ldots, \ell_N, m_1, \ldots, m_M} a_{\ell_1, \ldots, \ell_N}. \tag{3.6}$$

for some collection of coefficients $g_{\ell_1,\ldots,\ell_N,m_1,\ldots,m_M}$. Evidently, G is entirely defined by the constants $g_{\ell_1,\ldots,\ell_N,m_1,\ldots,m_M}$. In particular, we associate a *multiplication tensor* $T_G \in \mathcal{T}_{N+M}$ with G by setting

$$T_G := \sum_{\ell_1} \cdots \sum_{\ell_N} \sum_{m_1} \cdots \sum_{m_M} g_{\ell_1,\ldots,\ell_N,m_1,\ldots,m_M} e_1^{\ell_1} \otimes \ldots \otimes e_N^{\ell_N} \otimes \hat{e}_1^{m_1} \otimes \ldots \otimes \hat{e}_M^{m_M}. \tag{3.7}$$

It is immediate that any such G is linear in the sense that, for $A, B \in \mathcal{T}_N$ and $\alpha, \beta \in \mathbb{R}$, we have that $G(\alpha A + \beta B) = \alpha G(A) + \beta G(B)$.
Eigenvalues and eigentensors of a linear map $G : \mathcal{T}_N \to \mathcal{T}_N$ are defined as follows:

Definition 3.2.2 (Eigenvalues and Eigentensors)**.** *A nonzero tensor $A \in \mathcal{T}_N$ is an* eigentensor *of the linear map $G : \mathcal{T}_N \to \mathcal{T}_N$ with corresponding* eigenvalue *$\lambda \in \mathbb{R}$ if $GA = \lambda A$.*

The concept of positive definiteness for matrices, as discussed here [43], can easily be extended to mappings between tensors.

Definition 3.2.3 (Positive definite operator)**.** *A linear mapping $G : \mathcal{T}_N \to \mathcal{T}_N$ is* positive definite *if for any $0 \neq A \in \mathcal{T}_N$ there holds $\langle A, GA \rangle > 0$.*

Positive definite mappings between tensors have real eigenvalues:

Theorem 3.2.4. *If $G : \mathcal{T}_N \to \mathcal{T}_N$ is linear and positive definite then all its eigenvalues are positive.*

Proof. For any non-zero eigentensor $A_i \in \mathcal{T}_N$ of G with corresponding eigenvalue λ_i, we have that $GA_i = \lambda_i A_i$. Because G is positive definite we have

$$0 < \langle A_i, GA_i \rangle = \langle A_i, GA_i \rangle = \langle A_i, \lambda_i A_i \rangle = \lambda_i \langle A_i, A_i \rangle = \lambda_i \|A_i\|^2.$$

Since $\|A_i\| \neq 0$, we must have that $\lambda_i > 0$ for $i = 1, \ldots, N$. □

The next section will introduce tensor decompositions. For the remainder of this chapter it is assumed that the reader is familiar with matrix concepts such as rank, the Singular Value Decomposition and optimal rank approximations to matrices. These concepts are introduced in Appendix A.3, which can be referred to if necessary.

3.3 Introduction to Tensor Decompositions

Tensor Decompositions is a discipline of research that strives to develop tools that allow analysis and approximation of tensors. In this section, we will provide a general introduction to this field.
To explain what is meant by *tensor decompositions*, consider a tensor $W \in \mathcal{T}_N$. W operates on a collection of vector spaces, i.e. $W : \mathcal{X}_1 \times \cdots \times \mathcal{X}_N \to \mathbb{R}$ and is defined with respect to bases $\{f_k^{(\ell_k)}\}_{k=1}^{\dim(\mathcal{X}_k)}$, $k = 1, \ldots, N$. A decomposition of W is implied or defined by a basis change, such that the representation of W with respect to these new basis functions satisfies certain properties. Specifically, we look for sets of basis functions $\{\varphi_k^{(\ell_k)}\}_{\ell_k=1}^{L_k}$, $k = 1, \ldots, N$, such that the representation of W with respect to the newly defined basis functions, i.e.

$$W = \sum_{\ell_1=1}^{\dim(\mathcal{X}_1)} \cdots \sum_{\ell_N=1}^{\dim(\mathcal{X}_N)} w_{\ell_1 \cdots \ell_N} \varphi_1^{(\ell_1)} \otimes \cdots \otimes \varphi_N^{(\ell_N)} \tag{3.8}$$

satisfies certain properties.
Desirable properties of tensor decompositions could be the following

1. *Diagonality*, the core of the representation of W in (3.8) is diagonal, i.e.

$$w_{\ell_1 \cdots \ell_N} = 0, \quad \text{unless } \ell_1 = \cdots = \ell_N$$

2. *Orthonormality*, $\{\varphi_k^{(\ell_k)}\}_{k=1}^{\dim(\mathcal{X}_k)}$ is an orthogonal (orthonormal) basis for $\mathcal{X}_k$, $k = 1, \ldots, N$.

3. *Low Approximation Error*, the decomposition (3.8) of W may be used to construct approximations of W. Specifically, truncations in the summations of (3.8) result in small approximation errors between W and the truncated expansion.

To explain the last item, we will now show how a tensor decomposition (3.8) can be used to construct approximations to tensors. The way we define these approximations closely resembles the concepts discussed in Sec. 2.3 where we introduced approximations via projection of systems.
Consider a tensor W and its decomposition (3.8) and define an approximation degree r which is an N-dimensional vector of integers $r = (r_1, \ldots, r_N)$. As in (3.14) we define the following subspaces

$$\mathcal{M}_k^{(r_k)} = \operatorname{span}\{\varphi_k^{(1)}, \ldots, \varphi_k^{(r_k)}\}, \quad k = 1, \ldots, N.$$

An approximation $W_r \in \mathcal{T}_N$ of W is now defined as the restriction

$$W_r := W \mid_{\mathcal{M}_1^{(r_1)} \times \cdots \times \mathcal{M}_N^{(r_N)}} . \tag{3.9}$$

W_r has the following representation

$$W_r = \sum_{\ell_1=1}^{r_1} \cdots \sum_{\ell_N=1}^{r_N} w_{\ell_1 \cdots \ell_N} \varphi_1^{(\ell_1)} \otimes \cdots \otimes \varphi_N^{(\ell_N)}.$$

The approximation degree r is a measure for the complexity of the approximant W_r, since it indicates how many coefficients and basis elements must be stored to represent W_r.
Typically, it is not possible to construct tensor decompositions that satisfy all three properties mentioned. We will now discuss two types of tensor decompositions. The first generalizes the property of diagonality to tensor decompositions, the second type generalizes the property of orthogonality.
The first type of tensor decompositions is called the *Canonical Polyadic-*, or CP-decomposition. It was first defined in 1927 by Hitchcock [38] and became better-known when it was defined again in 1970 by Caroll and Chang and Richard Harshman. In this decomposition, the core $[[w_{\ell_1 \cdots \ell_N}]]$ of (3.8) is required to be diagonal, i.e. $w_{\ell_1 \cdots \ell_N} = 0$, unless $\ell_1 = \cdots = \ell_N$. Then, (3.8) is equivalent to

$$\begin{aligned} W &= \sum_{\ell=1}^{R} w_\ell \varphi_1^{(\ell)} \otimes \cdots \otimes \varphi_N^{(\ell)} \\ &= \sum_{\ell=1}^{R} w_\ell U_{(\ell)} \end{aligned} \tag{3.10}$$

The second type of tensor decompositions that will be examined is the *Tucker* decomposition [75]. This decomposition may be used to generalize the orthogonality property, though orthogonality is not strictly required. In the Tucker decomposition a tensor is represented as follows

$$W = \sum_{\ell_1=1}^{\dim(\mathcal{X}_1)} \cdots \sum_{\ell_N=1}^{\dim(\mathcal{X}_N)} w_{\ell_1 \cdots \ell_N} \varphi_1^{(\ell_1)} \otimes \cdots \otimes \varphi_N^{(\ell_N)}. \tag{3.11}$$

where the basis functions $\{\varphi_k^{(\ell_k)}\}$ may be orthogonal (orthonormal) sets.
Apart from these two general types, other types of tensor decompositions exist. These methods include *Tree-Tucker* decompositions, in which a tensor is decomposed into a

tree of order-3 tensors, [60, 62], and block-decompositions [53]. For more information on these and other tensor decompositions and the algorithms that may be used to compute these decompositions, we refer to [47].
Tensor decompositions originated in the field of psychometrics, where they were used for analysis purposes. Their applications have since been expanded to chemometrics, signal processing, numerical linear algebra and many more, see [47] and the references therein for an overview. An application area that receives a lot of attention lately is that of using tensors to reduce computation time of multidimensional functions on discretized grids. In [11] the authors introduce the concepts of separated rank and separated representations to accelerate computations of multidimensional functionals on discretized grids. It is their aim to arrive at function approximations, rather than tensor decompositions. Therefore, in the construction of the separated representation, the authors do not require minimality of rank, nor orthonormality of the decomposition in whatever sense, nor optimality of the approximation. Since it is ultimately the aim of this work to construct empirical projection spaces spanned by orthonormal bases, the work of [11] is not considered further here. In [45] the authors attempt to combine the strengths of the Tucker and CP decomposition to decrease the time involved in computations with function related multidimensional arrays. Numerical algorithms for tensor decompositions and approximations of Tucker type are presented, with the additional constraint that the core array is to be represented in a low-rank canonical format. As an application of tensors to accelerate computation of multidimensional functions on discretized grids, [36] uses tensors to solve elliptic eigenvalue problems.

3.4 Tensor Decompositions

The aim of this section is to make the tensor decompositions introduced in the previous section more specific. We will start with a discussion of the rank concepts that can be defined for tensors. Different rank concepts lead to different tensor decompositions. We will introduce modal rank decompositions of tensors, which are a subclass of the Tucker decomposition defined in (3.11). This section concludes with a formal problem statement for the remainder of this chapter.

3.4.1 Tensor rank and related decompositions

The concept of *tensor rank* is a highly non-trivial extension of the same concept for linear mappings and has been discussed in considerable detail in, for example, [26, 46, 48, 24, 25, 55]. As with orthogonality in the previous subsections, the different concepts of tensor rank lead to different types of tensor decompositions. The

rank of $W \in \mathcal{T}_N$, denoted $\mathrm{rank}(W)$, is the minimum integer R such that W can be decomposed as in (3.10). By definition, the rank of the zero tensor is 0. [46] also introduces the concepts of orthogonal and complete orthogonal rank. The *orthogonal rank* and *complete orthogonal rank* of a tensor W is the minimal integer R in decomposition (3.10) with the additional requirements that $w_\ell > 0$, $\|U_\ell\| = 1$ and $U_i \perp U_j$ (or $U_i \perp_c U_j$) for $1 \leq i, j \leq R$.
To define the *modal rank* of a tensor $W \in \mathcal{T}_N$, we first introduce the *k-mode kernel* of W to be the set

$$\mathrm{ker}_k(W) := \{x_k \in \mathcal{X}_k \mid W(x_1, \ldots, x_N) = 0,\ \forall x_p \in \mathcal{X}_p,\ p \neq k\}.$$

The multi-linearity of W implies that $\mathrm{ker}_k(W)$ is a linear subspace of $\mathcal{X}_k$. The *k-mode rank* of W, is defined by

$$R_k = \mathrm{rank}_k(W) := \dim(\mathcal{X}_k) - \dim(\mathrm{ker}_k(W)), \quad k = 1, \ldots, N,$$

and is coordinate free generalization of the k-rank in [24]. Note that $\mathrm{rank}_k(W)$ coincides with the dimension of the space spanned by stringing out all elements $w_{\ell_1,\ldots,1,\ldots\ell_N}$ till $w_{\ell_1,\ldots,N,\ldots\ell_N}$ (where the indices $1, \ldots, N$ are at the kth spot). Finally, the *modal rank* of W, denoted $\mathrm{modrank}(W)$, is the vector of all k-mode ranks, i.e., $\mathrm{modrank}(W) = (R_1, \ldots, R_N)$, $R_k = \mathrm{rank}_k(W)$. The modal rank is also referred to as *multi-linear* rank [42].
The rank and modal rank are well defined in that there exist unique numbers $R = \mathrm{rank}(W)$ and $R_k = \mathrm{rank}_k(W)$ for any $W \in \mathcal{T}_N$. Obviously, $\mathrm{rank}(U) = \mathrm{rank}_k(U) = 1$ for a rank-1 tensor U. For $W \in \mathcal{T}_N$ we have that $\mathrm{rank}_k(W) \leq \mathrm{rank}(W)$ and there exist examples with strict inequality for all k [24, 25]. For order-2 tensors (matrices) we have that $R = R_1 = R_2 = \mathrm{rank}(W)$ and the rank concept coincides with the usual notion of rank, row-rank or column-rank, of a matrix. The next example shows that the modal ranks of a tensor need not be the same.

Example 3.4.1. *This example is taken from [24]. Consider the tensor $W : \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3 \to \mathbb{R}$, where $\mathcal{X}_k = \mathbb{R}^2$, $k = 1, 2, 3$. The representation of W with respect to the standard bases is*

$$w_{111} = w_{221} = w_{112} = 1$$
$$w_{211} = w_{121} = w_{212} = w_{122} = w_{222} = 0.$$

The modal rank of W is given by $\mathrm{modrank}(W) = (2, 2, 1)$.

For $W \in \mathcal{T}_N$ of modal rank $\operatorname{modrank}(W) = (R_1, \ldots, R_N)$ the expression

$$W = \sum_{\ell_1=1}^{R_1} \cdots \sum_{\ell_N=1}^{R_N} w_{\ell_1 \ldots \ell_N} \varphi_1^{(\ell_1)} \otimes \cdots \otimes \varphi_N^{(\ell_N)} \tag{3.12}$$

is called a *modal rank decomposition* of W whenever

$$\langle \varphi_k^{(i)}, \varphi_k^{(j)} \rangle_k = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}, \quad \text{for all } 1 \leq i, j \leq R_k,\ k = 1, \ldots, N.$$

A modal rank decomposition is therefore a representation of W with respect to orthonormal bases

$$\{\varphi_1^{(\ell_1)}\}_{\ell_1=1}^{L_1}, \ldots, \{\varphi_N^{(\ell_N)}\}_{\ell_N=1}^{L_N} \tag{3.13}$$

of $\mathcal{X}_1 = \mathbb{R}^{L_1}, \ldots, \mathcal{X}_N = \mathbb{R}^{L_N}$, respectively. The modal rank decomposition is a higher-order extension of the Tucker decomposition introduced in [75] with additional orthogonality constraints.

Among the different notions of tensor rank that we define here, only the modal rank can actually be computed for arbitrary order-N tensors. The other rank concepts can only be determined for small academic examples such as $2 \times 2 \times 2$-tensors.

3.4.2 Modal Rank Approximations to Tensors

Since we view tensors as multi-linear functionals, rather than multi-dimensional arrays, we only consider decompositions that can be regarded as basis transformation of the multi-linear functional. Hence, we focus on modal rank decompositions of the form (3.12) and do not take decompositions of the form (3.10) into account. For order-2 tensors, matrices, modal rank decompositions are equal to the rank decompositions defined in (3.10) and can be computed via the Singular Value Decomposition (SVD), see Appendix A.3. Truncation of the SVD of a matrix yields optimal rank approximations of these matrices, as is discussed in Appendix A.3.

For higher-order tensors the situation is less straightforward. Several methods have been proposed to compute modal rank decompositions for tensors of order $N > 2$. Each of these methods generalizes different properties of the matrix SVD.

Consider a tensor $W \in \mathcal{T}_N$ of modal rank $(R_1, \ldots, R_N)$. We will now demonstrate how an approximation of this tensor can be computed from its modal rank decomposition. Using one of the modal rank decomposition algorithms, orthonormal basis functions $\{\varphi_k^{(\ell_k)}\}_{\ell_k=1}^{L_k}$ of $\mathcal{X}_k$ are computed as in (3.13). These basis functions define the modal rank decomposition, see (3.12). An approximation of W of degree $r = (r_1, \ldots, r_N)$ can be defined as follows. Define the subspaces

$$\mathcal{M}_k^{(r_k)} = \operatorname{span}\{\varphi_k^{(1)}, \ldots, \varphi_k^{(r_k)}\}, \quad k = 1, \ldots, N \tag{3.14}$$

Definition 3.4.2. *Given a tensor $W \in \mathcal{T}_N$. For a vector of integers $r = (r_1, \ldots, r_N)$, $r_k \leq R_k,\ k = 1, \ldots, N$, the modal truncation W_r is defined by the restriction $W_r := W\,|_{\mathcal{M}_1^{(r_1)} \times \cdots \times \mathcal{M}_N^{(r_N)}}$ and is represented by the expansion*

$$W_r = \sum_{\ell_1=1}^{r_1} \cdots \sum_{\ell_N=1}^{r_N} w_{\ell_1 \cdots \ell_N} \varphi_1^{(\ell_1)} \otimes \cdots \otimes \varphi_N^{(\ell_N)} \tag{3.15}$$

where $w_{\ell_1 \cdots \ell_N} := W(\varphi_1^{(\ell_1)}, \ldots, \varphi_N^{(\ell_N)})$.

3.4.3 Problem formulation

The problem of finding lower modal rank approximations of a given tensor is the prime motivation for the remainder of this chapter. A precise formulation is given as follows.

Problem 3.4.3. *Let $W \in \mathcal{T}_N$ be a given N-order tensor.*

P1: Given a vector of integers $r = (r_1, \ldots, r_N)$, $r_n \leq \operatorname{rank}_n(W)$, determine $\inf_{\operatorname{modrank}(W_r)=r} \|W - W_r\|$ and find, if possible, a tensor $W_r \in \mathcal{T}_N$ with $\operatorname{modrank}(W_r) = r$ such that $\|W - W_r\|$ is minimal.

P2: Given a vector of integers $r = (r_1, \ldots, r_N)$, $r_n \leq \operatorname{rank}_n(W)$, determine $\inf_{\operatorname{modrank}(W_r)=r} \|W - W_r\|_F$ and find, if possible, a tensor $W_r \in \mathcal{T}_N$ with $\operatorname{modrank}(W_r) = r$ such that $\|W - W_r\|_F$ is minimal.

P3: Given an integer $r \leq \operatorname{rank} W$, determine $\inf_{\operatorname{rank}(W_r)=r} \|W - W_r\|$ and find, if possible, a tensor $W_r \in \mathcal{T}_N$ of rank $\operatorname{rank}(W_r) = r$ such that $\|W - W_r\|$ is minimal.

P4: Given an integer $r \leq \operatorname{rank} W$, determine $\inf_{\operatorname{rank}(W_r)=r} \|W - W_r\|_F$ and find, if possible, a tensor $W_r \in \mathcal{W}_N$ of rank $\operatorname{rank}(W_r) = r$ such that $\|W - W_r\|_F$ is minimal.

For $N > 2$, Problem P4 has been studied in [46][48][55][26] by introducing orthogonal rank-1 tensor decompositions. It was found that the minimum rank r approximation problem is ill-posed in that optimal lower rank approximations do not need to exist. In [55] an example is given of a rank 6 tensor W for which $\inf_{\operatorname{rank}(W_2)=2} \|W - W_2\|_{\mathrm{F}} = 0$, showing that the space of lower rank tensors is not closed. For further discussions on Problem P4 we refer to [26, 18]. In this work we focus on the problems P1 and P2.

3.5 TSVD

This section presents a new method to compute modal rank approximations to tensors. We will first give the definitions of the method, then its properties will be discussed. The sections that follow this one will state low-rank approximation properties and give a numerical algorithm to compute the decomposition. The work presented here was published in [81].

Let $W \in \mathcal{T}_N$ be an order-N tensor defined on the finite dimensional vector spaces $\mathcal{X}_1, \ldots, \mathcal{X}_N$ where we suppose that $\dim(\mathcal{X}_k) = L_k$. The *singular values* of W, denoted $\sigma_m(W)$, with $m = 1, \ldots, K$ and $K = \min_k \operatorname{modrank}(W)$ are defined as follows.

For $k = 1, \ldots, N$ let

$$\mathcal{S}_k^{(1)} := \{x \in \mathcal{X}_k \mid \|x\|_k = 1\}$$

denote the unit sphere in $\mathcal{X}_k$. Define the first singular value of W by

$$\sigma_1(W) := \sup_{\substack{x_k \in \mathcal{S}_k^{(1)}, \\ 1 \le k \le N}} |W(x_1, \ldots, x_N)|. \tag{3.16}$$

Since W is continuous and the Cartesian product $\mathcal{S}^{(1)} := \mathcal{S}_1^{(1)} \times \cdots \times \mathcal{S}_N^{(1)}$ of unit spheres is a compact set, an extremal solution of (3.16) exists (i.e., the supremum in (3.16) is a maximum) and is attained by an N-tuple

$$(x_1^{(1)}, \ldots, x_N^{(1)}) \in \mathcal{S}^{(1)}.$$

Subsequent singular values of W are defined in an inductive manner by setting

$$\mathcal{S}_k^{(m)} := \{x \in \mathcal{X}_k \mid \|x\|_k = 1, \langle x, x_k^{(j)} \rangle_k = 0 \text{ for } j = 1, \ldots, (m-1)\} \tag{3.17}$$

for $k = 1, \ldots, N$, and by defining

$$\sigma_m(W) = \sup_{\substack{x_k \in \mathcal{S}_k^{(m)}, \\ 1 \le k \le N}} |W(x_1, \ldots, x_N)|, \qquad m \le K. \tag{3.18}$$

Again, since the Cartesian product

$$\mathcal{S}^{(m)} := \mathcal{S}_1^{(m)} \times \cdots \times \mathcal{S}_N^{(m)}$$

is compact, the supremum in (3.18) is a maximum that is attained by an N-tuple

$$(x_1^{(m)}, \ldots, x_N^{(m)}) \in \mathcal{S}^{(m)}.$$

It follows that the vectors $x_k^{(1)}, \ldots, x_k^{(K)}$ are mutually orthonormal in $\mathcal{X}_k$. If $K < L_k$ for any k, then we extend the collection of orthogonal elements $x_k^{(1)}, \ldots, x_k^{(K)}$ to a complete orthonormal basis of $\mathcal{X}_k$. This construction thus leads to a collection of orthonormal bases

$$\{x_1^{(\ell_1)}, \ell_1 = 1, \ldots, L_1\}, \ldots, \{x_N^{(\ell_N)}, \ell_N = 1, \ldots, L_N\} \tag{3.19}$$

for the vector spaces $\mathcal{X}_1$, …, $\mathcal{X}_N$, respectively.

Definition 3.5.1. *The* singular values *of an order-N tensor $W \in \mathcal{T}_N$ are the numbers $\sigma_1, \ldots, \sigma_K$ with $K = \min_k \operatorname{modrank}_k(W)$ defined by* (3.16) *and* (3.18)*. The* singular vectors of order m *are the extremal solutions $(x_1^{(m)}, \ldots, x_N^{(m)})$ in $\mathcal{S}^{(m)}$ that attain the maximum in* (3.18)*. A* singular value decomposition *(SVD) of the tensor W is a representation of W with respect to the basis* (3.19)*, i.e.,*

$$W = \sum_{\ell_1=1}^{L_1} \cdots \sum_{\ell_N=1}^{L_N} w_{\ell_1 \ldots \ell_N} x_1^{(\ell_1)} \otimes \cdots \otimes x_N^{(\ell_N)}. \tag{3.20}$$

The N-way array $[[w_{\ell_1,\ldots,\ell_N}]] \in \mathbb{R}^{L_1 \times \cdots \times L_N}$ in (3.20) *is called the* singular value core *of T.*

3.5.1 Characterization of singular vectors by duality

This section aims to characterize the singular values and singular vectors of tensors of any order. The idea of viewing singular values as defined in Definition 3.5.1 originates from [34]. Here, the duality properties have been extended to order $N > 3$. Let $L = L_1 + \cdots + L_N$ and associate with the optimization problem (3.16) the Lagrangian $L_1 : \mathbb{R}^{L+N} \to \mathbb{R}$ by setting

$$L_1(x, \lambda) := W(x_1, \ldots, x_N) + \sum_{k=1}^{N} \frac{1}{2} \lambda_k (1 - \langle x_k, x_k \rangle_k).$$

It has already been argued that an N-tuple $x^{(1)} = (x_1^{(1)}, \ldots, x_N^{(1)})$ exists that attains the maximum in (3.16). From the theory of variational analysis [10, 31], one then infers the existence of an N-tuple $\lambda^{(1)} = (\lambda_1^{(1)}, \ldots, \lambda_N^{(1)})$ of Lagrange multipliers such that

$$\nabla L_1(x^{(1)}, \lambda^{(1)}) = 0, \tag{3.21}$$

where ∇L_1 denotes the gradient of L_1. The k-mode Fréchet derivative $\partial_k W(x_1, \ldots, x_N)$ of W at the point $(x_1, \ldots, x_N)$ is an order-1 tensor (a linear functional) that maps $\mathcal{X}_k$ to $\mathbb{R}$ and satisfies

$$\partial_k W(x_1, \ldots, x_N) = W(x_1, \ldots, x_{k-1}, \cdot, x_{k+1}, \ldots, x_N)$$

where the 'dot' is at the kth spot. By the multi-linearity of the tensor, $\partial_k W(x_1, \ldots, x_N)$ is independent of $x_k \in \mathcal{X}_k$. Hence, rewriting (3.36) for each independent modal direction gives that $x^{(1)}, \lambda^{(1)}$ satisfies, for $k = 1, \ldots, N$,

$$W(x_1^{(1)}, \ldots, x_{k-1}^{(1)}, \cdot, x_{k+1}^{(1)}, \ldots, x_N^{(1)}) = \lambda_k^{(1)} \langle \cdot, x_k^{(1)} \rangle, \tag{3.22a}$$

$$\|x_k^{(1)}\|_k = 1. \tag{3.22b}$$

It follows that $W(x_1^{(1)}, \ldots, x_N^{(1)}) = \lambda_1^{(1)} = \cdots = \lambda_N^{(1)} = \sigma_1$, i.e., all Lagrange multipliers coincide. Moreover, (3.37a) implies that for each $k = 1, \ldots, N$,

$$W(x_1^{(1)}, \ldots, x_{k-1}^{(1)}, \xi_k, x_{k+1}^{(1)}, \ldots, x_N^{(1)}) = 0 \quad \text{whenever } \langle \xi_k, x_k^{(1)} \rangle_k = 0.$$

In a similar manner, for $m > 1$ we associate with the optimization problem (3.18) the Lagrangian $L_m : \mathbb{R}^{L+N+N(m-1)} \to \mathbb{R}$ defined by

$$L_m(x, \lambda, \mu) = W(x_1, \ldots, x_N) + \sum_{k=1}^{N} \frac{1}{2} \lambda_k (1 - \langle x_k, x_k \rangle_k) + \sum_{k=1}^{N} \langle g_k(x_k), \mu_k \rangle.$$

where $x_k \in \mathcal{X}_k$, $\lambda_k \in \mathbb{R}$, $\mu_k \in \mathbb{R}^{m-1}$ and $g_k : \mathcal{X}_k \to \mathbb{R}^{m-1}$ is given by

$$g_k(\xi_k) := \begin{pmatrix} \langle \xi_k, x_k^{(1)} \rangle_k \\ \vdots \\ \langle \xi_k, x_k^{(m-1)} \rangle_k \end{pmatrix}.$$

Again, there exist N-tuples $x^{(m)}$, $\lambda^{(m)}$ and $\mu^{(m)}$ that satisfy the stationarity condition

$$\nabla L_m(x^{(m)}, \lambda^{(m)}, \mu^{(m)}) = 0. \tag{3.23}$$

Rewriting (3.38) for each modal direction gives for $k = 1, \ldots, N$,

$$W(x_1^{(m)}, \ldots, x_{k-1}^{(m)}, \cdot, x_{k+1}^{(m)}, \ldots, x_N^{(m)}) = \lambda_k^{(m)} \langle \cdot, x_k^{(m)} \rangle + \langle g_k(\cdot), \mu_k^{(m)} \rangle, \tag{3.24a}$$

$$\|x_k^{(m)}\|_k = 1, \tag{3.24b}$$

$$g_k(x_k^{(m)}) = 0. \tag{3.24c}$$

This immediately implies that $W(x_1^{(m)}, \dots, x_N^{(m)}) = \lambda_1^{(m)} = \dots = \lambda_N^{(m)} = \sigma_m$ and we conclude again that, for fixed m, the Lagrange multipliers $\lambda_k^{(m)}$ coincide and are equal to the mth singular value. Moreover, for $k = 1, \dots, N$,

$$W(x_1^{(m)}, \dots, x_{k-1}^{(m)}, \xi_k, x_{k+1}^{(m)}, \dots, x_N^{(m)}) = \\ = \begin{cases} 0 & \text{whenever } \xi_k \perp \operatorname{span}(x_k^{(1)}, \dots, x_k^{(m)}) \\ \mu_{k,j}^{(m)} & \text{whenever } \xi_k = x_k^{(j)},\ j = 1, \dots, m-1 \end{cases} \tag{3.25}$$

where $\mu_{k,j}^{(m)}$ is the jth entry in the vector $\mu_k^{(m)}$.

3.5.2 TSVD properties

The following theorem summarizes a number of properties of the tensor singular value decomposition.

Theorem 3.5.2. *1. Every tensor $W \in \mathcal{T}_N$ admits a singular value decomposition. The singular value decomposition* (3.20) *is an orthogonal decomposition where the singular values are ordered according to $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_K > 0$. Here, $K = \min_k \operatorname{modrank}(T)$ and the singular vectors of any order m satisfy* (3.37) *and* (3.39)*.*

2. $W(x_1^{(m)}, \dots, w_N^{(m)}) = \sigma_m$.

3. For all $k = 1, \dots, N$ there holds

$$W(x_1^{(m)}, \dots, x_{k-1}^{(m)}, \xi, x_{k+1}^{(m)}, \dots, x_N^{(m)}) = 0$$

whenever $\xi \perp \operatorname{span}\{x_k^{(1)}, \dots, x_k^{(m)}\}$.

4. The singular value core of W satisfies

$$|w_{\ell_1 \cdots \ell_N}| = \begin{cases} \sigma_m & \text{if } \ell_1 = \dots = \ell_N = m \leq K \\ \leq \sigma_m & \text{if } m = \min\{\ell_1, \dots, \ell_N\} \\ 0 & \text{if } \ell_1 = \dots = \ell_N = m > K \\ 0 & \text{if } \ell_k > \ell_1 = \dots = \ell_{k-1} = \ell_{k+1} = \dots = \ell_N \end{cases}$$

5. *If $L_1 = L_2 = \cdots = L_N = L_0$ then the number of zeros in the singular value core of W is at least*
$$N\left(\frac{L_0(L_0-1)}{2}\right).$$

6. *In the general case, the number of zeros is at least equal to*
$$\sum_{k=1}^{N}\frac{L_k(L_k-1)}{2}.$$

Proof. The existence of the basis has been proven in the previous subsection. The ordering of the singular values and the fact that all rank-1 terms in the decomposition (3.20) are orthogonal is immediate from the definition (3.18). Item 2 follows from (3.37) and (3.39) and the observation that for fixed m, the Lagrange multipliers $\lambda_k^{(m)}$ coincide with σ_m (cf. subsection 3.5.1). Item 3 has been derived in (3.25). Since $w_{\ell_1\cdots\ell_N} = W(x_1^{(\ell_1)},\cdots,x_N^{(\ell_N)})$, it follows that $w_{mm\cdots m} = \sigma_m$ whenever $m \le \min_k \operatorname{modrank}(W)$. This is the first case in item 4. To prove the inequality in item 4, let $m = \min\{\ell_1,\ldots,\ell_N\}$ and suppose, without loss of generality, that $\ell_N = m$. Then, for all $v_k \in \mathcal{S}_k^{(m)}$, $k = 1,\ldots,N-1$,

$$\begin{aligned}\sigma_m = |W(x_1^{(m)},\ldots,x_N^{(m)})| = \\ = \max_{\substack{x_k\in\mathcal{S}_k^{(m)}\\ k=1,\ldots,N-1}} |W(x_1,\ldots,x_{N-1},x_N^{(m)})| \ge |W(v_1,\ldots,v_{N-1},x_N^{(m)})|.\end{aligned}$$

Substitute for v_k the singular vector $x_k^{(\ell_k)}$. Since $\ell_k \ge m$ we have that $x_k^{(\ell_k)} \in \mathcal{S}_k^{(\ell_k)} \subseteq \mathcal{S}_k^{(m)}$, i.e., $x_k^{(\ell_k)} \in \mathcal{S}_k^{(m)}$. It thus follows that
$\sigma_m \ge |W(x_1^{(\ell_1)},\ldots,x_{N-1}^{(\ell_{N-1})},x_N^{(m)})|$ as claimed. If $m > \min_k \operatorname{modrank}(W)$ then there exists $k \in \{1,\ldots,N\}$ for which $k > \operatorname{rank}_k(W) = R_k$. For this k we have $x_k^{(m)} \in \ker_k(W)$ and consequently, $w_{k\cdots k} = 0$. The fourth case in item 4 follows again from (3.25). Indeed, if $\ell_k > \ell_1 = \cdots = \ell_{k-1} = \ell_{k+1} = \cdots = \ell_N = k$ then $w_{\ell_1\cdots\ell_N} = W(x_1^{(m)},\cdots,x_k^{(\ell_k)},\cdots,x_N^{(m)})$ and, using orthonormality of the bases, $x_k^{(\ell_k)} \perp \operatorname{span}(x_k^{(1)},\ldots,x_k^{(m)})$ and hence $w_{\ell_1\cdots\ell_N} = 0$ by (3.25). Item 5 follows from (3.25). Indeed, if $L_1 = L_2 = \cdots = L_N = L_0$ then (3.25) shows that the singular value core tensor vanishes at $(L_0-1)+(L_0-2)+\cdots+1 = L_0(L_0-1)/2$ entries in its kth mode. The total number of zero entries of an order N tensor is therefore

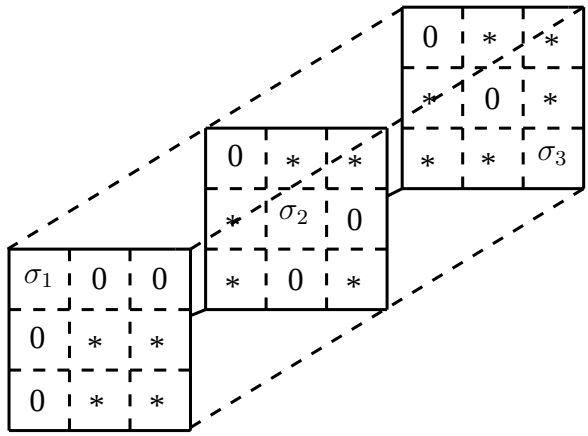


Figure 3.1: Visualization of the zero elements in the singular value core of an arbitrary order-3 tensor of dimensions $(3, 3, 3)$.

$\geq NL_0(L_0 - 1)/2$ as claimed. The result for the general case is immediate from this result. □

In words, any tensor W admits an SVD with at most $K = \min_k \operatorname{modrank}(W)$ non-zero singular values. In any singular value decomposition of W, the ordered singular values occur on the main diagonal of the N-way array $[[w_{\ell_1,\ldots,\ell_N}]]$ of elements of the tensor. In general, the singular value core tensor has non-zero entries on its non-diagonal elements. Absolute values of non-diagonal entries are bounded from above by the singular value of index equal to the smallest integer in the core index. Only if $N = 2$ (the matrix case) the singular value core tensor is diagonal. A visualization of the zero-structure in an order-3 singular value core is given in Figure 3.1.

Example 3.5.3. *Any $W \in \mathcal{T}_2$ admits a representation $W(u, v) = \langle u, Av\rangle = \langle A^\top u, v\rangle$ where $A \in \mathbb{R}^{m\times n}$. An SVD of W is then given by a representation of W with respect to the orthonormal bases $\{u_1, \ldots, u_m\}$ and $\{v_1, \ldots, v_n\}$ that consist of the ordered columns of the orthogonal matrices $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ that define a (any) singular value decomposition $A = U\Sigma V^\top$ of the matrix A. In particular, by item 4 of Theorem 3.5.2, the singular value core $[[w_{\ell_1\ell_2}]] \in \mathbb{R}^{m\times n}$ of W coincides with Σ as it has the $K = \operatorname{rank}(A)$ non-zero singular values of A on its main diagonal and is zero for all other elements. For $N = 2$ a tensor SVD therefore coincides with the matrix*

SVD.

Example 3.5.4. *Let the tensor $W : \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3 \to \mathbb{R}$ with vector spaces $\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{X}_3 = \mathbb{R}^2$ have coefficients $w_{111} = w_{211} = w_{112} = 1$ and $w_{121} = w_{221} = w_{212} = w_{122} = w_{222} = 0$ with respect to the standard Euclidean bases on $\mathbb{R}^2$. A computation of the singular vectors associated with W yields the orthonormal bases*

$$\left\{ \begin{pmatrix} \frac{\lambda}{\sqrt{\lambda^2+1}} \\ \frac{1}{\sqrt{\lambda^2+1}} \end{pmatrix}, \begin{pmatrix} \frac{1}{\sqrt{\lambda^2+1}} \\ \frac{-\lambda}{\sqrt{\lambda^2+1}} \end{pmatrix} \right\}; \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\}; \left\{ \begin{pmatrix} \frac{\lambda}{\sqrt{\lambda^2+1}} \\ \frac{1}{\sqrt{\lambda^2+1}} \end{pmatrix}, \begin{pmatrix} \frac{1}{\sqrt{\lambda^2+1}} \\ \frac{-\lambda}{\sqrt{\lambda^2+1}} \end{pmatrix} \right\}$$

of $\mathcal{X}_1$, $\mathcal{X}_2$ and $\mathcal{X}_3$, respectively, where $\lambda = \frac{1}{2} + \frac{1}{2}\sqrt{5}$. A representation of W with respect to this basis gives the singular value decomposition of W with singular values $\sigma_1 = \lambda$ and $\sigma_2 = 0$ and singular value core

$$\begin{aligned} &s_{111} = \sigma_1, \quad s_{121} = 0, \quad &&s_{112} = 0, \quad &&s_{122} = 0, \\ &s_{211} = 0, \quad s_{221} = 0, \quad &&s_{212} = -\frac{\lambda}{\sqrt{\lambda^2+1}}, \quad &&s_{222} = \sigma_2 = 0. \end{aligned}$$

Note that the singular values are on the 'main diagonal' entries s_{111}, s_{222} of the core, $|s_{212}| \leq \sigma_1$ and that not all off-diagonal entries are zero.

Example 3.5.5. *Consider a $2 \times 2 \times 2$ tensor W that is represented with respect to the standard bases in $\mathbb{R}^2 \times \mathbb{R}^2 \times \mathbb{R}^2$ with the elements $w_{111} = 2$, $w_{222} = \frac{1}{2}\sqrt{2}$, $w_{122} = \frac{1}{2}\sqrt{2}$ and with all other elements zero. Then W has singular values $\sigma_1 = w_{111} = 2$ and $\sigma_2 = w_{222} = \frac{1}{2}\sqrt{2}$ and it turns out that the standard basis defines a singular value decomposition of W. That is, W is already in SVD form and $[[w_{\ell_1\ell_2\ell_3}]] \in \mathbb{R}^{2\times2\times2}$ is the singular value core tensor of W. Observe that $\operatorname{modrank}(W) = (2,2)$, $\operatorname{rank}(W) = 3$, the singular value core is not diagonal and it has $5 > 3 = NL(L-1)/2$ zero entries.*

3.6 Low rank approximations

In this section we consider low-rank approximations as defined by the modal truncation in Definition 3.4.2. Here, we define subspaces $\mathcal{M}_k^{(r_k)}$ using the singular vectors defined in Definition 3.5.1 such that $\mathcal{M}_k^{(r_k)} = \operatorname{span}\{x_k^{(1)}, \ldots, x_k^{(r_k)}\}$ for $k =$

$1, \ldots, N$. This specific modal truncation will be referred to using the symbol W_r^*. W_r* is given by

$$W_r^* = \sum_{\ell=1}^{r_1} \cdots \sum_{\ell_N=1}^{r_N} w_{\ell_1 \cdots \ell_N} x_1^{(\ell_1)} \otimes \cdots \otimes x_N^{(\ell_N)}. \tag{3.26}$$

This section gives some low rank approximation properties that can be derived for the TSVD. These results have been published in [7] and [81].

3.6.1 Successive rank-1 approximations

The following theorem establishes that modal truncations of rank 1 are optimal solutions to problems P2 and P4.

Theorem 3.6.1. *Let $W \in \mathcal{T}_N$ and $r = (1, 1, \cdots, 1)$. Then the modal truncation W_r^* is a rank-1 tensor in $\mathcal{T}_N$ that is optimal in the sense that*

$$\|W - W_r^*\|_F \;=\; \inf\{\|W - W_1\|_F \mid W_1 \in \mathcal{T}_N \text{ has rank 1}\}$$

In particular, W_r^* is an optimal solution to problems P2 and P4. Moreover, the error $\|W - W_r^*\|_F^2 = \|W\|_F^2 - \sigma_1^2$ where σ_1 is the first singular value of W.

Proof. Let $W_1 \in \mathcal{T}_N$ be an arbitrary rank-1 tensor. Then W_1 can be written as $W_1 = \lambda U$ where $0 \neq \lambda \in \mathbb{R}$ and $U = u_1 \otimes \cdots \otimes u_N$ is a normalized rank-1 tensor in that $\|U\|_{\mathrm{F}} = 1$. Using the definition of the Frobenius norm, we have

$$\|W - \lambda U\|_{\mathrm{F}}^2 = \langle W - \lambda U, W - \lambda U\rangle = \langle W, W\rangle - 2\lambda\langle W, U\rangle + \lambda^2.$$

This is a convex function in λ that attains its minimum at $\lambda^* = \langle W, U\rangle$. But then

$$\begin{aligned}
\|W - \lambda^* U\|_{\mathrm{F}}^2 &= \langle W, W\rangle - 2\lambda^*\langle W, U\rangle + (\lambda^*)^2 = \\
&= \langle W, W\rangle - 2\langle W, U\rangle^2 + \langle W, U\rangle^2 = \\
&= \langle W, W\rangle - \langle W, U\rangle^2 = \\
&= \langle W, W\rangle - |W(u_1, \cdots, u_N)|^2
\end{aligned}$$

where the last equality follows from Lemma A.4.1. The latter expression shows that minimizing $\|W - \lambda^* U\|_{\mathrm{F}}$ over all rank-1 tensors U with $\|U\|_{\mathrm{F}} = 1$ is equivalent to

maximizing $|W(u_1, \cdots, u_N)|$ over all unit vectors u_n, $\|u_n\|_n = 1$, $n = 1, \ldots, N$. But this problem is (3.16) and has $U^* = x_1^{(1)} \otimes \cdots \otimes x_N^{(1)}$ as its optimal solution. Consequently, $\lambda^* = \langle W, U^* \rangle = \sigma_1$ and it follows that $W_1^* := \sigma_1 x_1^{(1)} \otimes \cdots \otimes x_N^{(1)}$ is the optimal rank-1 approximation of W. The error $\|W - W_1^*\|_{\mathrm{F}}^2 = \|W\|_{\mathrm{F}}^2 - \sigma_1^2$. $\square$

Theorem 3.6.1 is particularly useful to define an algorithm of successive rank-1 approximations of a given tensor $W \in \mathcal{T}_N$. Indeed, for given $W \in \mathcal{T}_N$, let $W_1^* := W_{(1,\ldots,1)}^*$ denote the optimal rank-1 tensor as defined in Theorem 3.6.1. The error $E_1 := W - W_1^*$ then belongs to $\mathcal{T}_N$ and is minimal in Frobenius norm when ranging over all tensors of the form $W - U \in \mathcal{T}_N$ with $U \in \mathcal{T}_N$ of rank-1. For successive values of $m > 1$, apply Theorem 3.6.1 to the error tensor E_{m-1} to define W_m^* as the optimal rank-1 tensor that minimizes the criterion $\|E_{m-1} - U\|_{\mathrm{F}}$ over all rank-1 tensors $U \in \mathcal{T}_N$. Then set $E_m := E_{m-1} - W_k^*$.

Definition 3.6.2. *Given $W \in \mathcal{T}_N$, the* rth order successive rank-1 approximation of W *is the tensor*

$$W^{(r)} := W_1^* + \cdots + W_r^* \tag{3.27}$$

where $W_1^*, \ldots, W_r^*$ are optimal rank-1 approximations of W, $E_1, \ldots, E_{r-1}$, respectively, as defined in the previous paragraph.

In this construction, the Frobenius norm of the error $E_m = W - W^{(m)}$ satisfies the recursion $\|E_m\|_{\mathrm{F}}^2 = \|E_{m-1}\|_{\mathrm{F}}^2 - \sigma_1^2(E_{m-1})$ with $\|E_1\|_{\mathrm{F}}^2 = \|W\|_{\mathrm{F}}^2 - \sigma_1^2(W)$. In particular, $\|E_m\|_{\mathrm{F}} \leq \|E_{m-1}\|_{\mathrm{F}}$ so that the norm of successive approximation errors is non-increasing.

Remark 3.6.3. *The rank-1 modal truncation $W_{1,\ldots,1}^*$ defined in Theorem 3.6.1 is not optimal in the induced norm. That is, the rank-1 modal truncation does not solve problems P1 and P3 for $r = (1, \ldots, 1)$.*

3.6.2 One modal-rank approximations

The following result establishes a lower bound on the approximation error between a tensor and its modal truncation when only one modal rank is reduced.

Theorem 3.6.4. *Let $W \in \mathcal{T}_N$ have modal rank* $\mathrm{modrank}(W) = (R_1, \ldots, R_N)$ *and let*

$$r = (R_1, \cdots, R_{k-1}, r_k, R_{k+1}, \cdots, R_N)$$

with $r_k \leq R_k$. Then

$$\|W - W_r^*\| \geq \sigma_{r_k+1}.$$

Proof. Without loss of generality assume $k = 1$ and $r_1 \leq R_1$. Define $E := W - W_r^*$. Then

$$\begin{aligned} E = &[W - W_r^*]|_{\mathcal{M}_1^{r_1} \times \mathcal{X}_2 \cdots \times \mathcal{X}_N} + [W - W_r^*]|_{\{\mathcal{M}_1^{r_1}\}^\perp \times \mathcal{X}_2 \cdots \times \mathcal{X}_N} \\ &- [W - W_r^*]|_{\{0\} \times \mathcal{X}_2 \cdots \times \mathcal{X}_N} = \\ &= 0 + W|_{\{\mathcal{M}_1^{r_1}\}^\perp \times \mathcal{X}_2 \cdots \times \mathcal{X}_N} - 0 \end{aligned}$$

where we used the definition of W_r^*, Lemma 3.25, and the fact that, $E|_{\{0\} \times \mathcal{X}_2 \cdots \times \mathcal{X}_N}$ satisfies $E(0, x_2, \cdots, x_N) = \sum_{\ell_1} \cdots \sum_{\ell_N} w_{\ell_1 \ldots \ell_N} \langle e_1^{\ell_1}, 0 \rangle \langle e_2^{\ell_2}, x_2 \rangle \cdots \langle e_N^{\ell_N}, x_N \rangle = 0$. Furthermore, since $\mathcal{S}_1^{r_1+1} \times \cdots \times \mathcal{S}_N^{r_1+1} \subset [\mathcal{M}_1^{r_1}]^\perp \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_N$ it follows that

$$\begin{aligned} \|E\| &= \max_{\substack{x_1 \in [\mathcal{M}_1^{r_1}]^\perp \\ x_k \in \mathcal{X}_k,\ \|x_k\| = 1 \\ k = 1 \ldots, N}} |W(x_1, \cdots, x_N)| \\ &\geq \max_{\substack{x_k \in \mathcal{S}_k^{(r_1+1)} \\ k = 1, \ldots, N}} |W(x_1, \cdots, x_N)| =: \sigma_{r_1+1} \end{aligned}$$

Consequently, $\|W - W_r^*\| \geq \sigma_{r_1+1}$. This yields the result. □

The following theorem states that when only one of the arguments of the tensor is approximated, the approximation error decreases for increasing approximation order.

Theorem 3.6.5. *Given $W \in \mathcal{T}_N$ of modal rank $R = (R_1, \ldots, R_N)$. Define for $r = (m, R_2, \ldots, R_N)$ the approximation error $E_m = W - W_r^*$, where W_r^* is defined in* (3.26)*. Then we have $\|E_{m+1}\|_F \leq \|E_m\|_F$.*

Proof. Without loss of generality let $p = 1$. Then $W_r^* = W|_{\mathcal{M}_1^{(k)} \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_N}$ with $\mathcal{M}_1^{(k)} = \operatorname{span}\{\varphi_1^{(1)}, \ldots, \varphi_1^{(k)}\}$. Then

$$\begin{aligned} W = W|_{\mathcal{M}_1^{(k)} \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_N} + W|_{\mathcal{M}_1^{(k)\perp} \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_N} - \\ W|_{(\mathcal{M}_1^{(k)} \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_N) \bigcap (\mathcal{M}_1^{(k)\perp} \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_N)}. \end{aligned}$$

Since

$$(\mathcal{M}_1^{(k)} \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_N) \bigcap (\mathcal{M}_1^{(k)^\perp} \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_N)$$

is equal to

$$(\mathcal{M}_1^{(k)} \cap \mathcal{M}_1^{(k)^\perp}) \times (\mathcal{X}_2 \cap \mathcal{X}_2) \times \ldots \times (\mathcal{X}_N \cap \mathcal{X}_N)$$

and $W|_{\emptyset \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_N} = 0$, we infer that

$$E_k = W|_{\mathcal{M}^{(k)^\perp} \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_N}.$$

Since $\mathcal{M}_1^{(k)} \subseteq \mathcal{M}_1^{(k+1)}$ and, consequently, $\mathcal{M}_1^{(k)^\perp} \supseteq \mathcal{M}_1^{(k+1)^\perp}$, we conclude $\|E_{k+1}\| \leq \|E_k\|$. □

3.6.3 Approximation of diagonalizable tensors

The diagonal of an arbitrary tensor $W \in \mathcal{T}_N$ represented with respect to the bases (3.1) is given by the elements $w_{\ell_1,\ldots,\ell_N}$ with $\ell_1 = \cdots = \ell_N$. We will say that a tensor is *diagonal* if only its diagonal elements are nonzero. Whenever a collection of bases can be found such that W is diagonal we will say that W is *diagonalizable*. When considering higher-order statistics in the problem of Independent Component Analysis, diagonal tensors are of considerable importance. See, e.g. [20, 23, 18].

Theorem 3.6.6. *Let $W \in \mathcal{T}_N$, then*

1. *Every $W \in \mathcal{T}_2$ is diagonalizable. Moreover, the singular value decomposition of W gives a singular value core tensor that is diagonal.*

2. *For $N > 2$ not every $W \in \mathcal{T}_N$ is diagonalizable. If $W \in \mathcal{T}_N$ is diagonalizable, then the singular value core tensor of W is, in general, not diagonal.*

3. *If W is diagonalizable with respect to a collection of orthogonal bases, then the singular value core tensor of W will be diagonal.*

Proof. 1. Every $W \in \mathcal{T}_2$ can be written as $W(x_1, x_2) = \langle x_1, Ax_2 \rangle$ for some matrix A. Let $A = U\Sigma V^\top$ be an SVD of A. Then the singular value core tensor of W is given by the diagonal matrix Σ.

2. A counterexample is given in Example 3.6.7 below.

3. Suppose that X_k is an orthogonal full rank matrix, $k = 1, \ldots, N$ such that $\overline{W} = W \bullet_1 X_1 \cdots \bullet_N X_N$ becomes diagonal. Then $S_k := X_k^\top X_k > 0$ and $Q_k := X_k S_k^{1/2} \Pi_k$ is unitary for any permutation matrix Π_k of dimension $L_k \times L_k$. Now, $\overline{W} = W \bullet_1 Q_1 \cdots \bullet_N Q_N$ remains diagonal and the permutation matrices Π_k can be chosen such that the diagonal elements of $\overline{W}$ are non-increasing. $\overline{W}$ is then the singular value core tensor of W and the columns of Q_k define the k-mode singular vectors.

□

Example 3.6.7. *An example of the second item of Theorem 3.6.6 is given by the tensor W defined in Example 3.5.5. As already shown, the singular value core of W is not diagonal. However, with respect to the bases*

$$\left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{1}{2}\sqrt{2} \\ \frac{1}{2}\sqrt{2} \end{pmatrix} \right\}, \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\}, \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\}.$$

one easily shows that W admits a diagonal representation with diagonal elements $w_{111} = 2$ and $w_{222} = 1$. Hence, a diagonalizable tensor will not necessarily have a diagonal singular value decomposition.

Theorem 3.6.8. *If $W \in \mathcal{T}_N$ is diagonalizable with respect to an orthonormal basis, then*

1. *the rank and orthogonal rank of W are equal.*
2. *the singular value decomposition of W is a completely orthogonal rank decomposition.*
3. *the singular value decomposition of W is given by*

$$W = \sum_{k=1}^{R} \sigma_k x_1^{(k)} \otimes \cdots \otimes x_N^{(k)}$$

where R is equal to the rank of W. The modal truncation W_r^* defined in (3.26) *is represented as*

$$W_r^* = \sum_{k=1}^{r} \sigma_k x_1^{(k)} \otimes \cdots \otimes x_N^{(k)}$$

and is an optimal rank-r approximation of W in the sense that

$$\inf_{\substack{W_r \in \mathcal{T}_N \\ \operatorname{modrank}(W_r)=r}} \|W - W_r\| \quad = \quad \|W - W_r^*\| \quad = \quad \sigma_{r+1}.$$

Moreover,

$$\inf_{\substack{W_r \in \mathcal{T}_N \\ \operatorname{modrank}(W_r)=r}} \|W - W_r\|_F^2 \;=\; \|W - W_r^*\|_F^2 \;=\; \sum_{k=r+1}^{R} \sigma_k^2.$$

That is, the modal truncation W_r^* is an optimal solution to problems P1,P2,P3 and P4.

Proof. Item 1 is proven in [46]. For item 2 see [85]. To see item 3, let $\{x_k^{\ell_k}\}_{\ell_k=1}^{L_k}$, $k = 1 \ldots, N$ be an orthonormal basis for which W has a diagonal representation where the diagonal entries $\sigma_k = w_{k\cdots k}$ are assumed to be ordered in that $\sigma_1 \geq \sigma_2 \geq \ldots$. The error tensor $E := W - W_r^*$ is then given by $E = \sum_{k>r_-}^{K} \sigma_k w_1^{(k)} \otimes \cdots \otimes w_N^{(k)}$. This gives $\|E\| = \sigma_{r_-+1}$ and $\|E\|_{\mathrm{F}}^2 = \sum_{k>r_-}^{K} \sigma_k^2$ which are minimal in view of the ordering of the singular values. See also [85]. □

3.7 Improved accuracy

One consequence of Definition 3.5.1 is that the computation of the singular vector $x_k^{(m)} \in \mathcal{X}_k$ of order m not only depends on singular vectors $x_k^{(j)}$ of order $j < m$ but also on the singular vectors $w_p^{(j)}$ for $p \neq k$ and $j < m$. If modal truncations in one specific modal direction, say the kth, are searched for, then the coupling of the constraints in the computation of the singular vectors of order m may actually prevent the modal truncation W_r^* defined in Theorem 3.6.4 to be optimal. A weakening of the constraints on the set $\mathcal{S}_k^{(m)}$ in (3.18) may then become an alternative. A modified singular value decomposition can be obtained by redefining the set $\mathcal{S}_k^{(m)}$ for $\min_k \operatorname{modrank}(W) < m < \max_k \operatorname{modrank}_k(W)$ by, for example,

$$\mathcal{S}_k^{(m)} := \{x \in \mathcal{X}_k \mid \|x\|_k = 1\}$$

and by performing the optimization in (3.18). The construction and properties of the decomposition that is thus obtained form the topic of this section. The results of this section have been published in [7].

The new construction for decomposing a tensor is as follows. Let $W \in \mathcal{T}_N$ be an order-N tensor defined on the finite dimensional vector spaces $\mathcal{X}_1, \ldots, \mathcal{X}_N$ where we suppose that $\dim(\mathcal{X}_n) = L_n$. Furthermore, let $\mathcal{X}' := \mathcal{X}_1 \times \ldots \times \mathcal{X}_i$ and $\mathcal{X}'' := \mathcal{X}_{i+1} \times \ldots \times \mathcal{X}_N$, where $0 < i < N$. The *dedicated singular values* of W, denoted $\hat{\sigma}_k(W)$, with $k = 1, \ldots, K$ and $K = \min_{p=1,\ldots,i} \operatorname{modrank}(W)$ are defined as follows.
Let

$$\begin{aligned} \mathcal{S}_k^{(1)} &:= \{x \in \mathcal{X}_k \mid \|x\|_k = 1\} \quad \text{for } k = 1, \ldots, i \\ \mathcal{S}_k &:= \{x \in \mathcal{X}_k \mid \|x\|_k = 1\} \quad \text{for } k = i+1, \ldots, N \end{aligned}$$

denote the unit sphere in $\mathcal{X}_k$. Define the first dedicated singular value of W by

$$\hat{\sigma}_1(W) := \sup_{\substack{x_k \in \mathcal{S}_k^{(1)},\ 1 \le k \le i \\ x_k \in \mathcal{S}_k,\ (i+1) \le k \le N}} |W(x_1, \ldots, x_N)|. \tag{3.28}$$

Since W is continuous and the Cartesian product $\mathcal{S}^{(1)} = \mathcal{S}_1^{(1)} \times \cdots \times \mathcal{S}_i^{(1)} \times \mathcal{S}_{i+1} \times \cdots \times \mathcal{S}_N$ of unit spheres is a compact set, an extremal solution of (3.28) exists (i.e., the supremum in (3.16) is a maximum) and is attained by an N-tuple

$$(\psi_1^{(1)}, \ldots, \psi_N^{(1)}) \in \mathcal{S}^{(1)}.$$

Subsequent dedicated singular values of W are defined in an inductive manner by setting

$$\mathcal{S}_k^{(m)} := \{x \in \mathcal{X}_k \mid \|x\|_k = 1,\ \langle x, \psi_k^{(j)} \rangle_k = 0 \text{ for } j = 1, \ldots, (m-1)\}$$

for $k = 1, \ldots, i$, and by defining

$$\hat{\sigma}_m(W) = \sup_{\substack{x_k \in \mathcal{S}_k^{(m)},\ 1 \le k \le i \\ x_k \in \mathcal{S}_k,\ (i+1) \le k \le N}} |W(x_1, \ldots, x_N)|, \quad k \le K. \tag{3.29}$$

Again, since the Cartesian product

$$\mathcal{S}^{(m)} = \mathcal{S}_1^{(m)} \times \cdots \times \mathcal{S}_i^{(m)} \times \mathcal{S}_{i+1} \times \cdots \times \mathcal{S}_N$$

is compact, the supremum in (3.29) is a maximum that is attained by an N-tuple

$$(\psi_1^{(m)}, \ldots, \psi_N^{(m)}) \in \mathcal{S}^{(m)}.$$

Note that the set over which the optimization takes place, $\mathcal{S}^{(m)}$, is in general a larger subset of the Cartesian product of unit balls than the set $\mathcal{S}^{(m)}$ as originally defined

in (3.17). It follows that the vectors $\psi_k^{(1)}, \ldots, \psi_k^{(K)}$ are mutually orthonormal in $\mathcal{X}_k$, for $k = 1, \ldots, i$. If $K < L_k$ for any $1 \leq k \leq i$, then we extend the collection of orthogonal elements $\psi_k^{(1)}, \ldots, \psi_k^{(K)}$ to a complete orthonormal basis of $\mathcal{X}_k$. This construction leads to a collection of orthonormal bases

$$\{\psi_1^{(\ell_1)}, \ell_1 = 1, \ldots, L_1\}, \ldots, \{\psi_i^{(\ell_i)}, \ell_i = 1, \ldots, L_i\} \tag{3.30}$$

for the vector spaces $\mathcal{X}_1, \ldots, \mathcal{X}_i$, respectively. We will call elements of these orthonormal bases *dedicated singular vectors* of the tensor W.
Since there is no construction of orthonormal bases for the vector spaces $\mathcal{X}_{i+1}, \ldots, \mathcal{X}_N$, it is not possible nor appropriate to define a singular-value-like decomposition of the tensor using dedicated singular vectors. Instead, we define a *dedicated representation* of the tensor, which can be used to define a *dedicated modal truncation*.

Definition 3.7.1. *Given an order-N tensor $W \in \mathcal{T}_N$, with $W : \mathcal{X}_1 \times \ldots \times \mathcal{X}_N \to \mathbb{R}$. Assume $\mathcal{X}' = \mathcal{X}_1 \times \ldots \times \mathcal{X}_i$ and $\mathcal{X}'' = \mathcal{X}_{i+1} \times \ldots \times \mathcal{X}_N$. Then, a* dedicated representation *of W can be defined as a representation of W with respect to the bases* (3.30) *for $\mathcal{X}'$, where the original bases for $\mathcal{X}''$ are kept intact, i.e.*

$$W^d = \sum_{\ell_1=1}^{L_1} \cdots \sum_{\ell_N=1}^{L_N} \tilde{w}_{\ell_1 \cdots \ell_N} \psi_1^{(\ell_1)} \otimes \cdots \otimes \psi_i^{(\ell_i)} \otimes f_{i+1}^{(\ell_{i+1})} \otimes \cdots \otimes f_N^{(\ell_N)} \tag{3.31}$$

$$= \sum_{\ell_1=1}^{L_1} \cdots \sum_{\ell_N=1}^{L_N} \tilde{w}_{\ell_1 \cdots \ell_N} U_{\ell_1 \cdots \ell_N}. \tag{3.32}$$

Using this representation of W, a dedicated modal truncation can be defined.

Definition 3.7.2. *Given an order-N tensor $W \in \mathcal{T}_N$, with dedicated representation W^d and a vector of integers $r = (r_1, \ldots, r_i)$ with $r_k \leq R_k$ for $k = 1, \ldots, i$. Let*

$$\mathcal{M}_k^{(m)} = \operatorname{span}\{\psi_k^{(1)}, \ldots, \psi_k^{(m)}\}, \quad k = 1, \ldots, i.$$

with $m \leq R_k$. A dedicated modal truncation *is then defined by the restriction $W_r^d := W^d|_{\mathcal{M}_1^{(r_1)} \times \cdots \mathcal{M}_i^{(r_i)}}$ and is represented by the expansion*

$$W_r^d = \sum_{\ell_1=1}^{r_1} \cdots \sum_{\ell_i=1}^{r_i} \sum_{\ell_{i+1}=1}^{R_{i+1}} \cdots \sum_{\ell_N=1}^{R_N} w_{\ell_1 \cdots \ell_N} \psi_1^{(\ell_1)} \otimes \cdots \otimes \psi_i^{(\ell_i)} \otimes f_{i+1}^{(\ell_{i+1})} \otimes \cdots \otimes f_N^{(\ell_N)}. \tag{3.33}$$

The following theorem states some properties of the dedicated representation of a tensor.

Theorem 3.7.3. *Consider $W \in \mathcal{T}_N$.*

1. *For all $1 \leq i \leq N$ the dedicated representation of W exists.*

2. *The dedicated representation is an orthogonal decomposition of W in the sense that the rank-one tensors $U_{\ell_1 \cdots \ell_N}$ in* (3.32) *are mutually orthogonal*

$$\langle U_{\ell_1 \cdots \ell_N}, U_{\ell'_1 \cdots \ell'_N} \rangle = 0, \quad \textit{unless } \ell_n = \ell'_n, \ \forall n = 1, \ldots, N.$$

3. $$\hat{\sigma}_1 \geq \cdots \geq \hat{\sigma}_K \geq 0 \tag{3.34}$$

4. *There exists an orthonormal basis $\{\psi_n^{(1)}, \ldots, \psi_n^{(L_n)}\}$ of X_n with $n > i$ such that*

$$W(\psi_1^{(k)}, \ldots, \xi, \ldots, \psi_i^{(k)}, \psi_{i+1}^{(\ell_{i+1})}, \ldots, \psi_N^{(\ell_N)}) = 0 \tag{3.35}$$

   for all $\xi \perp \operatorname{span}\{\psi_n^{(1)}, \ldots \psi_n^{(k)}\}$, where ξ is at the nth spot, with $1 \leq n \leq i$.

Proof. Proof of Theorem 3.7.3

1. Since extremal solutions to the optimization problems (3.28) and (3.29) are guaranteed to exist for any tensor $W \in \mathcal{T}_N$, also the dedicated representation is guaranteed to exist.

2. Using Lemma A.4.1 we have

$$\langle U_{\ell_1 \cdots \ell_N}, U_{\ell'_1 \cdots \ell'_N} \rangle = \prod_{n=1}^{i} \langle \psi_n^{(\ell_1)}, \psi_n^{(\ell'_n)} \rangle \prod_{n=i+1}^{N} \langle e_n^{(\ell_n)}, e_n^{(\ell'_n)} \rangle.$$

   Therefore, the inner product between unequal rank-one tensors is zero whenever one of the inner products on the right-hand side is zero. Since the bases are orthonormal, all rank-one tensors are orthogonal unless $\ell_n = \ell'_n$ for $1 \leq n \leq N$.

3. This is by construction

4. Let $L = L_1 + \cdots + L_N$ and associate with the optimization problem (3.28) the Lagrangian $L_1 : \mathbb{R}^{L+N} \to \mathbb{R}$ by setting

$$L_1(x, \lambda) := W(x_1, \ldots, x_N) + \sum_{n=1}^{N} \frac{1}{2}\lambda_n(1 - \langle x_n, x_n\rangle).$$

It has already been argued that an N-tuple $x^{(1)} = (\psi_1^{(1)}, \ldots, \psi_N^{(1)})$ exists that attains the maximum in (3.28). From the theory of variational analysis [10, 31], one then infers the existence of an N-tuple $\lambda^{(1)} = (\lambda_1^{(1)}, \ldots, \lambda_N^{(1)})$ of Lagrange multipliers such that

$$\nabla L_1(^{(1)}, \lambda^{(1)}) = 0, \tag{3.36}$$

where ∇L_1 denotes the gradient of L_1. The n-mode Fréchet derivative $\partial_n W(x_1, \ldots, x_N)$ of W at the point $(x_1, \ldots, x_N)$ is an order-1 tensor (a linear functional) that maps X_n to $\mathbb{R}$ and satisfies

$$\partial_n W(x_1, \ldots, x_N) = W(x_1, \ldots, x_{n-1}, \cdot, x_{n+1}, \ldots, x_N)$$

where the 'dot' is at the nth spot. By the multi-linearity of the tensor, $\partial_n W(x_1, \ldots, x_N)$ is independent of $x_n \in X_n$. Hence, rewriting (3.36) for each independent modal direction gives that $x^{(1)}, \lambda^{(1)}$ satisfies, for $n = 1, \ldots, N$,

$$W(\psi_1^{(1)}, \ldots, \psi_{n-1}^{(1)}, \cdot, \psi_{n+1}^{(1)}, \ldots, \psi_N^{(1)}) = \lambda_n^{(1)}\langle \cdot, \psi_n^{(1)}\rangle, \tag{3.37a}$$

$$\|\psi_n^{(1)}\| = 1. \tag{3.37b}$$

(3.37a) implies that for each $n = 1, \ldots, N$,

$$W(\psi_1^{(1)}, \ldots, \psi_{n-1}^{(1)}, \xi_n, \psi_{n+1}^{(1)}, \ldots, \psi_N^{(1)}) = 0 \quad \text{whenever } \langle \xi_n, \psi_n^{(1)}\rangle = 0.$$

In a similar manner, for $k > 1$ we associate with the optimization problem (3.29) the Lagrangian $L_k : \mathbb{R}^{L+N+i(k-1)} \to \mathbb{R}$ defined by

$$L_k(x, \lambda, \mu) = W(x_1, \ldots, x_n) + \sum_{n=1}^{N} \frac{1}{2}\lambda_n(1 - \langle x_n, x_n\rangle) + \sum_{n=1}^{i} \langle g_n(x_n), \mu_n\rangle.$$

where $x_n \in X_n$, $\lambda_n \in \mathbb{R}$, $\mu_n \in \mathbb{R}^{k-1}$ and $g_n : X_n \to \mathbb{R}^{k-1}$ is given by

$$g_n(\xi_n) := \begin{pmatrix} \langle \xi_n, \psi_n^{(1)}\rangle \\ \vdots \\ \langle \xi_n, \psi_n^{(k-1)}\rangle \end{pmatrix}.$$

Again, there exist N-tuples $x^{(k)}$, $\lambda^{(k)}$ and $\mu^{(k)}$ that satisfy the stationarity condition

$$\nabla L_k(x^{(k)}, \lambda^{(k)}, \mu^{(k)}) = 0. \tag{3.38}$$

Rewriting (3.38) for each modal direction gives, for $n = 1, \ldots, i$, that

$$W(\psi_1^{(k)}, \ldots, \psi_{n-1}^{(k)}, \cdot, \psi_{n+1}^{(k)}, \ldots, \psi_N^{(k)}) = \lambda_n^{(k)} \langle \cdot, \psi_n^{(k)} \rangle + \langle g_n(\cdot), \mu_n^{(k)} \rangle, \tag{3.39a}$$

$$\|\psi_n^{(k)}\| = 1, \tag{3.39b}$$

$$g_n(\psi_n^{(k)}) = 0. \tag{3.39c}$$

Now suppose, again for $n = 1, \ldots, i$, that $\xi \perp \operatorname{span}\{\psi_n^{(1)}, \ldots \psi_n^{(k)}\}$. Substituting ξ for the dotted argument in (3.39a) gives that

$$W(\psi_1^{(k)}, \ldots, \xi, \ldots, \psi_i^{(k)}, \psi_{i+1}^{(\ell_{i+1})}, \ldots, \psi_N^{(\ell_N)}) = 0 \tag{3.40}$$

for all $\xi \perp \operatorname{span}\{\psi_n^{(1)}, \ldots \psi_n^{(k)}\}$, where ξ is at the nth spot, with $1 \leq n \leq i$.

□

Remark 3.7.4. *Consider a tensor $W \in \mathcal{T}_N$. Item 4 of Thm. 3.7.3 immediately yields the following results regarding the zero structure of the dedicated representation* (3.31) *of W*

1. *The core $[[\tilde{w}_{\ell_1 \cdots \ell_N}]]$ of the dedicated representation of W satisfies*

$$\tilde{w}_{\ell_1 \cdots \ell_N} = \begin{cases} 0 & \text{if } \ell_1 = \ldots = \ell_N > K \\ 0 & \text{if } \ell_n > \ell_1 = \ldots = \ell_{n-1} = \ell_{n+1} = \ldots = \ell_i \end{cases} \tag{3.41}$$

2. *Consider the case that $L_1 = L_2 = \cdots = L_N = L$. Then the number of zeros in the core of the dedicated representation of W is at least*

$$i \left(\frac{L(L-1)}{2} \right).$$

The following theorem establishes a relationship between the original and dedicated singular values.

Theorem 3.7.5. *Consider $W \in \mathcal{T}_N$.*

1. *Both the original and dedicated singular values of a tensor are ordered*

$$\sigma_1 \geq \ldots \geq \sigma_K \geq 0 \tag{3.42}$$

$$\hat{\sigma}_1 \geq \ldots \geq \hat{\sigma}_K \geq 0 \tag{3.43}$$

2. *$\sigma_1 = \hat{\sigma}_1$ and*

$$\sigma \cdot \varphi_1^{(1)} \otimes \cdots \otimes \varphi_N^{(1)} = \hat{\sigma}_1 \cdot \psi_1^{(1)} \otimes \cdots \otimes \psi_N^{(1)}$$

3. *$\hat{\sigma}_2 \geq \sigma_2$*

Proof. Proof of Theorem 3.7.5.

1. This is by construction.

2. Since these optimization problems are identical, the first singular value and the singular vectors that are found will also be identical.

3. This uses the previous part of this theorem. Since the results from the first optimization are identical, the optimization domains $\mathcal{S}_n^{(2)},\ n = 1, \ldots, N$ will be the same for both optimization problems. As the dedicated SVD construction will incorporate less constraints for the second step, it only takes $\mathcal{S}_n^{(2)},\ n = 1, \ldots, i$ into account and uses the unit sphere for the rest of the vector spaces, $\hat{\sigma}_2 \geq \sigma_2$.

□

3.8 Algorithms and computational issues

In this section we propose an efficient algorithm for the computation of the TSVD as defined in Definition 3.5.1 of an arbitrary order-N tensor $W \in \mathcal{T}_N$. The algorithm is based on the fixed point properties of a contractive mapping G that is iterated in a power-type algorithm to compute the singular vectors of order m and the singular values as defined in Definition 3.5.1.

With $x_k \in \mathcal{X}_k$ and $\sigma \in \mathbb{R}$, we denote by x the vector $x = \mathrm{col}(x_1, \dots, x_N, \sigma)$ in $\mathbb{R}^{L_1+\cdots+L_N+1}$. To simplify notation, let $L = L_1 + \cdots + L_N$ and define the mapping $G : \mathbb{R}^{L+1} \to \mathbb{R}^{L+1}$ by

$$G(x) := \begin{pmatrix} \frac{1}{\sigma}\nabla_1 W(\frac{x_1}{\|x_1\|}, \dots, \frac{x_N}{\|x_N\|}) \\ \vdots \\ \frac{1}{\sigma}\nabla_N W(\frac{x_1}{\|x_1\|}, \dots, \frac{x_N}{\|x_N\|}) \\ W(\frac{x_1}{\|x_1\|}, \dots, \frac{x_N}{\|x_N\|}) \end{pmatrix} \tag{3.44}$$

Here, $\nabla_k W(x_1, \dots, x_N) = [\partial_k W(x_1, \dots, x_N)]^\top$ is the k-mode gradient of W in the point $(x_1, \dots, x_N)$ (i.e., the transpose of the n-mode Fréchet derivative of W). Then $G(x)$ is well defined provided that $x_k \neq 0$ for any k and $\sigma \neq 0$. The following theorem relates the fixed points of G to solutions of the $2N$ equations in the Lagrangian system (3.37).

Theorem 3.8.1. *x^* is a fixed point of G if and only if $x^* = \mathrm{col}(x_1^{(1)}, \dots, x_N^{(1)}, \sigma^{(1)})$ satisfies the Lagrangian conditions* (3.37) *with $\lambda_1^{(1)} = \cdots = \lambda_N^{(1)} = \sigma_1$.*

Proof. Only if: Suppose $x^* = \mathrm{col}(x_1, \dots, x_N, \sigma)$ is a fixed point of G. Then $G(x^*)$ $= x^*$ from which it follows that $\sigma = W(\frac{x_1}{\|x_1\|}, \dots, \frac{x_N}{\|x_N\|})$ and $\sigma x_k = \nabla_k W(x_1/\|x_1\|, \dots, x_N/\|x_N\|)$ for all $x = 1, \dots, N$. Consider these equalities for $x = 1$. Since the 1-mode Fréchet derivative $\partial_1 W(x_1, \dots, x_N) = W(\cdot, x_2, \dots, x_N)$ it follows that

$$\sigma\langle v_1, x_1\rangle = W(v_1, \frac{x_2}{\|x_2\|}, \cdots, \frac{x_N}{\|x_N\|} \quad \text{for all } v_1 \in \mathcal{X}_1. \tag{3.45}$$

By taking $v_1 = \frac{x_1}{\|x_1\|}$ we infer that $\sigma = \sigma\langle x_1/\|x_1\|, x_1\rangle = \sigma\|x_1\|$ so that we conclude that $\|x_1\| = 1$. In a similar fashion one shows that $\|x_k\| = 1$ for all $k = 1, \dots, N$. This gives (3.37b). But with unit norms, (3.45) reads $\sigma\langle \cdot, x_1\rangle = W(\cdot, x_2, \dots, x_N)$ which is (3.37a) with $k = 1$. the same argument applies to prove (3.37a) for other k.
If: Suppose a set of N vectors $x_k \in \mathcal{X}_k$, $k = 1, \dots, N$ satisfies $\|x_k\| = 1$ and

$$W(x_1, \dots, x_{k-1}, \cdot, x_{k+1}, \dots, x_N) = \sigma\langle \cdot, x_k\rangle.$$

Then

$$\frac{1}{\sigma}\nabla_k W(\frac{x_1}{\|x_1\|}, \dots, \frac{x_N}{\|x_N\|}) = \frac{1}{\sigma}\nabla_k W(x_1, \dots, x_N) = x_k,$$

for $k = 1, \dots, N$, which shows that $x = \mathrm{col}(x_1, \dots, x_N, \sigma)$ is fixed point of G. □

The result of Theorem 3.8.1 gives rise to the following TSVD algorithm for the computation of a singular value decomposition.

INPUT Tensor $W \in \mathcal{T}_N$ with k-mode dimension L_k.

DESIRED A singular value decomposition of W.

Step 0 (Initialization) Set tolerance level $\varepsilon_{\text{tol}} > 0$, order $m = 1$, and $W_m = W$.

Step 1 Select random elements $x_k^0 \in \mathcal{X}_k$, $k = 1, \ldots, N$ and σ^0 with $\|x_k^0\|_k = 1$ and $0 < \sigma^0 < 1$. Set $x^0 := \operatorname{col}(x_1^0, \ldots, x_N^0, \sigma^0)$.

Step 2 Let G be defined by (3.44) with $W = W_m$, and iterate the map

$$x^{i+1} = G(x^i), \qquad i = 0, 1, 2, \ldots, i^* \tag{3.46}$$

where i^* is such that $\|x^{i^*} - x^{i^*-1}\| < \varepsilon_{\text{tol}}$.

Step 3 Write $x^{i^*} = \operatorname{col}(x_1^*, \ldots, x_N^*, \sigma^*)$ and define, for $k = 1, \ldots, N$,

$$\begin{aligned} &\sigma_m = \sigma^*, \quad x_k^{(m)} = x_k^*, \\ &X_k^{(m)} = \begin{pmatrix} x_k^{(1)} & \cdots & x_k^{(m)} \end{pmatrix}, \\ &Q_k^{(m)} = I - X_k^{(m)} [X_k^{(m)}]^\top. \end{aligned}$$

Step 4 Define the tensor

$$W_{m+1} = W \,\blacksquare_1\, Q_1^{(m)} \cdots \blacksquare_N\, Q_N^{(m)}$$

and set m to $m + 1$.

Step 5 Repeat Step 1, Step 2, Step 3, Step 4 until $m = K = \min_k \operatorname{modrank}(W)$.

Step 6 For every k for which $K < L_k$ complement $X_k^{(K)}$ to an orthonormal matrix $X_k^{(L_k)}$.

Step 7 Define

$$\overline{W} = W \,\blacksquare_1\, X_1^{(L_1)} \cdots \blacksquare_N\, X_N^{(L_N)}.$$

Theorem 3.8.2. *Suppose that $G : \mathbb{R}^{L+1} \to \mathbb{R}^{L+1}$ maps a closed subset $\mathcal{D} \subset \mathbb{R}^{L+1}$ into itself and that G is contractive on $\mathcal{D}$ in the sense that there exists $\alpha < 1$ such that $\|Gx - Gy\| \leq \alpha \|x - y\|$ for all $x, y \in \mathcal{D}$. Then the iteration* (3.46) *converges to a unique fixed point x^* of G in $\mathcal{D}$. In that case, for each $m = 1, \ldots, K$ the vectors $x_k^{(m)}$ with $k = 1, \ldots, N$ satisfy the Lagrangian conditions* (3.39)*.*

The statements on the convergence of the iteration (3.46) can be found in [59]. Theorem 3.8.2 states that whenever (3.44) is a contractive mapping on a sufficiently large closed invariant set $\mathcal{D}$ then the iteration (3.46) converges to the solution of the Lagrangian systems (3.37) and (3.39). It is easy to see that contractivity of (3.44) with $W = W_1$ implies contractivity of (3.44) for $W = W_m$ with $m > 1$.
In practice it is not trivial to explicitly verify this condition and to find a closed invariant region $\mathcal{D}$ that makes G contractive. However, Theorem 3.8.2 promises that whenever the algorithm converges, it converges to a solution of the Lagrangian systems (3.37) and (3.39).

Remark 3.8.3. *We remark that singular vectors and singular values of* W necessarily *satisfy the Lagrangian systems* (3.37) *and* (3.39)*. If the Hessian* $\nabla_x^2 L$ *is positive definite, these conditions are also sufficient in which case one can conclude that the vectors* $x_k^{(m)}$ *with* $k = 1, \ldots, N$ *are indeed the desired singular vectors of order* m *and that* σ_m *is the corresponding singular value.*

Remark 3.8.4. *An algorithm for the computation of successive rank one approximations* $W^{(r)}$ *of* $W \in \mathcal{T}_N$ *as defined in Definition 3.6.2 is immediate from Algorithm TSVD. Indeed, for the computation of a rank-1 optimal approximant, only steps 1,2,3 of the TSVD algorithm are relevant. First apply the TSVD algorithm on* $E_0 := W$ *to result in the optimal approximation* W_1^**. Then repeat the TSVD algorithm on the error tensor* $E_m = E_{m-1} - W_m^*$ *for* $m \leq r$ *to define* (3.27)*.*

To investigate convergence properties of the TSVD algorithm further, let G_W^p denote the pth power of the operator G_W, i.e., the pth iterate of $x^{i+1} = G_W x^i$ with initial condition x^0 in (3.46) satisfies $x^p = G_W^p x^0$. The following theorem shows convergence of the above sequential series of iterations to the exact singular vectors and singular values of the tensor W.

Theorem 3.8.5. *Suppose that* $G_W : \mathbb{R}^{L+1} \to \mathbb{R}^{L+1}$ *maps a closed subset* $\mathcal{D} \subset \mathbb{R}^{L+1}$ *into itself and that*

$$\|G_W^p x - G_W^p y\| \leq \alpha_p \|x - y\|, \quad \text{for all } x, y \in \mathcal{D}, \quad p = 1, 2, \ldots \tag{3.47}$$

where $\beta = \sum_{p=1}^{\infty} \alpha_p < \infty$*. Then for every* $m = 1, \ldots K$*, with* $K = \min_k \operatorname{modrank}(W)$*, the operator* G_{W_m} *has a unique fixed point* $x^* \in \mathcal{D}$ *(depending on* m*) and the iteration* (3.46) *converges to* x^* *as* $i \to \infty$*. Moreover, every*

iterate establishes the error estimate

$$\|x^i - x^*\| \leq \beta \|x^i - x^{i-1}\|, \quad i = 1, 2, \ldots$$

and the components $\varphi_k^{(m)}$, $k = 1, \ldots, N$ *and* σ_m *of the fixed point* x^* *are extremal values of the optimizations* (3.16) *(if* $m = 1$*) and* (3.18) *(for* $1 < m \leq K$*).*

The proof of the above theorem is an application of Theorem 12.1.1 in [59] combined with the observation that the inequality (3.47) holds with G_W replaced by G_{W_k} with $k > 1$ whenever (3.47) for $G_W = G_{W_1}$. In particular, this observation makes the convergence rate β independent of k.
In practice it is not trivial to explicitly verify whether G_W satisfies (3.47) An interesting special case of Theorem 3.8.5 applies to tensors W for which G_W maps a closed subset $\mathcal{D} \subset \mathbb{R}^{L+1}$ into itself and is contractive on $\mathcal{D}$ in the sense that there exists $\alpha < 1$ such that $\|G_W x - G_W y\| \leq \alpha \|x - y\|$ for all $x, y \in \mathcal{D}$. In that case, the result of Theorem 3.8.5 simplifies to the contraction mapping theorem for nonlinear operators. Specifically, if G_W is contractive, (3.47) holds with $\alpha_p = \alpha^p$ and $\beta = \frac{\alpha}{1-\alpha}$ defines the convergence rate. This means that under the contractivity condition of G_W, the sequence (3.46) in step 2 of the TSVD algorithm converges to the unique fixed point of G_{W_m} in $\mathcal{D}$ whenever $x^0 \in \mathcal{D}$. More refinements of convergence conditions go in the direction of transforming G_W into $G'_W := TG_W T^{-1}$ where a suitable homeomorphism $T : \mathbb{R}^{L+1} \to \mathbb{R}^{L+1}$ is chosen so as to make G'_W contractive, or, alternatively, to study *iterated contractions* of the form $\|G_W(G_W x) - G_W x\| \leq \alpha \|G_W x - x\|$ where $\alpha < 1$ and $x \in \mathcal{D}$. We refer to [59] for more details.
Theorem 3.8.5 promises that whenever the algorithm converges, it converges to an extremal solution to the optimization problems (3.16) and (3.18) that define the singular value and singular vectors of order m. Here, by 'extremal solutions' we mean that the fixed points of G_{W_m} satisfy the *first order necessary conditions* for the optimal solution of the maximization problems formulated in (3.18). Solutions to the optimization problems (3.16) and (3.18) satisfy these conditions but we can not guarantee that the iterated map (3.46) converges to a fixed point x^* of G_{W_m} that also satisfies the *sufficient conditions* for the optima. This means that if G_{W_m} is contractive, the algorithm converges to a fixed point $x^* = \mathrm{col}(\varphi_1^{(m)}, \ldots, \varphi_N^{(m)}, \sigma_m)$ where the N-tuple $(\varphi_1^{(m)}, \ldots, \varphi_N^{(m)}) \in \mathcal{S}^{(m)}$ and where the gradient of the cost function $|W(x_1, \ldots, x_N)|$ vanishes in $(\varphi_1^{(m)}, \ldots, \varphi_N^{(m)})$.

Remark 3.8.6. *A numerical algorithm for the computation of a dedicated singular value decomposition requires a minor change to the TSVD algorithm. Indeed, if* $W \in \mathcal{T}_N$ *with* $\mathcal{X}' = \mathcal{X}_1 \times \ldots \times \mathcal{X}_i$ *and* $\mathcal{X}'' = \mathcal{X}_{i+1} \times \ldots \times \mathcal{X}_N$*. The dedicated singular value*

decomposition in Definition 3.7.1 is numerically calculated from the TSVD algorithm in which the definition of $Q_k^{(m)}$ in step 3 is replaced by

$$Q_k^{(m)} := \begin{cases} I - \Phi_k^{(m)}[\Phi_k^{(m)}]^\top & 1 \leq k \leq i \\ I & i+1 \leq k \leq N \end{cases}.$$

Remark 3.8.7. *In this section we have derived some convergence properties regarding the TSVD algorithm. When conducting numerical computations of modal rank approximations to tensors one experiences difficulties due to local optimal points. The reader is referred to [42] for an overview of this phenomenon and its implications. We did not investigate the implications of local optima for our algorithm.*

3.9 Numerical Example

To illustrate the methods discussed in this chapter, we consider a data compression problem in 3-D imaging[1]. The data consists of pixel intensities of an MRI scan of a human head in which each of the L_3 slices is an image of $L_1 \times L_2$ pixels. The original MRI scan has dimensions $L_1 = 262$, $L_2 = 262$ and $L_3 = 29$, consists of 1990676 pixels which corresponds to 2MB of storage. All pixel intensities are stored in an $L_1 \times L_2 \times L_3$ tensor $W \in \mathcal{T}_3$ of modal rank $\mathrm{modrank}(W) = (243, 199, 29)$. The 10th slice of the original data is shown in Fig. 3.2. We consider two kinds of approximations to this data. First, we discuss approximations of W by tensors of modal rank $r = (r_1, r_2, r_3)$. Second, we review approximations by tensors of modal rank $r = (r_1, r_2, L_3)$, i.e., only the first and second mode dimensions are approximated. For both types of approximations we compare the Higher-Order Singular Value Decomposition (HOSVD) [24], the Tensor SVD as introduced in this chapter and the method of Successive Rank-one approximations discussed in Section 3.6.1. We aim for a drastic compression of the data. All simulations discussed in this section have been carried out with an accuracy setting of $\varepsilon_{\text{tol}} = 1 \cdot 10^{-6}$ in the TSVD algorithm. Implementations of all algorithms use the tensor toolbox for Matlab [4]. More numerical examples can be found in [6].

[1]The data was obtained from TU/e-BME, Biomedical Image Analysis, in collaboration with Prof. Dr. med. Berthold Wein, Aachen, Germany

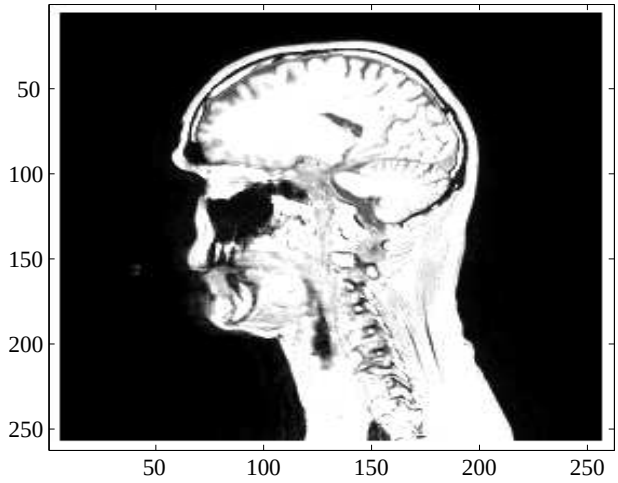


Figure 3.2: 10th slice of the original data.

3.9.1 Approximations of the form (r_1, r_2, r_3)

Table 3.1 shows the relative approximation errors in Frobenius norm that were obtained with HOSVD, Successive Rank-One and Tensor SVD respectively. From this it is obvious that HOSVD and Successive Rank-One give comparable approximation errors and outperform the Tensor SVD. The data compression in these approximations is substantial. Indeed, the modal rank approximation with $r = (10, 10, 10)$ implies a core tensor of 1000 elements, which is 0.05% of the number of entries in W and therefore amounts to a storage reduction from 2MB to 1KB. Table 3.2 lists the number of iterations i^* required in step 2 of the TSVD algorithm for the computation of the Tensor SVD and the Successive Rank-one approximations.

Due to space limitations only the number of iterations i^* for the first five steps of the respective algorithms are shown. The number of iterations seems to be increasing as the algorithm progresses, but this is pure coincidence. The number of iterations decreases and increases quite randomly. The time to compute the first five singular values and sets of singular vectors for this example was 74.05 seconds on a 1.83 GHz Intel Duo Processor T2400. The first five successive rank-one approximations have been computed in 46.21 seconds on the same PC.

Table 3.1: Relative Approximation Error, $\frac{\|W-W_r\|_F}{\|W\|_F}$.

(r_1, r_2, r_3)	HOSVD	Succ.R1	TSVD
$(1, 1, 1)$	0.5181	0.5181	0.5181
$(3, 3, 3)$	0.2648	0.2647	0.5126
$(5, 5, 5)$	0.2334	0.2306	0.5108
$(7, 7, 7)$	0.2111	0.2071	0.4280
$(10, 10, 10)$	0.1869	0.1857	0.4265

Table 3.2: Number of iterations i^* of (3.46).

	TSVD		Succ. Rank 1
σ_1	20	T_1^*	20
σ_2	56	T_2^*	13
σ_3	80	T_3^*	34
σ_4	130	T_4^*	95
σ_5	262	T_5^*	223

3.9.2 Approximations of the form $r = (r_1, r_2, L_3)$

In applications it may be desirable to leave the mode rank unchanged for one or more modal directions. For example, when considering spatial-temporal data, one may be interested in approximating spatial information only. To this end, this section gives simulation results for approximations of the form $r = (r_1, r_2, L_3)$ for the MRI data.
As in the previous section, a comparison is made between the HOSVD, Tensor SVD and the method of successive rank-one approximations. Furthermore, results are also included for the dedicated Tensor SVD as introduced in Sec. 3.7.
The numerical results of the computation of generic and dedicated singular values can be found in Table 3.3. From this table it is clear that the generic singular values decay much faster than the dedicated singular values. This would imply that the generic singular vectors give better results in approximation. However, examination of Table 3.4, which lists the approximation errors, shows that exactly the opposite is the case. Using the dedicated singular vectors for approximation gives approximation errors that are

much smaller than those obtained when using the generic singular vectors. A possible explanation for this is that since the dedicated singular value decomposition uses less constraints, more information of the original data is captured in the decomposition. Hence the larger dedicated singular values and the better approximations.

Table 3.3: Generic and dedicated singular values.

σ_1	102773.20	$\hat{\sigma}_1$	102773.20
σ_2	5815.49	$\hat{\sigma}_2$	49916.03
σ_3	3265.52	$\hat{\sigma}_3$	19275.82
σ_4	1948.73	$\hat{\sigma}_4$	11779.07
σ_5	1489.41	$\hat{\sigma}_5$	9920.99

Table 3.4: Relative Approximation Error, $\frac{\|W-W_r\|_F}{\|W\|_F}$, for approximations of the form (r_1, r_2, L_3).

(r_1, r_2)	HOSVD	Succ. Rank 1	TSVD	Dedicated TSVD
$(1,1)$	0.5181	0.5181	0.5181	0.5181
$(3,3)$	0.2647	0.2646	0.5126	0.2646
$(5,5)$	0.2331	0.2305	0.5108	0.2305
$(7,7)$	0.2108	0.2069	0.4280	0.2070
$(10,10)$	0.1868	0.1856	0.4265	0.1872

Figures 3.3 and 3.4 show the 10th slice of the rank-$(10, 10, 29)$ approximations to the original data. The figures illustrate the numbers given in Table 3.4. The figures show clearly that the performance of the HOSVD, successive rank-one approximations and dedicated TSVD is equivalent for this specific example. All three methods significantly outperform the TSVD.

3.10 Conclusions

This chapter considered the problem of finding low-rank approximations to tensors. We have formally introduced tensors and tensor concepts in a coordinate-free man-

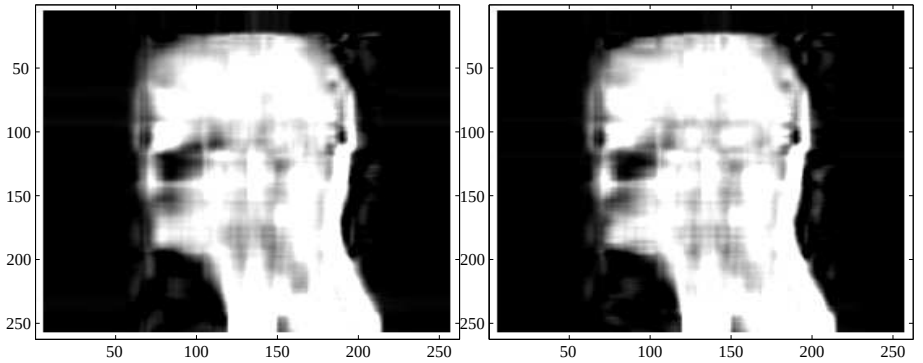


Figure 3.3: 10th slice of rank-$(10, 10, 29)$ approximant, computed using HOSVD (left) and Successive Rank-One approximations (right).

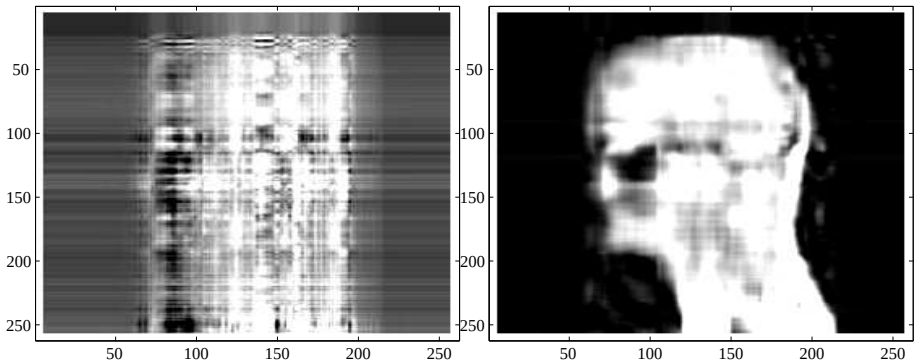


Figure 3.4: 10th slice of rank-$(10, 10, 29)$ approximant, computed using TSVD (left) and modified TSVD (right).

ner. Furthermore, we have given a brief overview of the field of tensor decompositions and defined the necessary concepts such as tensor rank and low-rank approximations. Then, we have presented a new method for the computation of modal rank approximations to tensors. We have given a thorough analysis of the properties of this method, referred to as TSVD, and presented a number of low-rank approximation results. We have defined an adaptation to the TSVD which may give better approximation results when not all modal directions are approximated. Finally, we have presented a numerical algorithm to compute the decomposition and analyzed its convergence properties. The chapter concludes with a numerical example. In the numerical example that was presented, the TSVD method was compared to the best-known existing modal rank approximation method, the HOSVD. The work presented in this chapter has previously appeared in [81, 7]. In future, a more thorough numerical analysis of the decomposition method is necessary, specifically with respect to the influence of local optima [42].

Chapter 4

Generalization of POD

4.1 Introduction

POD is a model reduction method that is applicable to systems, linear or non-linear, that evolve both over space and time. Specifically, Galerkin projections are used to derive approximate models. According to the definition of a Galerkin projection, the projection spaces for the signal and residual projection are chosen to be equal. The element that distinguishes POD from other Finite Element methods is the fact that the projection spaces are computed from measured or simulated process data. POD is used in many application areas such as fluid dynamics, process control and reservoir modeling.

As mentioned above, POD is a projection-based method that relies on the computation of empirical projection spaces from a representative set of measurement or simulation data. In its classical formulation, the projection spaces are used in a spectral expansion that separates space and time. No further structure is assumed for the spatial variables. This makes POD basically a two-variable method since it deals with an ND system by separating time and space. That is, the independent variables are assumed to reside in a Cartesian product of a temporal and spatial domain.

There are a number of limitations concerning the application of POD to true large-scale systems. Firstly, due to the Galerkin projection, when POD is applied to non-linear PDEs the computation time does not decrease significantly. This is because the original nonlinear equations still have to be evaluated at each time-step. This issue was addressed in [3, 17]. Secondly, in most model reduction applications, a separation is made between space and time, but no further structure is assumed for the spatial domain. In particular, for larger dimensional Euclidean geometries, all spatial

variables are lumped and this yields a large-dimensional data correlation operator Φ. In this way, the ND nature of the original problem is replaced by an artificial 2-D structure with space and time as independent variables. Thirdly, when multiple dependent variables, i.e. multiple physical quantities, are considered, the reduced model becomes very sensitive to scaling of these dependent variables.
In this chapter we propose an adaptation of POD to deal with the latter two of these issues. We will assume a more general Cartesian structure for the independent variables. This allows changes to be made to POD that allow, at least in principle, more flexibility in defining approximate models.
This chapter is structured as follows. First, we introduce the POD method as it is currently used. Then, we propose an adaptation which assumes a Cartesian structure of the independent variables. This allows projection spaces to be computed using the tensor decompositions methods discussed in the previous chapter. We discuss the spectral expansion and Galerkin projection that follow from the assumption of a Cartesian structure for the independent variables. The method proposed in this chapter is illustrated using two numerical examples. The material presented in this chapter was published in [7, 8].

4.2 Proper Orthogonal Decompositions

This section offers a brief introduction to the method of Proper Orthogonal Decompositions. More information on this topic can be found in [56, 52, 74].
Consider an arbitrary linear distributed system described by the following Partial Difference Equation

$$D(\varsigma_1, \varsigma_1^{-1}, \ldots, \varsigma_N, \varsigma_N^{-1})\underline{w} = 0. \tag{4.1}$$

Here $D \in \mathbb{R}^{\cdot\times n}[\xi_1, \ldots, \xi_N, \eta_1, \ldots, \eta_N]$ is a real matrix-valued polynomial in $2N$ indeterminates and ς_k (ς_k^{-1}) is the forward (backward) shift operator acting on the spatial discretization in the kth mode according to Definition A.2.1. The domain of the signal $\underline{w}$, $\mathbb{D}$, is assumed to have a Cartesian structure $\mathbb{D} = \mathbb{X} \times \mathbb{T}$, which is typically the product of a spatial and a temporal domain. Solutions $\underline{w}$ to this PDE assume the form $\underline{w} : \mathbb{X} \times \mathbb{T} \to \mathbb{R}$ where both $\mathbb{X}$ and $\mathbb{T}$ are sets of finite cardinality, say L_X and L_T, respectively. Specifically, we consider $\mathbb{X} = \{p_x^{(1)}, \ldots, p_x^{(L_X)}\}$ and $\mathbb{T} = \{p_t^{(1)}, \ldots, p_t^{(L_T)}\}$.
The POD method consists of three steps. First, the signal $\underline{w}$ is approximated by a signal $\underline{w}_r$ using a spectral expansion. Second, the reduced model is defined to consist of the signals $\underline{w}_r$ that satisfy a Galerkin projection. Third, the projection spaces in this Galerkin projection are empirical projection spaces, i.e. they are computed from measured or simulated process data.

Whenever $n > 1$, two different types of spectral expansions can be found in literature. These types will be referred to by *single-variable* and *lumped-variable* expansions. In single-variable expansions, each element of the vector-valued signal $\underline{w}$ is expanded separately. In lumped-variable expansions, the signal $\underline{w}$ is expanded as a whole, as we will show below.

4.2.1 Spectral Expansions

Lumped-variable expansions

Let $\mathcal{X}$ be the space of functions $f : \mathbb{X} \to \mathbb{R}^n$ with the following inner product

$$\langle \underline{f}, \underline{g} \rangle_{\mathcal{X}} = \sum_{k=1}^{L_X} \langle \underline{f}(p_x^{(k)}), \underline{g}(p_x^{(k)}) \rangle \tag{4.2}$$

for all $\underline{f}, \underline{g} \in \mathcal{X}$, where $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product in $\mathbb{R}^n$.
Solutions $\underline{w}$ of (4.1) are assumed to satisfy $\underline{w}(\cdot, p_t) \in \mathcal{X}$ for all $p_t \in \mathbb{T}$. Let $\{\underline{\varphi}_k,\ k = 1, 2, \ldots, L_X\}$ be an orthonormal basis for $\mathcal{X}$. Then, every solution $\underline{w}$ to (4.1) admits a spectral expansion

$$\underline{w}(p_x, p_t) = \sum_{k=1}^{L_X} a_k(p_t) \underline{\varphi}_k(p_x).$$

In this expansion the modal coefficients a_k are uniquely determined by $a_k(p_t) = \langle \underline{w}, \underline{\varphi}_k \rangle_{\mathcal{X}}$ for all $p_t \in \mathbb{T}$. For $0 < r < L_X$, a low-rank approximation to $\underline{w}$ is defined by the truncation

$$\underline{w}_r(p_x, p_t) = \sum_{k=1}^{r} a_k(p_t) \underline{\varphi}_k(p_x) \tag{4.3}$$

for all $p_x \in \mathbb{X}$ and $p_t \in \mathbb{T}$.

Single-variable expansions

In single-variable expansions each component of $\underline{w}$ is expanded individually. Specifically, for each of the components w_k, $k = 1, \ldots, n$, of $\underline{w} = [w_1, \ldots, w_n]^\top$ it is assumed that, for any time instant $p_t \in \mathbb{T}$, the component function $w_k(\cdot, p_t)$ belongs to a Hilbert space $\mathcal{X}_k$ of functions mapping $\mathbb{X} = \mathbb{R}^{L_X}$ to $\mathbb{R}$ with the following inner product

$$\langle f, g \rangle_{\mathcal{X}_k} = \sum_{k=1}^{L_x} f(p_x^{(k)}) g(p_x^{(k)}), \quad \forall f, g \in \mathcal{X}_k. \tag{4.4}$$

Let $\varphi_k^{(\ell_k)} : \mathbb{X} \to \mathbb{R}$ is a (countable) orthonormal set of basis functions of $\mathcal{X}_k$, then $\underline{w}$ admits an expansion of the form

$$\underline{w}(p_x, p_t) = \begin{bmatrix} \sum_{\ell_1} a_{\ell_1}^{(1)}(p_t)\varphi_1^{(\ell_1)}(p_x) \\ \vdots \quad \vdots \\ \sum_{\ell_n} a_{\ell_n}^{(n)}(p_t)\varphi_n^{(\ell_n)}(p_x) \end{bmatrix}$$

Here, the coefficients are uniquely determined by $a_{\ell_k}^{(k)}(p_t) = \langle w_k(\cdot, p_t), \varphi_k^{(\ell_k)} \rangle_j$.
If $r = (r_1, \ldots, r_n)$ is a vector of integers then the *truncated expansion* of order r is defined by the signal $\underline{w}_r(p_x, p_t)$ whose kth entry is given by the finite expansion

$$w_r^{(k)}(p_x, p_t) = \sum_{\ell_k=1}^{r_k} a_{\ell_k}^{(k)}(p_t)\varphi_k^{(\ell_k)}(p_x).$$

4.2.2 POD basis choice

POD basis choice for lumped-variable expansions

Clearly, the quality of the reduced order model (4.9) entirely depends on the choice of basis functions $\{\varphi_k\}$. In the POD method, the orthonormal basis functions $\{\varphi_k\}$ of $\mathcal{X}$ are determined empirically, either from measurements or data $w : \mathbb{D} \to \mathbb{R}$ simulated from (4.1). This set of measured or simulated data is assumed to contain a collection of trajectories that is representative of the system dynamics of interest. Specifically, for a Cartesian domain $\mathbb{D} = \mathbb{X} \times \mathbb{T}$ and given data $w : \mathbb{D} \to \mathbb{R}$ with $w(\cdot, p_t) \in \mathcal{X}$ and $p_t \in \mathbb{T}$, the basis functions φ_k are chosen so as to minimize the criterion function

$$J(\varphi_1, \ldots, \varphi_r) := \sum_{m=1}^{L_T} \|w(\cdot, p_t^{(m)}) - \sum_{k=1}^{r} \langle w(\cdot, p_t^{(m)}), \varphi_k \rangle \varphi_k\|^2 \tag{4.5}$$

subject to the constraint that

$$\langle \varphi_k, \varphi_m \rangle = \begin{cases} 1 & \text{if } k = m \\ 0 & \text{if } k \neq m \end{cases} \quad k, m = 1, \ldots, r. \tag{4.6}$$

Here, the inner product is the inner product of the Hilbert space $\mathcal{X}$ and the optimization is carried out for an arbitrary approximation degree r. The characterization of the POD basis that follows is valid for arbitrary Hilbert spaces $\mathcal{X}$, that may be infinite dimensional. This applies to this part of the chapter only.

Definition 4.2.1 (POD basis or order r)**.** *A POD basis of order r is defined to be a collection of functions $\{\varphi_1, \ldots, \varphi_r\}$, $\varphi_k : \mathbb{X} \to \mathbb{R}^n$, that minimizes the criterion* (4.5) *subject to the constraint* (4.6)*. Hence, a POD basis of order r minimizes the total error $\sum_{\mathbb{T}} \|w - w_r\|^2$ over all rank r approximations $\underline{w}_r$ of $\underline{w}$ of the form* (4.3)*.*

Definition 4.2.2 (POD basis)**.** *Let $\mathbb{I}$ be a countable set of indices with cardinality equal to the dimension of $\mathcal{X}$. A complete orthonormal basis $\{\varphi_k, k \in \mathbb{I}\}$ of $\mathcal{X}$ is said to be a* POD basis *if for all r the collection $\{\varphi_1, \ldots, \varphi_r\}$ is a POD basis of order r.*

The constrained optimization problem (4.5) has an elegant solution in terms of the data correlation operator $\Phi : \mathcal{X} \to \mathcal{X}$ that is implicitly defined as

$$\langle \psi_1, \Phi\psi_2 \rangle := \sum_{m=1}^{L_T} \langle \psi_1, w(\cdot, p_t^{(m)}) \rangle \cdot \langle w(\cdot, p_t^{(m)}), \psi_2 \rangle \quad \psi_1, \psi_2 \in \mathcal{X}. \tag{4.7}$$

Note that, Φ is a well defined linear, bounded, self-adjoint and non-negative operator on $\mathcal{X}$.

Theorem 4.2.3. *Suppose that $\{\varphi_k, k \in \mathbb{I}\}$ is an orthonormal basis of $\mathcal{X}$ and suppose that the eigenvalues of Φ are absolute summable. Then $\{\varphi_k, k \in \mathbb{I}\}$ is a POD basis if and only if $\Phi\varphi_k = \lambda_k\varphi_k$, $k \in \mathbb{I}$ where the eigenvalues λ_k are ordered according to $\lambda_1 \geq \lambda_2 \geq \cdots$. Moreover, in that case the error*

$$J(\varphi_1, \ldots, \varphi_r) = \sum_{k>r} \lambda_k$$

and is minimal for all truncation levels $r > 0$.

Hence, the eigenfunctions of the data correlation operator determine the POD basis.

Proof. If eigenvalues of Φ are absolute summable, Φ is self-adjoint and nuclear. This means that it admits a representation $\Phi = \sum_{k=1}^{N} \lambda_k \langle \psi_k, \cdot \rangle \psi_k$ where $1 \leq N \leq \infty$, the eigenvalues λ_k are positive, non-increasingly ordered and summable, and the eigenfunctions $\{\psi_k, k = 1, \ldots, N\}$ are orthonormal in $\mathcal{X}$. Moreover, for any orthonormal

basis $\{\varphi_k, k \in \mathbb{I}\}$ of $\mathcal{X}$ we have

$$\begin{aligned} J(\varphi_1, \ldots, \varphi_r) &= \sum_{m=1}^{L_T} \langle w - w_r, w - w_r \rangle \\ &= \sum_{m=1}^{L_T} \left\langle \sum_{k>r} \langle w(\cdot, p_t^{(m)}), \varphi_k \rangle \varphi_k, \sum_{k>r} \langle w(\cdot, p_t^{(m)}), \varphi_k \rangle \varphi_k \right\rangle \\ &= \sum_{m=1}^{L_T} \sum_{k>r} \left\langle w(\cdot, p_t^{(m)}), \varphi_k \right\rangle \cdot \left\langle w(\cdot, p_t^{(m)}), \varphi_k \right\rangle \\ &= \sum_{k>r} \sum_{m=1}^{L_T} \langle w(\cdot, p_t^{(m)}), \varphi_k \rangle \cdot \langle w(\cdot, p_t^{(m)}), \varphi_k \rangle \\ &= \sum_{k>r} \langle \varphi_k, \Phi \varphi_k \rangle. \end{aligned}$$

Now first suppose that $\varphi_k = \psi_k$ for $k = 1, \ldots, N$. Then, $\{\varphi_k, k = 1, \ldots, N\}$ is an orthonormal set of eigenfunctions of Φ and $J(\varphi_1, \ldots, \varphi_r) = \sum_{k>r} \lambda_k$ is finite and minimal for all r by the monotonicity of the sequence λ_k. Hence, $\{\varphi_k, k = 1, \ldots, r\}$ is a POD basis of order r for any r. Second, for any POD basis $\{\varphi_k, k \in \mathbb{I}\}$ the above expression for the error implies that

$$J(\varphi_1, \ldots, \varphi_r) = \sum_{k>r} \langle \varphi_k, \sum_{m=1}^{N} \lambda_m \langle \psi_m, \varphi_k \rangle \psi_m \rangle = \sum_{k>r} \sum_{m=1}^{N} \lambda_m \langle \varphi_k, \psi_m \rangle^2$$

which is minimal for all r only if $\langle \varphi_k, \psi_m \rangle = \delta_{k,m}$ for all integers k, m between 1 and N. But then it is immediate that $\{\varphi_k, k = 1, \ldots, N\}$ is also a set of orthonormal eigenvectors of Φ.

□

In the finite-dimensional discrete case, the data correlation operator Φ becomes a symmetric non-negative definite matrix

$$\Phi = W_{\text{snap}} W_{\text{snap}}^{\top} \tag{4.8}$$

where $W_{\text{snap}} \in \mathbb{R}^{L_X \times L_T}$ is a matrix that contains trajectories of the system (4.1) that

have been obtained either from measurements or simulations. W_{snap} is given by

$$\begin{aligned} W_{\text{snap}} &= [\underline{w}(p_x, p_t^{(1)}), \cdots, \underline{w}(p_x, p_t^{(L_T)})] \\ &= \begin{bmatrix} \underline{w}(p_x^{(1)}, p_t^{(1)}) & \cdots & \underline{w}(p_x^{(1)}, p_t^{(L_T)}) \\ \vdots & & \vdots \\ \underline{w}(p_x^{(L_X)}, p_t^{(1)}) & & \underline{w}(p_x^{(L_X)}, p_t^{(L_T)}) \end{bmatrix} \end{aligned}$$

POD basis functions can now be computed via the eigenvalue decomposition of Φ, i.e. the decomposition $\Phi = U\Lambda U^\top$. Specifically, the eigenvectors of Φ, stored in the columns of U, form the POD basis. These eigenvectors are equal to the left singular vectors of W_{snap}. Consider that the SVD of W_{snap} is given by $W_{\text{snap}} = U\Sigma V^\top$, then the following holds

$$\begin{aligned} W_{\text{snap}} W_{\text{snap}}^\top &= U\Sigma V^\top V \Sigma^\top U^\top \\ &= U\left(\Sigma\Sigma^\top\right) U^* \end{aligned}$$

Hence, computation of POD basis functions via the eigenvalues of the correlation matrix is equivalent to computation of the Singular Value Decomposition of the snapshot matrix.

POD basis choice for single-variable expansions

Similar to the lumped-variable case, for the single variable expansion a data correlation operator $\Phi_k : \mathcal{X}_k \to \mathcal{X}_k$ is defined for each $k = 1, \ldots, n$ with respect to $\underline{w}$ according to

$$\langle \psi_1, \Phi_k \psi_2 \rangle := \sum_{m=1}^{L_T} \langle \psi_1, w_k(\cdot, p_t^{(m)}) \rangle \langle \psi_2, w_k(\cdot, p_t^{(m)}) \rangle$$

for $\psi_1, \psi_2 \in \mathcal{X}_k$. Then Φ_k is a well defined linear, bounded, self-adjoint and non-negative operator on $\mathcal{X}_k$. The collection $\{\varphi_k^{(\ell)} \mid \ell = 1, 2, \ldots\}$ of ordered normalized eigenfunctions of Φ_k then defines an orthonormal basis of a subspace in $\mathcal{X}_k$. That is, let $\varphi_k^{(\ell)} : \mathbb{X} \to \mathbb{R}$ be the function that satisfies $\|\varphi_k^{(\ell)}\| = 1$ and

$$\Phi_k \varphi_k^{(\ell)} = \lambda_\ell \varphi_k^{(\ell)}$$

where λ_ℓ is the ℓth largest eigenvalue of Φ_k. Then $\{\varphi_k^{(\ell)}\}$ is a collection of orthonormal functions provided that the eigenvalues λ_ℓ are disjoint (for non-disjoint eigenvalues the eigenfunctions of Φ_k can be chosen to be orthonormal). This specific basis

is optimal for the given data in the sense that $\sum_{m=1}^{L_T} \|w_k(\cdot, p_t^{(m)}) - w_r^{(k)}(\cdot, p_t^{(m)})\|$ is minimal for all truncation levels r_k and for all $k = 1, \ldots, n$.
Since $\mathbb{X}$ consists of L_X disjoint samples, the spaces $\mathcal{X}_k$ are L_X-dimensional and Φ_k is a non-negative definite matrix of dimension $L_X \times L_X$ defined by $\Phi_k = W_k W_k^\top$ where $[W_k]_{\ell_1,\ell_2} = w_k(p_x^{(\ell_1)}, p_t^{(\ell_2)})$ is sometimes referred to as a *snapshot matrix*

$$W_k = \begin{bmatrix} w_k(p_x^{(1)}, p_t^{(1)}) & \cdots & w_k(p_x^{(1)}, p_t^{(L_T)}) \\ \vdots & & \vdots \\ w_k(p_x^{(L_X)}, p_t^{(1)}) & & w_k(p_x^{(L_X)}, p_t^{(L_T)}) \end{bmatrix}.$$

POD basis functions can now be computed via the eigenvalue decomposition of Φ_k, or the SVD of W_k, as in the lumped-variable case.

4.2.3 Galerkin projecion

Galerkin projection for lumped-variable expansion

For $r > 0$, the reduced order model is then defined by the collection of solutions $\underline{w}_r(p_x, p_t) = \sum_{k=1}^{r} a_k(p_t)\varphi_k(p_x)$ that satisfy the Galerkin projection

$$\left\langle D(\varsigma_1, \varsigma_1^{-1}, \ldots, \varsigma_N, \varsigma_N^{-1})\underline{w}_r, \varphi \right\rangle = 0, \quad \forall \varphi \in \mathcal{X}_r \tag{4.9}$$

where $\mathcal{X}_r$ is the finite dimensional projection space $\mathcal{X}_r = \operatorname{span}\{\varphi_1, \ldots, \varphi_r\}$. Whenever the spectral expansion of $\underline{w}_r$ is substituted in (4.9), (4.9) becomes a system of r ordinary difference equations in the modal coefficients a_k, $k = 1, \ldots, r$. This reduces the PDE (4.1) to an approximate model of r ordinary differential equations.

Galerkin projection for single-variable expansion

In the single-variable case, the reduced model is defined for each row of (4.1) as follows. Consider a vector of integers $r = (r_1, \ldots, r_n)$. For $1 \leq k \leq n$, the reduced order model is defined by the collection of solutions $w_r^{(k)}(p_x, p_t) =$
$\sum_{\ell_k=1}^{r_k} a_{\ell_k}^{(k)}(p_t)\varphi_k^{(\ell_k)}(p_x)$ that satisfy

$$\langle [D(\varsigma_1, \varsigma_1^{-1}, \ldots, \varsigma_N, \varsigma_N^{-1})\underline{w}_r]^{(k)}, \varphi \rangle = 0, \quad \forall \varphi \in \operatorname{span}\{\varphi_k^{(1)}, \ldots, \varphi_k^{(r_k)}\} \tag{4.10}$$

Again, the spectral expansion of $\underline{w}_r$ is substituted in (4.10),this becomes a system of r_k ordinary differential equations in $a_{\ell_k}^{(k)}$.

4.3 Adaptation of POD

The previous section introduced the method of Proper Orthogonal Decompositions as it can be found in literature. As mentioned in the introduction to this chapter, there are a couple of issues when applying POD to large-scale systems. Therefore, we describe an adaptation of the method in this section. Instead of the separation of time and space, we propose a more general Cartesian structure of the independent variables. This structure allows to define an alternative spectral expansion, Galerkin projection and projection basis. Specifically, we assume that the domain $\mathbb{X}$ of (4.1) has Cartesian structure $\mathbb{X} = \mathbb{X}_1 \times \cdots \times \mathbb{X}_N$. Furthermore, we assume that each $\mathbb{X}_k$ has finite cardinality L_k, i.e. $\mathbb{X}_k = \{p_k^{(1)}, \ldots, p_k^{(L_k)}\}$ for $k = 1, \ldots, N$ and let $\mathcal{X}_k := \mathbb{R}^{L_k}$ be equipped with the standard Euclidean inner product. Finally, we assume $Y := \mathbb{R}^n$ to also be equipped with the standard Euclidean product.
The proposed approach can be summarized as follows. The assumption of a Cartesian domain $\mathbb{X}$ allow a tensor $W : \mathcal{X}_1 \times \cdots \times \mathcal{X}_N \times \mathcal{Y} \to \mathbb{R}$ to be associated with the signal $\underline{w}$. This also means that we can associate a tensor with the measured or simulated data that will be used to compute a projection basis. Approximations to this tensor provide a projection basis for each of the independent variables. This leads to a spectral expansion which is an alternative to the single- and lumped-variable expansions introduced in the previous section. Since the spectral expansion and the projection basis have both changed, an adapted Galerkin projection is needed to be able to define a reduced model.

4.3.1 Spectral Expansion

The solution $\underline{w}$ of the difference equation (4.1) can be viewed as a mapping $\underline{w} : \mathbb{X}_1 \times \cdots \times \mathbb{X}_N \to \mathbb{R}^n$. Therefore, we can define the set $\hat{W}$ of mappings $\hat{W} : \mathcal{X}_1 \times \cdots \times \mathcal{X}_N \to \mathcal{Y}$. $\hat{W}$ is a map $\mathcal{T}_N \to \mathcal{T}_1$ and as described in Section 3.2.1, we can associate a tensor $W \in \mathcal{T}_{N+1}$ with $\hat{W}$ and therefore with $\underline{w}$. W is a tensor $W : \mathcal{X}_1 \times \cdots \times \mathcal{X}_N \times \mathcal{Y} \to \mathbb{R}$. W is represented with respect to the standard bases as follows

$$W = \sum_{\ell_1=1}^{L_1} \cdots \sum_{\ell_N=1}^{L_N} \sum_{\ell_{N+1}=1}^{n} w_{\ell_1 \cdots \ell_{N+1}} e_1^{(\ell_1)} \otimes \cdots \otimes e_{N+1}^{(\ell_{N+1})} \tag{4.11}$$

where the coefficients $w_{\ell_1 \cdots \ell_{N+1}}$ are defined as follows

$$w_{\ell_1 \cdots \ell_{N+1}} := \underline{w}_{\ell_{N+1}}(p_1^{(\ell_1)}, \ldots, p_N^{(\ell_N)}) \tag{4.12}$$

that is, $w_{\ell_1 \cdots \ell_{N+1}}$ takes the value of the ℓ_{N+1}-th element of $\underline{w}$ at grid-point $(p_1^{(\ell_1)}, \ldots, p_N^{(\ell_N)})$.

We will work towards a spectral expansion via a basis change for the tensor W as follows. Consider the following orthonormal bases for $\mathcal{X}_1, \ldots, \mathcal{X}_N, \mathcal{Y}$

$$\{\varphi_1^{(1)}, \ldots, \varphi_1^{(L_1)}\}, \ldots, \{\varphi_N^{(1)}, \ldots, \varphi_N^{(L_N)}\}, \{\varphi_{N+1}^{(1)}, \ldots, \varphi_{N+1}^{(n)}\}. \tag{4.13}$$

We now represent the tensor W with respect to (4.13), which leads to the following representation

$$\tilde{W} = \sum_{\ell_1=1}^{L_1} \cdots \sum_{\ell_{N+1}=1}^{n} \tilde{w}_{\ell_1 \cdots \ell_{N+1}} \varphi_1^{(\ell_1)} \otimes \cdots \otimes \varphi_{N+1}^{(\ell_{N+1})}. \tag{4.14}$$

$\tilde{w}_{\ell_1 \cdots \ell_{N+1}}$ denotes the elements of W with respect to the new bases (4.13), i.e
$\tilde{w}_{\ell_1 \cdots \ell_{N+1}} := W(\varphi_1^{(\ell_1)}, \ldots, \varphi_{N+1}^{(\ell_{N+1})})$
From now on, we assume that $\mathcal{X}_N$ refers to the time variable, which we do not wish to approximate in the remainder of the model reduction approach. We now define a spectral expansion for $\underline{w}$ as follows. Let one component, w_k, of the signal $\underline{w}$ be defined as follows

$$w_k(k_1, \ldots, k_N) := \tilde{W}(e_1^{(k_1)}, \ldots, e_N^{(k_N)}, e_{N+1}^{(k)}). \tag{4.15}$$

Now, $\underline{w}$ is defined as

$$\begin{aligned} \underline{w}(k_1, \ldots, k_N) = \sum_{\ell_1=1}^{L_1} \cdots \sum_{\ell_{N+1}=1}^{n} \tilde{w}_{\ell_1 \cdots \ell_{N+1}} \langle \varphi_1^{(\ell_1)}, e_1^{(k_1)} \rangle \\ \cdots \langle \varphi_N^{(\ell_N)}, e_N^{(k_N)} \rangle \begin{bmatrix} \langle \varphi_{N+1}^{(\ell_{N+1})}, e_{N+1}^{(1)} \rangle \\ \vdots \\ \langle \varphi_{N+1}^{(\ell_{N+1})}, e_{N+1}^{(n)} \rangle \end{bmatrix}. \end{aligned} \tag{4.16}$$

Now define coefficients $\underline{b}_{\ell_1 \cdots \ell_{N-1}}$ as follows

$$\underline{b}_{\ell_1 \cdots \ell_{N-1}}(p_N^{(k_N)}) := \begin{bmatrix} b_{\ell_1 \cdots \ell_{N-1}}^{(1)}(p_N^{(k_N)}) \\ \vdots \\ b_{\ell_1 \cdots \ell_{N-1}}^{(n)}(p_N^{(k_N)}) \end{bmatrix} \tag{4.17}$$

$$= \sum_{\ell_N=1}^{L_N} \tilde{w}_{\ell_1 \cdots \ell_N} \langle \varphi_N^{(\ell_N)}, e_N^{(k_N)} \rangle \begin{bmatrix} \langle \varphi_{N+1}^{(\ell_{N+1})}, e_{N+1}^{(1)} \rangle \\ \vdots \\ \langle \varphi_{N+1}^{(\ell_{N+1})}, e_{N+1}^{(n)} \rangle \end{bmatrix}. \tag{4.18}$$

Then we arrive at the following spectral expansion for $\underline{w}$

$$\underline{w}(p_1^{(k_1)}, \dots, p_N^{(k_N)}) = \sum_{\ell_1=1}^{L_1} \cdots \sum_{\ell_{N-1}=1}^{L_{N-1}} \underline{b}_{\ell_1 \cdots \ell_{N_1}}(p_N^{(k_N)}) \\ \langle \varphi_1^{(\ell_1)}, e_1^{(k_1)} \rangle \cdots \langle \varphi_{N-1}^{(\ell_{N-1})}, e_{N-1}^{(k_{N-1})} \rangle. \quad (4.19)$$

So now we have a spectral expansion for $\underline{w}$, with vector-valued coefficients and a scalar basis functions. A rank-$r = (r_1, \dots, r_{N-1})$ approximation of $\underline{w}$ is defined by truncation of the sums in (4.19).

Remark 4.3.1. *Due to practical motivations, we have chosen here not to incorporate approximations in time or the space of dependent variables $\mathcal{Y}$. From a mathematical point of view it is indeed possible to also approximate time and the space of dependent variables. One would need to consider the physical significance of such approximations.*

4.3.2 Projection basis

Assume a FE solution $\underline{w}_{\text{snap}}$ of (4.1) is available. We can associate a tensor $W_{\text{snap}} : \mathcal{X}_1 \times \cdots \times \mathcal{X}_N \times \mathcal{Y} \to \mathbb{R}$ with $\underline{w}_{\text{snap}}$ as outlined above. W_{snap} has the following representation with respect to the standard bases

$$W_{\text{snap}} = \sum_{\ell_1=1}^{L_1} \cdots \sum_{\ell_{N+1}=1}^{n} w_{\ell_1 \cdots \ell_{N+1}} e_1^{(\ell_1)} \otimes \cdots \otimes e_{N+1}^{(\ell_{N+1})}$$

where $w_{\ell_1 \cdots \ell_{N+1}}$ takes the value of the ℓ_{N+1}-th element of $\underline{w}$ on the sample point with index $(\ell_1, \dots, \ell_N)$.
Our aim here is to generalize the idea of a POD basis for spatial domains to a higher dimensional Euclidean product space. For this, following 3.5.1, the data dependent tensor W_{snap} is assumed to be decomposed in SVD from according to (3.20). Let

$$\mathcal{M}_k^{(r_k)} = \operatorname{span}\{\varphi_k^{(1)}, \dots, \varphi_k^{(r_n)}\}$$

for $k = 1, \dots, N$ and $r_k \leq L_k$ define a r_k-dimensional projection space in $\mathcal{X}_k$. Then, $\varphi_k^{(1)}, \dots, \varphi_k^{(r_k)}$ form an orthonormal basis of $\mathcal{M}_k^{(r_k)}$. Let W_r^* be the modal truncation of W_{snap} as defined in Def. 3.4.2. Then, a projection basis can be defined as follows.

Definition 4.3.2. *The vector of integers $r = (r_1, \ldots, r_N)$ with $r_n \leq R_n$ is said to achieve a relative approximation error $\epsilon > 0$ if*

$$\frac{\|W_{snap} - W_r^*\|_F}{\|W_{snap}\|_F} \leq \epsilon \tag{4.20}$$

In that case, we say that the basis functions $\{\varphi_k^{(1)}, \ldots, \varphi_k^{(r_k)}\}$ for $k = 1, \ldots, N$ constitute a generalized projection basis for the model (4.1) *derived from the tensor $W_{textsnap}$.*

4.3.3 Galerkin projection

In the previous subsections a spectral expansion and a projection basis have been defined. What remains to be done is the definition of a Galerkin projection concept that fits this framework. To avoid confusion because of complex index-notation, we restrict the discussion to scalar $\underline{w}$, i.e. from now on $n = 1$. That is, we first consider the following difference equation

$$D(\varsigma_1, \ldots, \varsigma_N)w = 0.$$

Th residual of this difference equation, $R := D(\varsigma_1, \ldots, \varsigma_N)w$ is a signal $w : \mathbb{X}_1 \times \cdots \times \mathbb{X}_N \to \mathbb{R}$. We associate a $\hat{D}$ with R as follows. Let $\hat{D} : \mathcal{X}_1 \times \cdots \times \mathcal{X}_N \to \mathbb{R}$ be represented as

$$\hat{D} = \sum_{\ell_1} \cdots \sum_{\ell_N} \hat{d}_{\ell_1 \ldots \ell_N} e_1^{(\ell_1)} \otimes \cdots \otimes e_N^{(\ell_N)}. \tag{4.21}$$

Here, given a point $(p_1^{(\ell_1)}, \ldots, p_N^{(\ell_N)})$ in the domain $\mathbb{X}$, coefficients $\hat{d}_{\ell_1 \ldots \ell_N}$ are defined as

$$\hat{d}_{\ell_1 \ldots \ell_N} = R(\ell_1, \ldots, \ell_N) = [D(\varsigma_1, \ldots, \varsigma_N)w](p_1^{(\ell_1)}, \ldots, p_N^{(\ell_N)}). \tag{4.22}$$

That is, $\hat{d}_{\ell_1 \ldots \ell_N}$ takes the value of the residual at grid-point $(p_1^{(\ell_1)}, \ldots, p_N^{(\ell_N)})$.
Given sets of projection basis functions $\left\{\varphi_k^{(\ell_k)}\right\}_{\ell_k=1}^{r_k}$ a Galerkin projection of the discrete time model $D(\varsigma_1, \ldots, \varsigma_N)w = 0$ is defined as

$$\langle \varphi_1^{(k_1)}, \cdots, \langle \varphi_{N-1}^{(k_{N-1})}, D(\varsigma_1, \ldots, \varsigma_N)w \rangle_{N-1} \cdots \rangle_1 = 0 \tag{4.23}$$

for $k_m = 1, \ldots, r_m$ and $m = 1, \ldots, N-1$. Here, the expression of nested inner products should be interpreted as follows. Firstly, $R = D(\varsigma_1, \ldots, \varsigma_N)w$ is to be associated with the tensor $\hat{D} : \mathcal{X}_1 \times \cdots \times \mathcal{X}_N \to \mathbb{R}$ as above so that $\hat{D}_{N-1} :=$

$\langle \varphi_{N-1}, \hat{D} \rangle$ becomes a linear functional $\hat{D}_{N-1} : \mathcal{X}_1 \times \ldots \times \mathcal{X}_{N-2} \times \mathcal{X}_N \to \mathbb{R}$, etc. Note that the Nth independent variable is not projected. Throughout we assume this independent variable corresponds to time. Equation (4.23) defines a collection $\mathcal{D}_G$ of tensors $\hat{D}_G : X_N \to \mathbb{R}$ defined by

$$\begin{aligned} \hat{D}_G &:= \hat{D}(\varphi_1^{(k_1)}, \ldots, \varphi_{N-1}^{(k_{N-1})}, \cdot) \\ &= \sum_{\ell_1} \cdots \sum_{\ell_N} \hat{d}_{\ell_1 \cdots \ell_N} \langle e_1^{(\ell_1)}, \varphi_1^{(k_1)} \rangle \cdots \langle e_{N-1}^{(\ell_{N-1})}, \varphi_{N-1}^{(k_{N-1})} \rangle \langle e_N^{(\ell_N)}, \cdot \rangle \end{aligned} \tag{4.24}$$

for $1 \leq k_m \leq r_m$, $m = 1, \ldots, N-1$, see Lemma A.4.1 in the appendix for background. Now, $\mathcal{D}_G$ is defined by

$$\mathcal{D}_G = \{ \hat{D}_G : X_N \to \mathbb{R} \mid 1 \leq k_m \leq r_m,\ m = 1, \ldots, N-1 \}. \tag{4.25}$$

Equation (4.24) can be simplified by defining

$$b_{k_1 \cdots k_{N-1} \ell_N} = \sum_{\ell_1} \cdots \sum_{\ell_{N-1}} \hat{d}_{\ell_1 \cdots \ell_N} \langle e_1^{(\ell_1)}, \varphi_1^{(k_1)} \rangle \cdots \langle e_{N-1}^{\ell_{N-1}}, \varphi_{N-1}^{(k_{N-1})} \rangle. \tag{4.26}$$

This gives

$$\hat{D}_G(\cdot) = \sum_{\ell_N} b_{k_1 \cdots k_{N-1} \ell_N} \langle e_N^{(\ell_N)}, \cdot \rangle \tag{4.27}$$

Now, we are in a position to define a reduced order model. Given a time instance $t = p_N^{(k_N)}$, the reduced model is given by the following equations

$$D_G(p_N^{(k_N)}) = b_{k_1 \cdots k_N} = 0 \tag{4.28}$$

for $1 \leq k_m \leq r_m$, $m = 1, \ldots, N-1$. Given the order of the reduced model, $r = (r_1, \ldots, r_{N-1})$, the spectral expansion used for $w(p_1, \ldots, p_N)$ is given by

$$\begin{aligned} w(p_1^{(k_1)}, \ldots, p_N^{(k_N)}) = \sum_{\ell_1=1}^{L_1} \cdots \sum_{\ell_{N-1}=1}^{L_{N-1}} & b_{\ell_1 \cdots \ell_{N_1}}(p_N^{(k_N)}) \\ & \langle \varphi_1^{(\ell_1)}, e_1^{(k_1)} \rangle \cdots \langle \varphi_{N-1}^{(\ell_{N-1})}, e_{N-1}^{(k_{N-1})} \rangle. \end{aligned} \tag{4.29}$$

Trajectories of the reduced model are thus formed by the coefficients $b_{\ell_1 \cdots \ell_{N-1}}$ that satisfy the residuals (4.28) for all $t = p_N^{(k_N)}$.

Remark 4.3.3. *The approach to Galerkin projection presented here is similar for the case $n \geq 2$. That case requires another vector space to be taken into account, but the approach remains identical.*

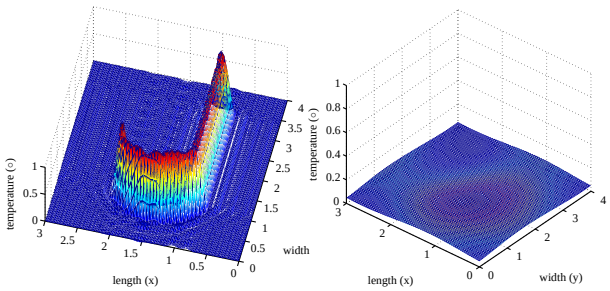


Figure 4.1: First and final time slices of the FE solution of (5.27).

4.3.4 Model reduction of a heat transfer process

Consider the following model of a heat transfer process on a rectangular plate of size $L_x \times L_y$:

$$0 = -\rho c_p \frac{\partial w}{\partial t} + \kappa_x \frac{\partial^2 w}{\partial x^2} + \kappa_y \frac{\partial^2 w}{\partial y^2}. \tag{4.30}$$

Here, $w(x, y, t)$ denotes temperature on position (x, y) and time $t \in \mathbb{T} := [0, T_f]$ and the rectangular spatial geometry defines the Cartesian product $\mathbb{X} \times \mathbb{Y} := [0, L_x] \times [0, L_y]$. The plate is assumed to be insulated from its environment. Let $\mathcal{X} = \mathcal{L}_2(\mathbb{X} \times \mathbb{Y})$ be the Hilbert space of square integrable functions on $\mathbb{X} \times \mathbb{Y}$ and let $\mathcal{X}_r = \mathcal{X}_{r_1} \times \mathcal{Y}_{r_2}$ with $\mathcal{X}_{r_1} \subseteq \mathcal{X} = \mathcal{L}_2(\mathbb{X})$ and $\mathcal{Y}_{r_2} \subseteq \mathcal{Y} = \mathcal{L}_2(\mathbb{Y})$ be finite dimensional subspaces spanned by r_1 and r_2 orthonormal bases functions $\{\varphi_1^{(\ell_1)}\}$ and $\{\varphi_2^{(\ell_2)}\}$, respectively. Solutions of the reduced model are then given by $w_r(x, y, t) = \sum_{\ell_1=1}^{r_1} \sum_{\ell_2=1}^{r_2} a_{\ell_1 \ell_2}(t) \varphi_1^{(\ell_1)}(x) \otimes \varphi_2^{(\ell_2)}(y)$ with $a_{\ell_1 \ell_2}(t) = [A(t)]_{\ell_1 \ell_2}$ a solution of the matrix differential equation

$$0 = -\rho c_p \dot{A} + \kappa_x F A + \kappa_y A P. \tag{4.31}$$

Here, F and P are defined as:

$$F = \begin{bmatrix} \langle \varphi_1^{(1)}, \ddot{\varphi}_1^{(1)} \rangle & \dots & \langle \varphi_1^{(1)}, \ddot{\varphi}_1^{(r_1)} \rangle \\ \vdots & & \vdots \\ \langle \varphi_1^{(r_1)}, \ddot{\varphi}_1^{(1)} \rangle & \dots & \langle \varphi_1^{(r_1)}, \ddot{\varphi}_1^{(r_1)} \rangle \end{bmatrix}; \; P = \begin{bmatrix} \langle \varphi_2^{(1)}, \ddot{\varphi}_2^{(1)} \rangle & \dots & \langle \varphi_2^{(1)}, \ddot{\varphi}_2^{(r_2)} \rangle \\ \vdots & & \vdots \\ \langle \varphi_2^{(r_2)}, \ddot{\varphi}_2^{(1)} \rangle & \dots & \langle \varphi_2^{(r_2)}, \ddot{\varphi}_2^{(r_2)} \rangle \end{bmatrix}$$

Alternatively, $a_{\ell_1\ell_2}(t)$ is the solution of the ordinary differential equation

$$\rho c_p \dot{a}_{\ell_1\ell_2}(t) = \kappa_x \sum_{k_1=1}^{r_1} a_{k_1\ell_2}(t)\langle \ddot{\varphi}_1^{(k_1)}(x), \varphi_1^{(\ell_1)}(x)\rangle + \kappa_y \sum_{k_2=1}^{r_2} a_{\ell_1 k_2}(t)\langle \ddot{\varphi}_2^{(k_2)}(y), \varphi_2^{(\ell_2)}(y)\rangle \quad (4.32)$$

for $1 \leq \ell_1 \leq r_1$ and $1 \leq \ell_2 \leq r_2$. A FE solution of (5.27) has been computed with physical and discretization parameters as given in Table 4.1. Time slices, including the initial condition, of the simulation data can be seen in Fig. 4.1. The boundary conditions are chosen so as to represent that the plate is insulated from its environment.

Table 4.1: PDE Parameter Values

Parameter	Value	Unit
ρC_p	5	$\frac{J}{m^3 \cdot K}$
κ_x	0.5	$\frac{W}{m \cdot K}$
κ_y	0.5	$\frac{W}{m \cdot K}$
L_x	3	m
L_y	4	m
T_f	3.6	s
Δ_x	0.05	m
Δ_y	0.05	m
Δ_t	0.05	s

In this example the original orders $(L_1, L_2, L_3) = (61, 81, 72)$ are reduced to (r_1, r_2, L_3) where we take $r_1 = r_2$. The orthonormal bases $\{\varphi_1^{(\ell_1)}\}$ and $\{\varphi_2^{(\ell_2)}\}$ have been computed using Tensor SVD and dedicated Tensor SVD construction, where in the latter time was not orthonormalized, since these basis functions will not be used in the reduced model. The first basis functions for $\mathcal{X}$ and $\mathcal{Y}$ computed using Tensor SVD described in Section 3.5 are shown in Fig. 4.2.
The simulation time of the FE implementation is 17.22s, the reduced models have a simulation time of approximately 0.35s. Table 4.2 gives the simulation error of the reduced model for different model orders. The reduced models were given the same initial condition as the was used to collect the snapshot data. Simulation errors are

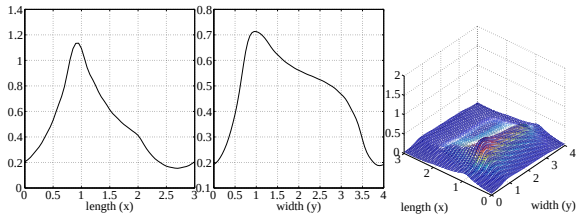


Figure 4.2: $\varphi_1^{(1)}$ (left), $\varphi_2^{(1)}$ (middle) and $\varphi_1^{(1)} \otimes \varphi_2^{(1)}$ (right). These basis functions were computed using the Tensor Singular Value Decomposition.

given for models that use basis functions computed using TSVD and basis functions computed using the dedicated construction. As can be seen in Table 4.2 using a dedicated construction to compute basis functions does not give a more accurate reduced model for this example.

Table 4.2: Reduced model simulation error results, basis functions were computed using TSVD (left) and dedicated construction (right).

r	$\frac{\|W-W_r\|_F}{\|W\|_F}$	$\frac{\|W-W_r^d\|_F}{\|W\|_F}$
$(2,2)$	0.366	0.366
$(3,3)$	0.347	0.336
$(5,5)$	0.239	0.205
$(7,7)$	0.174	0.162
$(10,10)$	0.137	0.079

4.4 Simulation example

The aim of this section is to show how the different options for spectral expansions introduced in Section 4.2.1 and Section 4.3 perform in a benchmark example. To this end, we consider their application in the reduced order modeling of a non-isothermal tubular reactor, where a first order irreversible exothermic reaction takes place [40]. The reactor is illustrated in Figure 4.3. This model has two independent variables, namely one spatial variable and time, this gives $N = 2$. The two dependent variables are temperature and concentration, therefore $n = 2$.

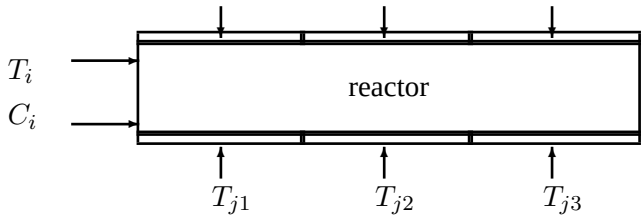


Figure 4.3: Tubular reactor

4.4.1 The model

The jacket temperatures T_{j1}, T_{j2} and T_{j3} are considered to be three independent inputs that serve as control variables. At the inlet side of the reactor, the temperature and concentration of the reactant are two additional disturbance inputs. The mathematical model of the reactor describes the (normalized) temperature $T(z,t)$ and the (normalized) concentration $C(z,t)$ of the reactant at an arbitrary location $z \in [0,1]$ of the reactor and at arbitrary time instants $t \geq 0$. The model is given by the partial differential equations.

$$\frac{\partial T}{\partial t} = \frac{1}{P_{\text{eh}}}\frac{\partial^2 T}{\partial z^2} - \frac{1}{L_e}\frac{\partial T}{\partial z} + \nu C e^{\gamma\left(1-\frac{1}{T}\right)} + \mu(T_{\text{wall}} - T)$$
$$\frac{\partial C}{\partial t} = \frac{1}{P_{\text{em}}}\frac{\partial^2 C}{\partial z^2} - \frac{\partial C}{\partial z} - D_{\text{a}} C e^{\gamma\left(1-\frac{1}{T}\right)}$$

subject to the boundary conditions

$$\text{at } z = 0: \begin{cases} \frac{\partial T}{\partial z} = P_{\text{eh}}(T - T_i) \\ \frac{\partial C}{\partial z} = P_{\text{em}}(C - C_i) \end{cases} \qquad \text{at } z = 1: \begin{cases} \frac{\partial T}{\partial z} = 0 \\ \frac{\partial C}{\partial z} = 0 \end{cases}$$

Here, the wall temperature T_{wall} is given by

$$T_{\text{wall}}(z,t) = \begin{cases} T_{j1}(t) & \text{if } 0 \leq z \leq 1/3 \\ T_{j2}(t) & \text{if } 1/3 \leq z \leq 2/3 \\ T_{j3}(t) & \text{if } 2/3 \leq z \leq 1 \end{cases}$$

where T_{j1}, T_{j2}, T_{j3} are the jacket temperatures. The physical parameters of the model are given in Table 4.3.

P_{eh}	P_{em}	L_{e}	D_{a}	γ	ν	μ
5	5	1.0	0.875	15	0.84	13.0

Table 4.3: Physical parameters

Let

$$\underline{w}(z,t) := \begin{pmatrix} T(z,t) \\ C(z,t) \end{pmatrix},\ \underline{u}(t) := \begin{pmatrix} T_{j1}(t) \\ T_{j2}(t) \\ T_{j3}(t) \end{pmatrix},\ \underline{d}(t) := \begin{pmatrix} T_i(t) \\ C_i(t) \end{pmatrix}$$

denote the state, control input and disturbance input of the model, respectively.

4.4.2 The data

A steady state operating condition has been determined for the model by carrying out an optimization on the three jacket temperatures $\underline{u} = \text{col}(T_{j1}, T_{j2}, T_{j3})$ under the assumption that the temperature and concentration inlets are given by the normalized values of the disturbances $\underline{d}(t) = \text{col}(T_i(t), C_i(t)) = \text{col}(1,1)$ for $t \geq 0$. The optimization has been performed by minimizing a criterion function that expresses a trade-off between a minimal energy consumption in the reactor and a maximal production (i.e., a minimum of reactant concentration) under the constraint that the temperature in the reactor does not exceed a certain upper limit [72]. This resulted in optimal steady state jacket temperatures

$$\underline{u}^* = \begin{pmatrix} T_{j1}^* \\ T_{j2}^* \\ T_{j3}^* \end{pmatrix} = \begin{pmatrix} 0.9970 \\ 1.0475 \\ 1.0353 \end{pmatrix}$$

and corresponding steady state temperature and concentration profiles
$(T^*(z), C^*(z))$ as shown in Figure 4.4.
This optimal steady state operating condition turns out to be asymptotically stable. However with a very small region of attraction. Indeed, a 3% perturbation on the steady state inlet temperature or inlet concentration of the reactant brings the state $\underline{w}(z,t)$ of the reactor in a periodic limit-cycle.
For this, the spatial configuration of the reactor has been discretized on a uniform spatial grid of 100 points and we applied the method of lines to approximate solutions of the distributed model by a discrete iteration of the sampled state vector

$$\underline{\hat{w}}(t) := \text{col}(T(z_1,t), \cdots, T(z_{100},t), C(z_1,t), \cdots, C(z_{100},t))$$

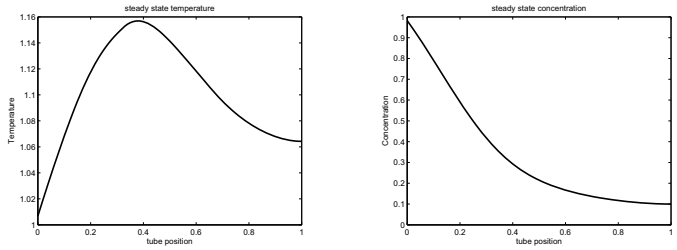


Figure 4.4: Steady state profiles for temperature and concentration

with the steady state profile as initial condition and with the perturbed inputs $T_{j1}(t) = T_{j2}(t) = T_{j3}(t) = 1$ and

$$T_i(t) = \begin{cases} 1 & \text{if } t < 5 \\ 1.038 & \text{if } t \geq 5 \end{cases}, \quad C_i(t) = 1$$

State data $\underline{w}(z,t)$ has been collected on the discretized spatial samples z_i and at 5000 equidistant time samples in the interval $0 \leq t \leq 20$. The evolution over time of temperature and concentration at point $z = 0.5$ can be seen in Fig. 4.7 (left).
All reduced models are derived from snapshot data described in this subsection, the performance of the reduced models will be evaluated using a validation data set.

4.4.3 Reduced order model performance

To assess the performance of the reduced order models, a data set is generated as described in Sec. 4.4.2, except for the inlet temperature and inlet concentration, which are disturbed as follows

$$\begin{aligned} T_i(t) &= \begin{cases} 1 & \text{if } t{<}4,\ t{>}18 \\ 1{+}0.04\mathrm{e}^{0.045(t-4)}\sin(2(t{-}4)) & \\ \quad +0.01\sin(5(t{-}4)) & \text{if } 4{\leq}t{\leq}18 \end{cases} \\ C_i(t) &= \begin{cases} 1 & \text{if } t{<}4 \\ 1{+}0.015\sin(5t){+}0.02sin(t) & \text{if } 4{\leq}t{\leq}18 \\ 1.02 & \text{if } t{>}18 \end{cases}. \end{aligned}$$

Figure 4.7 (right) shows these inlet trajectories.
For this benchmark problem a number of reduced models have been constructed. We will compare the performance of the single-variable, lumped-variable and tensor approaches, where in the tensor approach basis functions are generated using both the Higher Order Singular Value Decomposition (HOSVD) [24] and Successive Rank One approximations. The orders of the reduced models are chosen to be comparable.

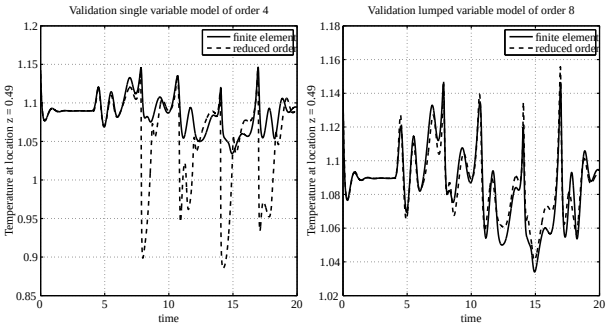


Figure 4.5: Time evolution of temperature of single-variable reduced model (left) and lumped-variable reduced model (right).

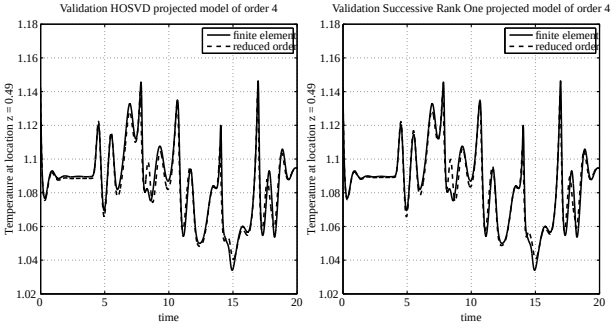


Figure 4.6: Time evolution of temperature of tensor-based reduced model with basis functions computed using HOSVD (left) and tensor-based reduced model with basis functions computed using successive rank-one approximations (right).

For the single-variable approach the order is chosen to be $(4, 4)$, the lumped variable reduced order model has order 8, and both tensor-based reduced order models have order $(4, 4)$.

Figures 4.5 and 4.6 show the time evolution of temperature at point $z = 0.5$ for the four different reduced models. The time evolution of concentration shows similar behavior for each of the reduced models. The performance of the single-variable reduced model is inferior to the performance of the other models, see also Table 4.4. This table gives the relative error of the total signal $s(z, t)$ in Frobenius norm and the worst-case errors of temperature and concentration. The table shows that the performance of the three remaining reduced models is comparable.

Table 4.4: Reduced model simulation error results

Method	$\frac{\|W-\hat{W}\|_F}{\|W\|_F}$	$\|T-\hat{T}\|_\infty$	$\|C-\hat{C}\|_\infty$
Single	0.142	0.43	0.74
Lumped	0.024	0.17	0.19
Tensor - HOSVD	0.027	0.30	0.11
Tensor - Succ. R1	0.029	0.32	0.12

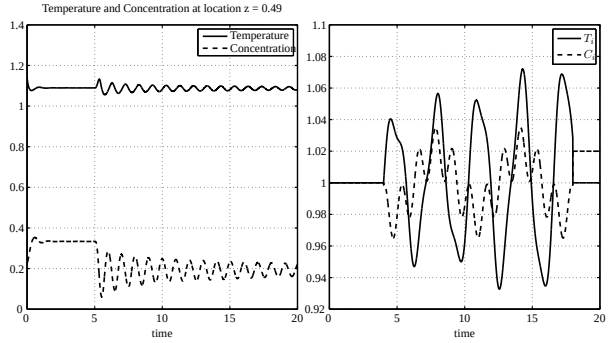


Figure 4.7: Time-evolution of temperature and concentration from snapshot data at point $z = 0.5$ (left), inlet temperature and concentration used for reduced model validation (right).

4.5 Conclusions

This chapter considered the construction of reduced order models for multi-variable distributed systems. Starting point is the method of Proper Orthogonal Decompositions. We have introduced a new method for the construction of projection spaces from measurement or simulation data of these processes, whenever a Cartesian structure can be assumed for the independent variables. These projection spaces lead to new construction of spectral expansions and Galerkin projection.The approach allows inner products for dependent and independent variables to be chosen separately, this will provide some additional freedom when dealing with scaling problems. Furthermore, in the proposed model reduction scheme reduction levels can be determined for each independent variable separately.

The tensor-based method has been applied to a tubular reactor example and compared

to single-variable and lumped-variable techniques for obtaining reduced order models. The simulation results in this paper support earlier findings in that the lumped-variable spectral expansions perform better than single-variable expansions. For this example, the performance of the tensor-based approach introduced in this paper is comparable to the performance of the lumped-variable approach. This makes the tensor approach an interesting alternative in applications of very high dimensionality, or high number of physical variables.

Chapter 5

Reconstruction and Approximation of Multidimensional Signals Described by Proper Orthogonal Decompositions

This chapter was published as the paper [5].

5.1 Introduction

The question to recover or approximate an unknown analytic function from a number of measurements has led to a substantial body of literature in the theory of interpolation, identification and function approximation. This problem has been thoroughly studied in digital signal processing and interpolation theory and its solution is key to many questions in optimization, estimation, reduction, data compression, information retrieval, filtering and optimal control. Initiated by the pioneering work of E.T. Whittaker [82], Kotelnikhov [49] and Shannon [70], the question of when a signal can be completely recovered from its samples led to a development of information theory and communication engineering that is nowadays known as sampling theory. See [84, 57] for some authorative overviews on this development. The decomposition of

analytic functions in spectral components is key to the understanding of milestone contributions such as Shannon's sampling theorem in its many variations.

This chapter considers the problem of reconstruction and approximation of multi-dimensional signals that are sampled with non-uniformly distributed sensors. More specifically, we assume signals to be defined on a N-dimensional Cartesian domain and consider multi-dimensional spectral decompositions by orthonormal functions in each co-ordinate of the signal domain. Such spectral decompositions are also called tensor decompositions as they involve the representation of a multi-linear functional in terms of orthonormal basis functions. With partial information available on the signal or the tensor, we address the problem to reconstruct or approximate the signal (or tensor) by suitably defining the spectral coefficients of a reconstructed signal on the basis of partial information only. Unlike prevailing approaches [84, 64, 57, 14, 63, 73, 35, 32], we will not consider spectral expansions with specific basis functions such as polynomials, harmonic functions, splines or shifted-modulated Gaussian functions (Gabor expansions) but, instead, decompose signals (and tensors) in a set of *empirical* basis functions and develop reconstruction strategies by taking appropriate linear combinations of these functions. See [29] for a similar approach for uniform sampled signals.

The motivation to consider empirical (or arbitrary orthonormal) basis functions stems from applications in model reduction where the aim is to find simple substitute models for complex, large-scale finite element models. The method of Proper Orthogonal Decompositions (POD), also known as Principal Component Analysis (PCA) or the Karhunen-Loève expansion is popular in the fluid dynamics community, and uses spectral decompositions and Galerkin projections to project the solution of partial differential equations onto a set of basis functions that is derived from empirical or simulated data [71, 56, 39]. In the POD method, the idea is to determine a set of empirical basis functions such that the error obtained by projecting simulated or measured data onto the span of such functions is minimal. This method has led to a substantial reduction of complexity of large-scale systems in computational fluid dynamics and has proven very useful for identifying coherent patterns in turbulent fluids. See for example [2, 3, 39, 68, 83] for some large-scale POD applications. However, despite the complexity reduction, the gain in computational speed is rather moderate for large-scale nonlinear systems due to the high dimensionality of the data. Efforts to address the problem of high computational cost include trajectory piecewise-linear approximation schemes [67, 66], spatial-temporal correlation schemes [13], Gappy POD techniques [30], missing point estimations [3] and the exploitation of symmetries [9, 69]. Each of these methods aims to remove latent variables and latent equations from large systems of differential-algebraic equations.

This chapter will focus on the missing point estimation technique as proposed in [30]

and further developed in [3, 83]. The technique is based on signal reconstruction properties from sampled data and is mainly developed for one-dimensional signals. The purpose of this chapter is to approach this reconstruction problem in a more fundamental way and employ multidimensional spectral analysis for studying reconstruction and approximation questions. The focus will be on reconstruction schemes for two-dimensional spectral expansions, with empirical orthonormal basis functions. We address the problems of exact and approximate reconstruction of sampled signals and provide expressions for alias errors and the alias sensitivity.
The chapter is organized as follows. In Section 5.2 we formulate the signal reconstruction problem that is considered in this chapter. Section 5.3 discusses the conditions for exact reconstruction of a multidimensional signal from its discrete measurements. In Section 5.4 we proceed by deriving an expression for the alias error for situations where exact reconstruction is not possible. In Section 5.5, the derived results will be illustrated on a heat transfer model. Conclusions and recommendations for further research are collected in Section 5.6. The appendix contains a review of some tensor concepts.

Preliminaries and notation

For a matrix $A \in \mathbb{R}^{n\times m}$ its transpose is denoted by $A^\top$. The left- and right inverse of A are defined by $A^{-L} = (A^\top A)^{-1}A^\top$ and $A^{-R} = A^\top(AA^\top)^{-1}$, respectively. Furthermore, recall that $(A^{-L})^\top = (A^\top)^{-R}$. For a function $f : \mathcal{A} \to \mathcal{B}$ and a set $\mathcal{A}_0 \subseteq \mathcal{A}$, we will denote by $f|_{\mathcal{A}_0}$ the restriction of f to $\mathcal{A}_0$ defined as $f|_{\mathcal{A}_0}(x) = f(x)$ for $x \in \mathcal{A}_0$. If $\mathcal{A}$ is a Hilbert space and $\mathcal{A}_0 \subseteq \mathcal{A}$ a subspace, then $\Pi_{\mathcal{A}_0} : \mathcal{A} \to \mathcal{A}_0$ denotes the canonical projection of $\mathcal{A}$ onto $\mathcal{A}_0$. The operator $\mathrm{col}(\cdot)$ stacks its arguments in a vector. The set of positive integers is denoted by $\mathbb{Z}_+$.

5.2 Problem formulation

We consider signals $w : \mathbb{D} \to \mathbb{R}$ defined on a N-dimensional domain $\mathbb{D} \subset \mathbb{R}^N$ and assume that such signals are continuous on the interior of $\mathbb{D}$. Following standard terminology in engineering, for an arbitrary finite set of points $\mathbb{D}_0$ in $\mathbb{D}$ we call the restriction $\tilde{w} := w|_{\mathbb{D}_0}$ a *sampling of* w and refer to $\mathbb{D}_0$ as a collection of *sample points*. A central paradigm in digital signal processing deals with the problem of reconstructing the signal w from its samples $\tilde{w}$. In its traditional formulation, the reconstructed signal interpolates the function values $\tilde{w}$ on the sample points. In a more general setting, the reconstructed signal is assumed to belong to a subspace spanned by a set of *reconstruction functions*. See [29, 28]. The reconstructed signal

$\hat{w}$ is then selected as a linear combination of reconstruction functions such that the error $\|w - \hat{w}\|$ is small or minimal in some norm.
Throughout this chapter, we consider the case where the domain $\mathbb{D}$ is a Cartesian product of N intervals $\mathbb{X}_i$, i.e., $\mathbb{D} = \mathbb{X}_1 \times \cdots \times \mathbb{X}_N$. The set $\mathbb{D}_0$ of sample points is assumed to have a Cartesian structure $\mathbb{D}_0 = \mathbb{X}_1^0 \times \cdots \times \mathbb{X}_N^0$ where $\mathbb{X}_i^0$ is a finite subset of $\mathbb{X}_i$. These assumptions allow us to consider multidimensional spectral decompositions in a natural way as follows. For $i = 1, \ldots, N$, let $\mathcal{W}_i = \mathcal{L}_2(\mathbb{X}_i)$ be the Hilbert space of square integrable mappings $\mathbb{X}_i \to \mathbb{R}$ with the usual inner product $\langle \cdot, \cdot \rangle_i$ and norm $\| \cdot \|_i$. Similarly, let $\mathcal{W} := \mathcal{L}_2(\mathbb{D})$ be the set of square integrable functions on $\mathbb{D}$. We assume that the function $w(x_1, \ldots, x_{i-1}, \cdot, x_{i+1}, \ldots, x_N)$, viewed as a mapping from $\mathbb{X}_i$ to $\mathbb{R}$, belongs to $\mathcal{W}_i$ for all choices of $x_k \in \mathbb{X}_k$, $k \neq i$. Because of the Cartesian structure of $\mathbb{D}$ this is equivalent to saying that $w \in \mathcal{W}$.
If

$$\{\varphi_1^{(\ell_1)}, \quad \ell_1 \in \mathbb{Z}_+\}, \quad \ldots \quad , \{\varphi_N^{(\ell_N)}, \quad \ell_N \in \mathbb{Z}_+\}$$

defines a collection of orthonormal bases for $\mathcal{W}_1, \ldots, \mathcal{W}_N$, then w admits a multidimensional spectral decomposition of the form

$$w(x_1, \ldots, x_N) = \sum_{\ell_1=1}^{\infty} \cdots \sum_{\ell_N=1}^{\infty} a_{\ell_1 \ldots \ell_N} \varphi_1^{(\ell_1)}(x_1) \cdots \varphi_N^{(\ell_N)}(x_N) \tag{5.1}$$

where the expansion coefficients are defined according to

$$a_{\ell_1 \cdots \ell_N} = \int_{\mathbb{D}} \langle w, \varphi_1^{(\ell_1)}(x_1) \cdots \varphi_N^{(\ell_N)}(x_N) \rangle \mathrm{d}x_1 \cdots \mathrm{d}x_N$$

Here, convergence of the series in (5.1) is understood in the strong $\mathcal{L}_2$ sense.
Apart from spectral decompositions (5.1), the expansion coefficients also define *tensor decompositions*. Specifically, a *tensor decomposition* of a multi-linear map $W : \mathcal{W}_1 \times \ldots \times \mathcal{W}_N \to \mathbb{R}$ operating on elements $w_i \in \mathcal{W}_i,\ i = 1, \ldots, N$ is defined by

$$W(w_1, \ldots, w_N) = \sum_{\ell_1=1}^{\infty} \cdots \sum_{\ell_N=1}^{\infty} a_{\ell_1, \ldots, \ell_N} \langle w_1, \varphi_1^{\ell_1} \rangle_1 \cdots \langle w_N, \varphi_N^{\ell_N} \rangle_N. \tag{5.2}$$

Note that W is linear in each of its arguments. The Frobenius norm of tensors and signals is defined in Sec. 3.2 and equals

$$\|w\|_{\mathrm{F}} = \|W\|_{\mathrm{F}} = \left(\sum_{\ell_1=1}^{\infty} \cdots \sum_{\ell_N=1}^{\infty} a_{\ell_1 \ldots \ell_N}^2 \right)^{1/2}.$$

We will say that the signal w or the tensor W has *finite or limited bandwidth* $(L_1, \ldots, L_N)$ if the summation index ℓ_i in (5.1) or (5.2) ranges from 1 to L_i for $i = 1, \ldots, N$. Clearly, finite bandwidth signals (or tensors) are obtained from (5.1) (or (5.2)) by projecting w on the span of the first L_i basis functions in the ith coordinate direction, $i = 1, \ldots, N$. Evidently, if a signal or a tensor has finite bandwidth $(L_1, \ldots, L_N)$ then it is uniquely defined by the N-way array $[[a_{\ell_1 \ldots \ell_N}]] \in \mathbb{R}^{L_1 \times \cdots \times L_N}$.

Let $\mathcal{D}_0$ be the set of functions $\mathbb{D}_0 \to \mathbb{R}$ and let $\mathcal{D} \subset \mathcal{W}$ be a *finite dimensional* subspace of $\mathcal{W}$ that we refer to as the *reconstruction space*. The reconstruction of a sampled signal $\tilde{w} \in \mathcal{D}_0$ is then defined by the signal $\hat{w} = R(\tilde{w})$ where

$$R : \mathcal{D}_0 \to \mathcal{D}$$

is the *reconstruction map*. Evidently, for any reconstruction map R and any signal $w \in \mathcal{W}$, the error $\|w - R(\tilde{w})\|_{\mathrm{F}}$ satisfies

$$\|w - R(\tilde{w})\|_{\mathrm{F}} \;\geq\; \|w - \Pi_{\mathcal{D}} w\|_{\mathrm{F}}$$

where $\Pi_{\mathcal{D}}$ is the orthogonal projection of $\mathcal{W}$ onto $\mathcal{D}$. In words: the projection error between w and $\Pi_{\mathcal{D}} w$ is a lower bound for the error incurred by *any* reconstruction map.

In this chapter we investigate a specific reconstruction map R for multi-variate signals that admit non-uniformly distributed samples on a Cartesian domain of sample points $\mathbb{D}_0$ and investigate the error $\|w - R(\tilde{w})\|_{\mathrm{F}}$ in terms of the *alias sensitivity*, which will be defined below. It is assumed that the sampled signal is not corrupted by noise. The reconstruction map R is well understood for signals in one independent variable ($N = 1$) and for specific collections of orthonormal basis functions $\{\varphi_i^{\ell_i}\}$ in $L_2(\mathbb{X}_i)$, including harmonic functions, Laguerre polynomials, Chebyshev polynomials, Hermite polynomials or Jacobi polynomials [63, 14, 35]. Here we consider arbitrary orthonormal bases of the Hilbert spaces $\mathcal{W}_i$, sample points that may be non-uniformly distributed and multi-dimensional spectral decompositions of signals.

In order to expose ideas clearly, we will deal with the two-dimensional (planar) case where $N = 2$. However, generalizations to higher dimensions are obtained in a straightforward manner. We therefore consider the problem to reconstruct a signal $w : \mathbb{D} \to \mathbb{R}$ with $\mathbb{D} = \mathbb{X} \times \mathbb{Y}$ from its restriction $\tilde{w}$ on a finite set of sample points $\mathbb{D}_0 = \mathbb{X}_0 \times \mathbb{Y}_0$. Here, $\mathbb{X}_0 = \{x_1, \ldots, x_N\}$ and $\mathbb{Y}_0 = \{y_1, \ldots, y_M\}$ are non-empty sets of N and M distinct samples of $\mathbb{X}$ and $\mathbb{Y}$, respectively. We assume that $w(\cdot, y) \in \mathcal{X} = \mathcal{L}_2(\mathbb{X})$ for all $y \in \mathbb{Y}$ and $w(x, \cdot) \in \mathcal{Y} = \mathcal{L}_2(\mathbb{Y})$ for all $x \in \mathbb{X}$. Equivalently, we assume that $w \in \mathcal{L}_2(\mathbb{D})$.

To define the finite dimensional reconstruction space $\mathcal{D}$, suppose that $\{\varphi_k,\ k \in \mathbb{Z}^+\}$ and $\{\psi_\ell,\ \ell \in \mathbb{Z}^+\}$ are orthonormal bases for $\mathcal{X}$ and $\mathcal{Y}$, respectively and let $\mathcal{X}_n =$

$\mathrm{span}(\varphi_1, \ldots, \varphi_n)$ and $\mathcal{Y}_m = \mathrm{span}(\psi_1, \ldots, \psi_m)$ denote a pair of n and m dimensional subspaces of $\mathcal{X}$ and $\mathcal{Y}$. For 2-D signals, the reconstruction space is then defined as

$$\mathcal{D} = \mathcal{D}_{nm} = \{\hat{w} \mid \hat{w}(\cdot, y) \in \mathcal{X}_n,\ \hat{w}(x, \cdot) \in \mathcal{Y}_m \text{ for all } (x, y) \in \mathbb{D}\}.$$

Furthermore, let

$$\tilde{\varphi}_k := \varphi_k|_{\mathbb{X}_0}, \qquad \tilde{\psi}_\ell := \psi_\ell|_{\mathbb{Y}_0}$$

denote the restrictions of the basis functions to $\mathbb{X}_0$ and $\mathbb{Y}_0$, respectively. The multidimensional spectral expansion (5.1) reads

$$w(x, y) = \sum_{k=1}^{\infty} \sum_{\ell=1}^{\infty} a_{k\ell} \varphi_k(x) \psi_\ell(y) \tag{5.3}$$

where

$$a_{k\ell} = \langle \langle w(x, y), \varphi_k(x) \rangle_{\mathcal{X}}, \psi_\ell(y) \rangle_{\mathcal{Y}} = \langle \langle w(x, y), \psi_\ell(y) \rangle_{\mathcal{Y}}, \varphi_k(x) \rangle_{\mathcal{X}}.$$

The projection of w onto $\mathcal{D}_{nm}$ is defined by the finite bandwidth (n, m) signal

$$w_{nm}(x, y) := \Pi_{\mathcal{D}_{nm}} w = \sum_{k=1}^{n} \sum_{\ell=1}^{m} a_{k\ell} \varphi_k(x) \psi_\ell(y). \tag{5.4}$$

For a given collection $\hat{a}_{k\ell}$ of real valued coefficients we define the reconstructed signal

$$\hat{w}_{nm}(x, y) := \sum_{k=1}^{n} \sum_{l=1}^{m} \hat{a}_{k\ell} \varphi_k(x) \psi_\ell(y), \qquad x \in \mathbb{X} \text{ and } y \in \mathbb{Y}. \tag{5.5}$$

Note that $\hat{w}_{nm}$ belongs to $\mathcal{D}_{nm}$. Conversely, any element of $\mathcal{D}_{nm}$ can be represented in the form (5.5) for suitable coefficients $\hat{a}_{k\ell}$. Since

$$\begin{aligned} \|w - \hat{w}_{nm}\|_F^2 &= \| \underbrace{w - w_{nm}}_{\in \mathcal{D}_{nm}^{\perp}} + \underbrace{w_{nm} - \hat{w}_{nm}}_{\in \mathcal{D}_{nm}} \|_F^2 \\ &= \underbrace{\|w - w_{nm}\|_F^2}_{\text{projection error}} + \underbrace{\|w_{nm} - \hat{w}_{nm}\|_F^2}_{\text{reconstruction error}} \end{aligned} \tag{5.6}$$

it is clear that the reconstruction error is independent of the projection error. The latter originates in the truncation of the spectral expansion (5.3). The projection error is determined by the pair (n, m), which is assumed to be fixed. This is visualized in Figure 5.1

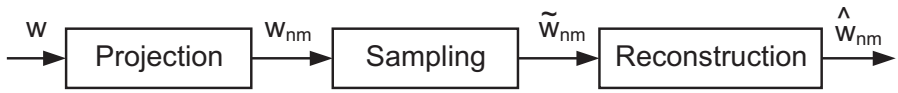


Figure 5.1: Visualisation of signal flow

5.3 Exact reconstruction

In this section we will show that under certain conditions it is possible to exactly reconstruct a continuous function $w \in \mathcal{W}$ from its samples $\tilde{w}$. We will first define two bilinear forms needed to compute the expansion coefficients $\hat{a}_{kl}$ in (5.5) and then discuss the conditions for exact reconstruction. We assume that the sets of basis functions, $\{\varphi_k\}$ and $\{\psi_\ell\}$ are known. The truncation levels n and m are given and constant.
Define:

$$\tilde{\Phi} := \begin{pmatrix} \varphi_1(x_1) & \dots & \varphi_n(x_1) \\ \vdots & & \vdots \\ \varphi_1(x_N) & \dots & \varphi_n(x_N) \end{pmatrix} \in \mathbb{R}^{N \times n}$$

and

$$\tilde{\Psi} := \begin{pmatrix} \psi_1(y_1) & \dots & \psi_m(y_1) \\ \vdots & & \vdots \\ \psi_1(y_M) & \dots & \psi_m(y_M) \end{pmatrix} \in \mathbb{R}^{M \times m}$$

i.e., the columns of $\tilde{\Phi}$ and $\tilde{\Psi}$ are sampled basis functions for $\mathcal{X}$ and $\mathcal{Y}$, respectively. Define for every $v, w \in \mathcal{X}$ and $r, s \in \mathcal{Y}$ the following bilinear forms:

$$\langle v, w \rangle_N := \sum_{i,j=1}^{N} v(x_i) q_{ij} w(x_j) \tag{5.7}$$

$$\langle r, s \rangle_M := \sum_{i,j=1}^{M} r(y_i) p_{ij} s(y_j) \tag{5.8}$$

where q_{ij} is the (i,j)th entry of

$$Q := (\tilde{\Phi}^{-L})^\top (\tilde{\Phi}^{-L}) \tag{5.9}$$

and p_{ij} is the (i,j)th entry of

$$P := (\tilde{\Psi}^{-L})^T \tilde{\Psi}^{-L} \tag{5.10}$$

The bilinear forms have the following property:

Lemma 5.3.1. *If $\tilde{\Phi}$ has full column rank, i.e. $\tilde{\Phi}$ is injective, then*

$$\langle v, w\rangle_N = \langle v, w\rangle_{\mathcal{X}}$$

for all $v, w \in \mathcal{X}_n$, where $\mathcal{X}_n = \operatorname{span}(\varphi_1, \ldots, \varphi_n)$. If $\tilde{\Psi}$ has full column rank, i.e. $\tilde{\Psi}$ is injective, then

$$\langle r, s\rangle_M = \langle r, s\rangle_{\mathcal{Y}}$$

for all $r, s \in \mathcal{Y}_m$, where $\mathcal{Y}_m = \operatorname{span}(\psi_1, \ldots, \psi_m)$

This lemma implies that if $\tilde{\Phi}$ and $\tilde{\Psi}$ have full column rank, the bilinear forms (5.7) and (5.8) define inner products for the Hilbert spaces $\mathcal{X}_n \subset \mathcal{X}$ and $\mathcal{Y}_m \subset \mathcal{Y}$. In particular, from Lemma 5.3.1 it follows that for $v, w \in \mathcal{X}_n$ the inner product $\langle v, w\rangle_{\mathcal{X}}$ can be computed from samples $\tilde{v}$ and $\tilde{w}$. The same goes for computing the inner product $\langle r, s\rangle_{\mathcal{Y}}$ when $r, s \in \mathcal{Y}_m$. This means that under certain conditions the inner product of the infinite dimensional Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$ can be computed from the sampled elements that require only N or M samples. Note that full column rank of $\tilde{\Phi}$ implies that $N \geq n$ and that full column rank of $\tilde{\Psi}$ implies that $M \geq m$.

5.3.1 Conditions for exact reconstruction

Define the expansion coefficients $\hat{a}_{kl}$ by setting

$$\hat{a}_{k\ell} = \langle\langle w(\cdot, \cdot), \varphi_k\rangle_N, \psi_\ell\rangle_M, \qquad 1 \leq k \leq n, \quad 1 \leq \ell \leq m. \tag{5.11}$$

These coefficients are actually functions of $\tilde{w}$ since (5.11) only requires knowledge of w on the sample points $\mathbb{D}_0$. This means that the reconstruction map $R : \mathcal{D}_0 \to \mathcal{D}_{nm}$

$$\hat{w}_{nm}(x, y) = R(\tilde{w})(x, y) = \sum_{k=1}^{n}\sum_{\ell=1}^{m} \hat{a}_{k\ell}\varphi_k(x)\psi_\ell(y) \tag{5.12}$$

is well defined.

Theorem 5.3.2. *Let $\mathbb{X}_0 = \{x_1, \ldots, x_N\}$ be N distinct points in $\mathbb{X}$ and $\mathbb{Y}_0 = \{y_1, \ldots, y_M\}$ be M distinct points in $\mathbb{Y}$. Furthermore, let $\{\varphi_k, k \in \mathbb{Z}^+\}$ be an orthonormal basis of $\mathcal{X}$ and $\{\psi_\ell, \ell \in \mathbb{Z}^+\}$ be an orthonormal basis of $\mathcal{Y}$. Suppose that $\tilde{\Phi}$ has rank n and $\tilde{\Psi}$ has rank m. If*

$$w \in \mathcal{D}_{nm} = \{w \mid w(\cdot, y) \in \mathcal{X}_n, w(x, \cdot) \in \mathcal{Y}_m \text{ for all } (x, y) \in \mathbb{D}\} \tag{5.13}$$

then

$$\hat{w}_{nm} = R(\tilde{w}) = w$$

where R is the reconstruction map defined in (5.12)*. That is, w can be exactly reconstructed from its samples $\tilde{w}$ by the reconstruction map* (5.12)*.*

The conditions for exact reconstruction offer an interesting interpretation. The condition (5.13) is a limitation on the bandwidth of w. Therefore, exact reconstruction is possible if w has bandwidth of at most (n, m) in terms of the basis functions for $\mathcal{X}$ and $\mathcal{Y}$. Hence, this provides a co-ordinate dependent notion of bandwidth. The other condition for exact reconstruction concerns the ranks of $\tilde{\Phi}$ and $\tilde{\Psi}$ and imply that necessarily the sample densities $N \geq n$ and $M \geq m$ need to be sufficiently large. In other words, the bandwidth of w may not exceed the number of samples to allow exact reconstruction. This is an interesting generalization of Shannon's sampling theorem that states that exact reconstruction of a signal from its samples is possible if the sampling frequency f_s is at least twice as large as the bandwidth of the signal, where the traditional concept of bandwidth in terms of harmonic functions is used.

5.4 Approximate reconstruction

Naturally, there are cases where (5.13) does not hold. Exact reconstruction is then no longer possible. In this section we derive expressions for the alias error in these cases.

5.4.1 Alias error in the expansion coefficients

From the definition of the expansion coefficients (5.11) we can derive a connection between the expansion coefficients of the original signal, a_{kl}, and the coefficients $\hat{a}_{kl}$ which are inferred from a sampled signal.

Theorem 5.4.1. *Let $\hat{a}_{k\ell}$ be defined by* (5.11) *and let $a_{k\ell}$ be the expansion coefficients of a signal w as defined in* (5.3)*. Then:*

$$\hat{a}_{k\ell} = a_{k\ell} + a_{k\ell}^{alias}, \qquad 1 \leq k \leq n, \quad 1 \leq \ell \leq m \tag{5.14}$$

where

$$\begin{aligned} a_{k\ell}^{alias} = &\sum_{p>n} a_{p\ell} \langle \varphi_p, \varphi_k \rangle_N + \sum_{q>m} a_{kq} \langle \psi_q, \psi_\ell \rangle_M \\ &+ \sum_{p>n} \sum_{q>m} a_{pq} \langle \varphi_p, \varphi_k \rangle_N \langle \psi_q, \psi_\ell \rangle_M \end{aligned} \tag{5.15}$$

The expression (5.15) consists of three terms which offer an interesting interpretation. The first term represents the alias-error originating from the sampling in $\mathbb{X}$. The second term represents the alias-error caused by sampling in $\mathbb{Y}$, while the third is a cross-term, originating from sampling in both $\mathbb{X}$ and $\mathbb{Y}$. The expression (5.14) considerably simplifies if the signal w is band-limited in either of the coordinates x and/or y. Specifically:

Corollary 5.4.2. *1. If $w(x,\cdot) \in \mathcal{Y}_m$ for all $x \in \mathbb{X}$ then the alias coefficient becomes:*

$$a_{kl}^{alias} = \sum_{p>n} a_{pl} \langle \varphi_p, \varphi_k \rangle_N.$$

2. If $w(\cdot,y) \in \mathcal{X}_n$ for all $y \in \mathbb{Y}$ then the alias coefficient becomes:

$$a_{kl}^{alias} = \sum_{q>m} a_{kq} \langle \psi_q, \psi_l \rangle_M. \tag{5.16}$$

5.4.2 The alias error

We showed that in certain cases w can be reconstructed exactly from $\tilde{w}$, i.e. w is equal to the reconstruction $\hat{w}$. In this section we examine the alias error when exact reconstruction is not possible.
The alias-coefficients are linearly dependent on the expansion coefficients a_{kl}, where $k > n$ and $l > m$. We define the alias operator S which maps the expansion coefficients, $\{a_{k\ell}, (k,\ell) \in \mathbb{Z}_+^2\}$ represented by $A \in \ell_2(\mathbb{Z}_+^2)$ in the sense that $a_{kl} = A(k,l)$, to the corresponding alias error coefficients a_{kl}^{alias}. The alias coefficients are stored in an $n \times m$ matrix A^{alias}. Therefore $S : \ell_2(\mathbb{Z}_+^2) \to \mathbb{R}^{n \times m}$ is defined by:

$$SA := A^{\text{alias}} \tag{5.17}$$

with $A^{\text{alias}}(k,l) = a_{kl}^{\text{alias}}$. The operator norm of S, $\|S\|$, is the induced norm

$$||S|| := \sup_{0 \neq A \in \ell_2(\mathbb{Z}_+^2)} \frac{||SA||_F}{||A||_F}. \tag{5.18}$$

Since $w = \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} a_{kl} \varphi_k \psi_l$ and the basis functions are fixed, we have that:

$$||S||^2 := \sup_{0 \neq A \in \ell_2(\mathbb{Z}_+^2)} \frac{||SA||_F^2}{||A||_F^2} = \sup_{0 \neq w \in \mathcal{L}_2(\mathbb{X} \times \mathbb{Y})} \frac{||w_{nm} - \hat{w}_{nm}||_F^2}{||w||_F^2}. \tag{5.19}$$

Consequently, the operator norm of the alias operator is a measure of the aliasing error. We will therefore call $\|S\|$ the *alias sensitivity*.

Theorem 5.4.3. *The alias sensitivity $\|S\|$ is given by:*

$$\|S\| = \lambda_{max}^{1/2}(G) \tag{5.20}$$

where G is an operator $G : \mathcal{T}_2 \to \mathcal{T}_2$. The coefficients of the tensor T_G associated with G with respect to the standard Euclidean bases are given by:

$$g_{rsvw} = \sum_{q>m} \langle \psi_q, \psi_s \rangle_M \langle \psi_q, \psi_w \rangle_M + \sum_{p>n} \langle \varphi_p, \varphi_r \rangle_N \langle \varphi_p, \varphi_v \rangle_N + \sum_{p>n} \langle \varphi_p, \varphi_r \rangle_N \langle \varphi_p, \varphi_v \rangle_N \sum_{q>m} \langle \psi_q, \psi_s \rangle_M \langle \psi_q, \psi_w \rangle_M \tag{5.21}$$

for $r = 1, \ldots, n$, $s = 1, \ldots, m$, $v = 1, \ldots, n$ and $w = 1, \ldots, m$. λ_{max} is the largest eigenvalue of G.

Remark 5.4.4. *The tensor representation of G with respect to the standard Euclidean bases is symmetric in its first and third and in its second and fourth coefficient. That is, $g_{rsvw} = g_{vwrs}$.*

5.4.3 Finite dimensional case

In this section we consider the case where $\mathcal{X}$ and $\mathcal{Y}$ are finite dimensional and equipped with the Euclidean inner product. We will assume that $\mathcal{X}$ is K dimensional with $K > n$ and that $\mathcal{Y}$ is L dimensional with $L > m$. We will derive an expression for the elements g_{rsvw} of G with respect to the standard Euclidean bases.
We have defined G to be an operator $G : \mathcal{T}_2 \to \mathcal{T}_2$. In this particular case G is a mapping $G : \mathbb{R}^{n \times m} \to \mathbb{R}^{n \times m}$. As shown in the Appendix, G admits a tensor representation T_G given by

$$T_G = \sum_{r=1}^{n} \sum_{s=1}^{m} \sum_{v=1}^{n} \sum_{w=1}^{m} g_{rsvw} e_n^r \otimes e_m^s \otimes e_n^v \otimes e_m^w \tag{5.22}$$

The 4-way array representation of T_G with respect to the standard Euclidean bases will be denoted by $[[g_{rsvw}]] \in \mathbb{R}^{n \times m \times n \times m}$. We will now return to the three scenarios discussed earlier and give an expression for $[[g_{rsvw}]]$ for all three scenarios:

Theorem 5.4.5. *1. If $w(x, \cdot) \in \mathcal{Y}_m$ for all $x \in \mathbb{X}$ then the alias sensitivity is given by:*

$$\|S\| = \lambda_{max}^{1/2} \left\{ (\tilde{\Phi}^\top \tilde{\Phi})^{-1} - I \right\} \tag{5.23}$$

2. *If $w(\cdot, y) \in \mathcal{X}_n$ for all $y \in \mathbb{Y}$ then the alias sensitivity is given by:*

$$\|S\| = \lambda_{max}^{1/2} \left\{ (\tilde{\Psi}^\top \tilde{\Psi})^{-1} - I \right\} \tag{5.24}$$

3. *If*

$$\begin{cases} w(\cdot, y) \notin \mathcal{X}_n & \text{for all } y \in \mathbb{Y} \\ w(x, \cdot) \notin \mathcal{Y}_m & \text{for all } x \in \mathbb{X} \end{cases} \tag{5.25}$$

then the alias sensitivity is given by:

$$\|S\| = \lambda_{max}^{1/2}([[G_y(s,w) + G_x(r,v) + G_y(s,w)G_x(r,v)]]) \tag{5.26}$$

where $G_y(s,w)$ is the (s,w)th entry of $\left\{ (\tilde{\Psi}^\top \tilde{\Psi})^{-1} - I \right\}$ and $G_x(r,v)$ is the (r,v)th entry of $\left\{ (\tilde{\Phi}^\top \tilde{\Phi})^{-1} - I \right\}$.

Theorem 5.4.6. *The operator $G : \mathcal{T}_2 \to \mathcal{T}_2$ as defined in the preceding Theorem is positive definite.*

5.5 Illustrative example

As mentioned in the introduction, the motivation to consider empirical basis functions originates from applications in the field of model reduction. Model reduction aims to find substitute models for complex, large-scale finite element models. The often excessive computation and simulation time of such models makes model-based control design, prediction or real-time monitoring virtually impossible. The method of Proper Orthogonal Decompositions (POD) is particularly popular in the fluid dynamics community and derives low order substitute models from model equations and an empirical set of basis functions. These basis functions are derived either from empirical or simulated data. The reduction process is carried out such that the error between the outputs of the original and substitute model is small. The substitute model is then used for applications such as real-time monitoring.
The theoretical results presented in the preceding section can be used either to reduce the dimensionality of the substitute model or to asses the effect of different sensor locations. To show how the theoretical results relate to a real-life example we present the following two-dimensional heat transfer process:

$$\rho c_p \frac{\partial w}{\partial t}(x,y,t) = \kappa_x \frac{\partial^2 w}{\partial x^2}(x,y,t) + \kappa_y \frac{\partial^2 w}{\partial y^2}(x,y,t) + u(x,y,t) \tag{5.27}$$

where $w(x, y, t)$ denotes the temperature at position (x, y) and at time t of some medium with heat capacity c_p, material density ρ and thermal conductivity κ, $\kappa = \kappa_x = \kappa_y$. For each time instance, solutions are defined on a closed set $\mathbb{D}$ which is assumed to be a Cartesian product $\mathbb{D} := \mathbb{X} \times \mathbb{Y}$ with boundary $\Gamma := \partial(\mathbb{D})$. Here, $\mathbb{X} = [0, L_x]$ and $\mathbb{Y} = [0, L_y]$, where $L_x > 0$ and $L_y > 0$ denote the length and width of the medium. The last term in the partial differential equation (5.27), $u(x, y, t)$, is a heat-source input, which is assumed to be factorized as

$$u(x, y, t) = s(x, y)v(t)$$

where $s(x, y)$ is an indicator function representing the source locations and $v(t)$ is the time-dependent heat input. The following initial conditions apply

$$\begin{aligned} w(x, y, 0) &= w_0(x, y) && (x, y) \in \mathbb{D} \\ \left.\frac{\partial w}{\partial x}\right|_{\Gamma} &= \gamma_1(x, y, t) && (x, y) \in \Gamma,\ t \geq 0 \\ \left.\frac{\partial w}{\partial y}\right|_{\Gamma} &= \gamma_2(x, y, t) && (x, y) \in \Gamma,\ t \geq 0. \end{aligned}$$

The first initial condition specifies the temperature profile at time $t = 0$. The other initial conditions prescribe the boundary conditions. If we consider the heating of a rectangular (length $L_x = 0.5$, $L_y = 1$ [m]) piece of aluminium, we have $c_p = 963\frac{J}{kg \cdot K}$, $\rho = 2700\frac{kg}{m^3}$ and $\kappa = 155.8\frac{W}{m \cdot K}$. The heat source is applied in a rectangular area in the center of the plate. The initial temperature profile is constant, the boundary conditions are chosen such that the rectangular plate is insulated from its environment, i.e. $\gamma_1 = \gamma_2 = 0$.

The domains $\mathbb{X}$ and $\mathbb{Y}$ are assumed to be gridded in K and L grid points, $\{x_1, \ldots, x_K\}$ and $\{y_1, \ldots, y_L\}$ with $K = 50$ and $L = 100$. Sample points are taken to be the subsets $\mathbb{X}_0 \subset \mathbb{X}$ and $\mathbb{Y}_0 \subset \mathbb{Y}$ which consist, respectively, of $N \leq K$ and $M \leq L$ inhomogeneously distributed points over the rectangular grid, (with a higher density in the vicinity of the heat source, i.e. placing a sample point at each grid point where the heat source is located and distributing the remaining sample points evenly over the other gridpoints). Finite element functions $\{\mu_k \in \mathcal{L}_2(\mathbb{X}),\ k = 1, \ldots, K\}$ and $\{\nu_\ell \in \mathcal{L}_2(\mathbb{Y}),\ \ell = 1, \ldots, L\}$ are chosen as the piecewise constant harmonic functions

$$\mu_k(x_\ell) = \begin{cases} \frac{1}{\sqrt{L_x}} & \text{for } k = 1 \\ \sqrt{\frac{2}{L_x}} \cos(\frac{(k-1)\pi x_\ell}{L_x}) & \text{for } k > 1 \end{cases}$$

$$\nu_\ell(y_k) = \begin{cases} \frac{1}{\sqrt{L_y}} & \text{for } \ell = 1 \\ \sqrt{\frac{2}{L_y}} \cos(\frac{(\ell-1)\pi y_k}{L_y}) & \text{for } \ell > 1 \end{cases};$$

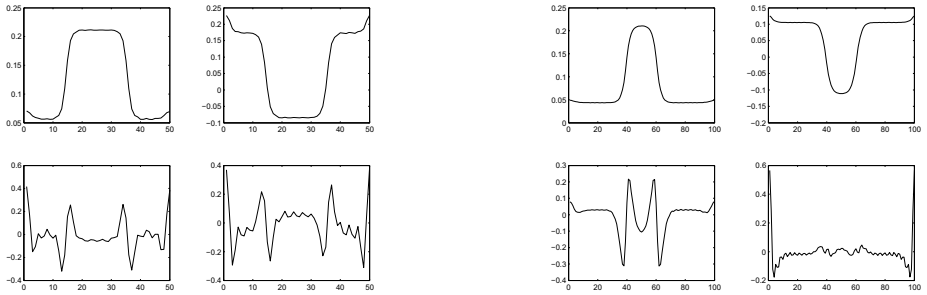


Figure 5.2: First basis functions $\varphi_1, \ldots, \varphi_4$ (left) and $\psi_1, \ldots, \psi_4$ (right)

To approximate the finite element solution of the PDE (5.27), solutions are represented in the basis functions according to

$$w^{FE}(x, y, t) = \sum_{k=1}^{K} \sum_{\ell=1}^{L} a_{kl}(t) \mu_k(x) \nu_\ell(y). \tag{5.28}$$

Two data-dependent sets of basis functions $\{\varphi_k\}_{k=1}^n$ and $\{\psi_l\}_{l=1}^m$ are determined from the finite element simulation (5.28). The first 4 basis functions are displayed in Figure 5.2. The sets $\{\varphi_k\}_{k=1}^n$ and $\{\psi_l\}_{l=1}^m$ are orthonormal and span finite dimensional subspaces $\mathcal{X}_n \subset \mathcal{L}_2(\mathbb{X})$ and $\mathcal{Y}_m \subset \mathcal{L}_2(\mathbb{Y})$, respectively. An approximate solution is obtained by truncating the spectral expansion (5.28) as

$$w_{nm}(x, y, t) = \sum_{k=1}^{n} \sum_{l=1}^{m} a_{kl}(t) \varphi_k(x) \psi_l(y) \tag{5.29}$$

with degrees $n \leq N$ and $m \leq M$.

Since all theoretical results in the preceding sections of the chapter concern the reconstruction error, we do not consider the projection error, see also Figure 5.1. We assume throughout that the projection step is carried out such that the projection error is small.

5.5.1 Exact reconstruction

In the exact reconstruction case, the original signal and the reconstructed signal have the same bandwidth:

$$\begin{aligned} w(x,y,t) &= \sum_{k=1}^{n}\sum_{l=1}^{m} a_{kl}(t)\varphi_k(x)\psi_l(y) = w_{nm}(x,y,t) \\ \hat{w}_{nm}(x,y,t) &= \sum_{k=1}^{n}\sum_{l=1}^{m} \hat{a}_{kl}(t)\varphi_k(x)\psi_l(y) \end{aligned}$$

The expansion coefficients in the reconstruction, $\hat{a}_{kl}(t)$, were determined from the sampled signal $\tilde{w}(x,y,t)$, $(x,y) \in \mathbb{X}_0 \times \mathbb{Y}_0$ by calculating (5.11). This, in turn, defines the reconstruction map.

We examined the alias error $e^{\text{alias}} = \|w_{nm} - \hat{w}_{nm}\|_F$ and averaged this error over 500 time-steps. We give the error for two different combinations of N and M, see Table 5.1.

Table 5.1: Simulation results for exact reconstruction, non-homogeneously distributed samples

n	m	N	M	Average temperature error
31	61	35	80	$1.25 \cdot 10^{-8}$
5	50	6	60	$1.35 \cdot 10^{-9}$

From the lower half of Table 5.1 we can conclude that it is possible to use a much lower spectral resolution in one dimension, i.e. take $n \ll m$, without influencing the reconstruction error. However, using a much lower spectral resolution in one dimension may influence the projection error (not shown in Table 5.1). Using a lower spectral resolution may be useful in applications where there are more high-frequent variations in one coordinate direction than in the other.

5.5.2 Approximate reconstruction

To illustrate the results on approximate reconstruction derived in subsection 5.4.3, we consider bandlimited signals of the form

$$\begin{aligned} w(x,y,t) &= \sum_{k=1}^{n}\sum_{l=1}^{m} a_{kl}(t)\varphi_k(x)\psi_l(y) = w_{nm}(x,y,t) \\ \hat{w}_{n'm'}(x,y,t) &= \sum_{k=1}^{n'}\sum_{l=1}^{m'} \hat{a}_{kl}(t)\varphi_k(x)\psi_l(y) \end{aligned}$$

where $n' < n < N$ and $m' < m < M$ and where $\hat{w}_{n'm'}$ is the signal $R(\tilde{w})$ with $w = w_{nm}$. This means that reconstructed signal has lower bandwidth than the original signal. The expansion coefficients, $\hat{a}_{kl}$, of the reconstruction map were defined and computed in the same way as for exact reconstruction. We consider the relative Frobenius norm temperature error:

$$e_{\%} = 100 \cdot \frac{\|W_{nm} - \hat{W}_{n'm'}\|_{\mathrm{F}}}{\|W_{nm}\|_{\mathrm{F}}}$$

This error was averaged over 500 time-steps. For all simulations the number of samples were set to $N = 35$ and $M = 80$. The sample locations were distributed inhomogeneously over the grid, with a higher density in the vicinity of the heat source. We considered the three different scenarios from Theorem 5.4.5:

1. Scenario I: $w(x,\cdot) \in \mathcal{Y}_m$ for all $x \in \mathbb{X}$.
   To simulate this scenario, we fixed $m' = m = 61$ and varied n' from $n' = 1$ to $n' = n = 31$. The results of this simulation are displayed in Figure 5.3. In this Figure it is clearly shown that the alias error decreases as n' increases and is zero for $n' = 31$.

2. Scenario II: $w(\cdot,y) \in \mathcal{X}_n$ for all $y \in \mathbb{Y}$.
   For this scenario, we proceeded similarly as in the simulations for scenario 2. $n' = n = 31$ remained fixed, whereas m' was varied from $m' = 1$ to $m' = m = 61$. The results are displayed in Figure 5.3. Again, the alias error decreases as m' increases and is zero for $m' = 61$.

3. Scenario III: there exist $(x,y) \in \mathbb{X} \times \mathbb{Y}$ for which

$$\begin{cases} w(\cdot,y) \notin \mathcal{X}_n \\ w(x,\cdot) \notin \mathcal{Y}_m \end{cases}$$

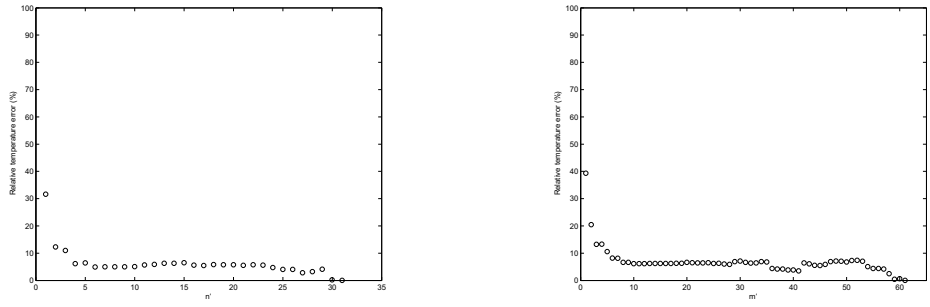


Figure 5.3: Simulation results for scenario 1 (left) and scenario 2 (right)

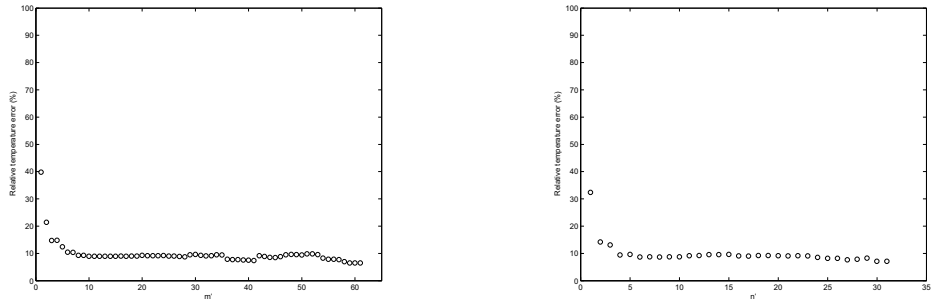


Figure 5.4: First result for scenario 3 (left) and second result for scenario 3 (right)

For this scenario, simulation results are shown in Figure 5.4. On the left the alias error is shown for a simulation, where n' was fixed at 15 and m' was varied from 1 to 61. On the right, the alias error is shown for a simulation where m' was fixed at 30 and n' was varied from 1 to 31. Both Figures show clearly that a full bandwidth in one dimension is not sufficient for the alias error to become zero.

5.6 Conclusions

In this chapter we considered the problem to recover or approximate signals defined on a multi-dimensional domain from non-uniform samples. The domain of the signal has been assumed to have a Cartesian structure and this coordinate structure has been used in a multi-dimensional spectral decomposition that uses empirical orthonormal

basis functions. This means that no assumptions on the structure or analyticity of the basis functions have been made, other than their orthonormality with respect to the Hilbert space of square integrable functions. For band-limited signals we showed that exact recovery of the signal is possible from its samples by introducing a suitable bilinear form from which the Fourier coefficients of a reconstruction function have been inferred. For non-bandlimited signals we derived an explicit alias expression for the Fourier coefficients of the alias error. We introduced an alias sensitivity operator that reflects the size of the alias error between signal and reconstruction and characterized the maximum alias sensitivity in terms of the maximal eigenvalue of a suitably defined tensor operator. It is shown that for planar signals (i.e., signals on a two-dimensional domain) and for finite dimensional inner product spaces, the alias sensitivity is computable from a matrix eigenvalue decomposition in a number of special cases.

Results in this chapter have been developed primarily for planar signals. However, the exact reconstruction result stated in Theorem 5.3.2 admits a straightforward generalization to signals defined on higher dimensional domains. The same remark applies to the result of Theorem 5.4.1 on the alias expressions. Theorem 5.4.3 generalizes to signals on N-dimensional domains by introducing a linear operator $G : \mathcal{T}_N \to \mathcal{T}_N$ in a similar manner as in the proof of Theorem 5.4.3. The alias sensitivity $\|S\|$ then becomes the maximum eigenvalue of G. An efficient numerical scheme for the computation of eigenvalues of tensorial operators does not seem to exist and is an interesting topic of future research.

In this chapter we assumed a Cartesian structure on the domain of the signals. This assumptions can not be weakened to more general (non-Cartesian) signal domains without compromising the structure that is assumed in the spectral decompositions (5.1) of signals or tensors.

The results on the characterization of the alias sensitivity operator can be applied in an algorithm to optimize the selection of sample points in each coordinate. The alias sensitivity is then used as a measure to select a suitable set of sample points which achieve a minimum alias error. See [3] for an application on sample point selection in computational fluid dynamics models.

5.7 Proofs

Proof. Proof of lemma 5.3.1.

Because of the symmetry in coordinate directions, it suffices to only prove the second part. Let $r, s \in \mathcal{Y}_m$. Then $r = \sum_{k=1}^{m} a_k \psi_k$ and $s = \sum_{l=1}^{m} b_l \psi_l$, where $a_k = \langle r, \psi_k \rangle$

for $k = 1, \dots, m$ and $b_l = \langle s, \psi_l \rangle$ for $l = 1, \dots, m$. Then

$$\langle r, s \rangle_{\mathcal{Y}} = \langle \sum_{k=1}^{m} a_k \psi_k, \sum_{l=1}^{m} b_l \psi_l \rangle = \sum_{k=1}^{m} a_k b_k = a^\top b$$

where $a = \operatorname{col}(a_1, \dots, a_m)$ and $b = \operatorname{col}(b_1, \dots, b_m)$. Now use the fact that $\tilde{r} = \tilde{\Psi} a$ and $\tilde{s} = \tilde{\Psi} b$. Since $\tilde{\Psi}$ has full column rank, a and b are uniquely determined by $\tilde{r}$ and $\tilde{s}$ and given by $a = \tilde{\Psi}^{-L} \tilde{r}$ and $b = \tilde{\Psi}^{-L} \tilde{s}$. Substitution yields

$$\langle r, s \rangle_{\mathcal{Y}} = a^\top b = \tilde{r}^\top \underbrace{(\tilde{\Psi}^{-L})^\top \tilde{\Psi}^{-L}}_{P} \tilde{s} = \tilde{r}^\top P \tilde{s} = \langle r, s \rangle_M$$

which gives the result. □

Proof. Proof of Theorem 5.3.2.
The assumption (5.13) implies that $a_{k\ell} = 0$ for $k > n$ and $\ell > m$. Hence, the signal w or the tensor W admit representations

$$w(x, y) = \sum_{k=1}^{n} \sum_{\ell=1}^{m} a_{k\ell} \varphi_k(x) \psi_\ell(y)$$
$$W = \sum_{k=1}^{n} \sum_{\ell=1}^{m} a_{k\ell} \varphi_k \otimes \psi_\ell$$

where, for $1 \le k \le n$ and $1 \le \ell \le m$,

$$a_{k\ell} = \langle \langle w, \varphi_k \rangle_{\mathcal{X}}, \psi_\ell \rangle_{\mathcal{Y}}.$$

Since both $\tilde{\Phi}$ and $\tilde{\Psi}$ have full rank, Lemma 5.3.1 promises that for all $(x, y) \in \mathbb{X} \times \mathbb{Y}$, $1 \le k \le n$ and $1 \le \ell \le m$ we have

$$\langle w(\cdot, y), \varphi_k \rangle_N = \langle w(\cdot, y), \varphi_k \rangle_{\mathcal{X}}$$
$$\langle w(x, \cdot), \psi_\ell \rangle_M = \langle w(x, \cdot), \psi_\ell \rangle_{\mathcal{Y}}.$$

Using (5.11), this yields that

$$a_{k\ell} = \langle \langle w, \varphi_k \rangle_{\mathcal{X}}, \psi_\ell \rangle_{\mathcal{Y}} = \langle \langle w, \varphi_k \rangle_N, \psi_\ell \rangle_M = \hat{a}_{k\ell}.$$

But then

$$\hat{w}_{nm}(x, y) = \sum_{k=1}^{n} \sum_{\ell=1}^{m} a_{k\ell} \varphi_k(x) \psi_l(y) = w(x, y)$$

and

$$\hat{W}_{nm} = \sum_{k=1}^{n}\sum_{\ell=1}^{m} a_{k\ell}\varphi_k \otimes \psi_\ell = W$$

as claimed. □

Proof. Proof of Theorem 5.4.1.
For $k = 1, \ldots, n$ we have that

$$\begin{aligned}\langle w, \varphi_k\rangle_N &= \Big\langle \sum_{p=1}^{\infty}\sum_{q=1}^{\infty} a_{pq}\varphi_p\psi_q, \varphi_k\Big\rangle_N = \sum_{q=1}^{\infty}\psi_q \underbrace{\sum_{p=1}^{\infty} a_{pq}\langle\varphi_p, \varphi_k\rangle_N}_{a_{kq}+\sum_{p>n} a_{pq}\langle\varphi_p,\varphi_k\rangle_N} \\ &= \sum_{q=1}^{\infty}\psi_q a_{kq} + \sum_{q=1}^{\infty}\psi_q \sum_{p>n} a_{pq}\langle\varphi_p, \varphi_k\rangle_N.\end{aligned}$$

Here, in the second equality we used Lemma 5.3.1 which states that $\langle\varphi_p, \varphi_k\rangle_N = \langle\varphi_p, \varphi_k\rangle_{\mathcal{X}} = \delta_{pk}$ for $p \le n$ and $k \le n$. Using this expression together with the orthonormality of the basis functions, we obtain that for $k = 1, \ldots, n$ and $\ell = 1, \ldots, m$,

$$\begin{aligned}\hat{a}_{kl} &= \langle\langle w, \varphi_k\rangle_N, \psi_l\rangle_M \\ &= \langle\sum_{q=1}^{\infty}\psi_q a_{kq}, \psi_l\rangle_M + \langle\sum_{q=1}^{\infty}\psi_q\sum_{p>n} a_{pq}\langle\varphi_p, \varphi_k\rangle_N, \psi_l\rangle_M \\ &= \sum_{q=1}^{\infty} a_{kq}\langle\psi_q, \psi_l\rangle_M + \sum_{p>n}\langle\varphi_p, \varphi_k\rangle_N \sum_{q=1}^{\infty} a_{pq}\langle\psi_q, \psi_l\rangle_M \\ &= \underbrace{\sum_{q=1}^{m} a_{kq}\langle\psi_q, \psi_l\rangle_M}_{a_{kl}} + \sum_{q>m} a_{kq}\langle\psi_q, \psi_l\rangle_M \\ &\quad + \sum_{p>n}\langle\varphi_p, \varphi_k\rangle_N \left\{\underbrace{\sum_{q=1}^{m} a_{pq}\langle\psi_q, \psi_l\rangle_M}_{a_{pl}} + \sum_{q>m} a_{pq}\langle\psi_q, \psi_l\rangle_M\right\}.\end{aligned}$$

For the last equality we again used Lemma 5.3.1 to infer that $\langle\psi_q, \psi_l\rangle_M = \delta_{ql}$. This gives that $\hat{a}_{kl} = a_{kl} + a_{kl}^{\text{alias}}$ with a_{kl}^{alias} as given in (5.15). □

Proof. Proof of Theorem 5.4.3.
The operator norm of the alias sensitivity S satisfies

$$\begin{aligned}\|S\|^2 &= \sup_{0\neq A\in\ell_2(\mathbb{Z}_+^2} \frac{\|SA\|_F^2}{\|A\|_F^2} \\ &= \sup_{0\neq A,\|A\|_F=1} \|SA\|^2 \\ &= \sup_{0\neq A,\|A\|_F=1} \langle SA, SA\rangle \\ &= \sup_{0\neq A,\|A\|_F=1} \langle S^*SA, A\rangle\end{aligned}$$

It follows that $\|S\|^2 = \lambda_{\max}(S^*S) = \lambda_{\max}(SS^*)$, where $\lambda_{\max}$ is the largest number λ of the eigenvalue problem

$$SS^*Z = \lambda Z.$$

Here, $SS^* : \mathbb{R}^{n\times m} \to \mathbb{R}^{n\times m}$ or, equivalently, $SS^* : \mathcal{T}_2 \to \mathcal{T}_2$ with $\mathcal{T}_2$ the set of order-2 tensors on $\mathbb{R}^n \times \mathbb{R}^m$.
Since $S : \ell_2(\mathbb{Z}_+^2) \to \mathbb{R}^{n\times m}$, its Hilbert adjoint [51] is $S^* : \mathbb{R}^{n\times m} \to \ell_2(\mathbb{Z}_+^2)$ and defined, for arbitrary $A \in \ell_2(\mathbb{Z}_+^2)$ and $B \in \mathbb{R}^{n\times m}$, by the property

$$\langle SA, B\rangle_{\mathbb{R}^{n\times m}} = \langle A, S^*B\rangle_{\ell_2(\mathbb{Z}_+^2)}.$$

Using the definition of S we find

$$\begin{aligned}\langle SA, B\rangle &= \sum_{k=1}^{n}\sum_{l=1}^{m} a_{kl}^{\text{alias}} b_{kl} \\ &= \sum_{k=1}^{n}\sum_{l=1}^{m}\Bigg[\sum_{p>n} a_{pl}\langle\varphi_k,\varphi_p\rangle_N + \sum_{q>m} a_{kq}\langle\psi_l,\psi_q\rangle_M \\ &\quad + \sum_{p>n}\sum_{q>m} a_{pq}\langle\varphi_k,\varphi_p\rangle_N\langle\psi_l,\psi_q\rangle_M\Bigg] b_{kl} \\ &= \sum_{p\in\mathbb{Z}_+}\sum_{q\in\mathbb{Z}_+} a_{pq}\left[(S^*B)(p,q)\right]\end{aligned}$$

where the (p,q)th entry of S^*B is

$$(S^*B)(p,q) = \begin{cases} 0 & 1 \le p \le n \text{ and } 1 \le q \le m \\ \sum_{k=1}^{n}\sum_{l=1}^{m} \delta_{kp}\langle \psi_q, \psi_l\rangle_M b_{kl} & 1 \le p \le n, q > m \\ \sum_{k=1}^{n}\sum_{l=1}^{m} \delta_{ql}\langle \varphi_p, \varphi_k\rangle_N b_{kl} & p > n, 1 \le q \le m \\ \sum_{k=1}^{n}\sum_{l=1}^{m} \langle \varphi_p, \varphi_k\rangle_N \langle \psi_q, \psi_l\rangle_M b_{kl} & p > n, q > m \end{cases} \tag{5.30}$$

It remains to obtain a representation for $G := SS^* : \mathcal{T}_2 \to \mathcal{T}_2$ as linear operator on the set of order-2 tensors on $\mathbb{R}^n \times \mathbb{R}^m$. For this, let $T_G \in \mathcal{T}_4$ be the multiplicative tensor associated with G (See appendix I). G is then defined by the elements g_{rsvw} of T_G, which are obtained by evaluating

$$g_{rsvw} = \langle T^{rs}, SS^*T^{vw}\rangle_{\mathbb{R}^{n\times m}} = \langle S^*T^{rs}, S^*T^{vw}\rangle_{\ell_2}$$

where T^{rs} and T^{vw} are rank-1 tensors defined by:

$$T^{ij} = e^i_n \otimes e^j_m \tag{5.31}$$

where the vectors e^i_n and e^j_m for $i = 1, \ldots, n$ and $j = 1, \ldots, m$ form the standard Euclidean basis for $\mathbb{R}^n$ and $\mathbb{R}^m$. For this, substitute $B = T^{rs}$ in (5.30) to infer that

$$(S^*T^{rs})(p,q) = \begin{cases} 0 & 1 \le p \le n, 1 \le q \le m \\ \delta_{rp}\langle \psi_q, \psi_s\rangle_M & 1 \le p \le n, q > m \\ \delta_{qs}\langle \varphi_p, \varphi_r\rangle_N & p > n, 1 \le q \le m \\ \langle \varphi_p, \varphi_r\rangle_N \langle \psi_q, \psi_s\rangle_M & p > n, q > m \end{cases}$$

Hence,

$$\begin{aligned} g_{rsvw} &= \langle S^*T^{rs}, S^*T^{vw}\rangle_{\ell_2} \\ &= \sum_{p=1}^{n}\sum_{q>m} \delta_{rp}\delta_{vp}\langle \psi_q, \psi_s\rangle_M \langle \psi_q, \psi_w\rangle_M \\ &+ \sum_{p>n}\sum_{q=1}^{m} \delta_{qs}\delta_{qw}\langle \varphi_p, \varphi_r\rangle_N \langle \varphi_p, \varphi_v\rangle_N \\ &+ \sum_{p>n}\sum_{q>m} \langle \varphi_p, \varphi_r\rangle_N \langle \psi_q, \psi_s\rangle_M \langle \varphi_p, \varphi_v\rangle_N \langle \psi_q, \psi_w\rangle_M. \end{aligned}$$

which rewrites as (5.21). This gives the result. □

Proof. Proof of Theorem 5.4.5

1. If $\dim\mathcal{X} = K$, the expression (5.21) simplifies to

$$g_{rsvw} = \sum_{p=n+1}^{K} \langle\varphi_p, \varphi_r\rangle_N \langle\varphi_p, \varphi_v\rangle_N.$$

and no longer depends on s and w. The elements of the four-dimensional array $[[g_{rsvw}]]$ can therefore be equivalently represented by two-dimensional array, the matrix $G_x \in \mathbb{R}^{n\times n}$, say. Define $\tilde{\Phi}_{\text{tail}}$ as the $N \times (K-n)$ matrix whose $(k-n)$th column is the vector of restrictions $\tilde{\varphi}_k = \varphi_k|_{\mathbb{X}_0}$, $n < k \leq K$. Then, using the orthonormality of the basis $\{\varphi_k, k = 1, \ldots, K\}$, we have that $\Phi^\top\Phi = \Phi\Phi^\top = I_K$ and

$$\begin{pmatrix}\tilde{\Phi} & \tilde{\Phi}_{\text{tail}}\end{pmatrix}\begin{pmatrix}\tilde{\Phi} & \tilde{\Phi}_{\text{tail}}\end{pmatrix}^\top = I_N. \tag{5.32}$$

With Q the matrix defined in (5.7), this implies that

$$\begin{aligned} G_x &= \tilde{\Phi}Q\tilde{\Phi}_{\text{tail}}\tilde{\Phi}_{\text{tail}}^\top Q\tilde{\Phi}^\top = \tilde{\Phi}^\top Q\left(I_N - \tilde{\Phi}\tilde{\Phi}^\top\right)Q\tilde{\Phi} = \\ &= \tilde{\Phi}^\top\tilde{\Phi}(\tilde{\Phi}^\top\tilde{\Phi})^{-2}\tilde{\Phi}\left(I_N - \tilde{\Phi}\tilde{\Phi}^\top\right)\tilde{\Phi}(\tilde{\Phi}^\top\tilde{\Phi})^{-2}\tilde{\Phi}^\top\tilde{\Phi} = \\ &= (\tilde{\Phi}^\top\tilde{\Phi})^{-1}\left(\tilde{\Phi}^\top\tilde{\Phi} - (\tilde{\Phi}\tilde{\Phi}^\top)^2\right)(\tilde{\Phi}^\top\tilde{\Phi})^{-1} = \\ &= (\tilde{\Phi}^\top\tilde{\Phi})^{-1} - I_n \end{aligned}$$

where, in the second equality we used that (5.32) implies $\tilde{\Phi}_{\text{tail}}\tilde{\Phi}_{\text{tail}}^\top = I_N - \tilde{\Phi}\tilde{\Phi}^\top$. Hence, the alias sensitivity is given by:

$$\|S\| = \lambda_{\max}^{1/2}(G_x) = \lambda_{\max}^{1/2}\left\{\left(\tilde{\Phi}^\top\tilde{\Phi}\right)^{-1} - I\right\} \tag{5.33}$$

2. The proof is similar to the previous case.

3. In the third case, the summations in the expression (5.21) for g_{rsvw} run to K or L. Specifically, the multiplication tensor T_G is a sum of three tensors, given by:

$$G = \sum_r\sum_s\sum_v\sum_w [G_y(s,w) + G_x(r,v) + G_x(r,v)G_y(s,w)] \\ e_n^r \otimes e_m^s \otimes e_n^v \otimes e_m^w$$

where $G_y(s,w)$ is the (s,w)th entry of $(\tilde{\Psi}^\top\tilde{\Psi})^{-1}-I$ and $G_x(r,v)$ is the (r,v)th entry of $(\tilde{\Phi}^\top\tilde{\Phi})^{-1}-I$. Hence, the representation of G with respect to the standard Euclidean bases, $[[g_{rsvw}]] \in \mathbb{R}^{n\times m\times n\times m}$, is given by the four-way array

$$[[g_{rsvw}]] = [[G_y(s,w) + G_x(r,v) + G_y(s,w)G_x(r,v)]]. \tag{5.34}$$

The alias sensitivity thus becomes

$$\|S\| = \lambda_{\max}^{1/2}\left([[G_y(s,w) + G_x(r,v) + G_y(s,w)G_x(r,v)]]\right) \tag{5.35}$$

□

Proof. Proof of Theorem 5.4.6 We need to show that $\langle A, GA\rangle > 0$ for all $A \in \mathcal{T}_2, A \neq 0$. Since

$$\langle A, GA\rangle = \langle A, SS^*A\rangle_{\mathbb{R}^{n\times m}} = \langle S^*A, S^*A\rangle_{\ell_2(\mathbb{Z}_+^2,\mathbb{R})} = \|S^*A\| \geq 0$$

it suffices to prove that $\dim(\ker S^*) = 0$. To see this let $B \neq 0$ and consider S^*B as defined in (5.30). Since $B \neq 0$, there exists $(\hat{k},\hat{l})$ such that $b_{\hat{k}\hat{l}} \neq 0$. Set $q = \hat{l}$, let

$$\tilde{\Phi}_{\text{tail}} = [\tilde{\varphi}_{n+1}\ldots\tilde{\varphi}_K].$$

and define $X \in \mathbb{R}^{n\times(K-n)}$ to be the matrix whose (k,ℓ)th entry is $\langle\varphi_k,\varphi_{\ell+n}\rangle_N^2$ for $1 \leq k \leq n$ and $1 \leq \ell \leq K-n$. Then

$$X = \tilde{\Phi}^\top Q\tilde{\Phi}_{\text{tail}}\tilde{\Phi}_{\text{tail}}^\top Q\tilde{\Phi} = \left(\tilde{\Phi}^\top\tilde{\Phi}\right)^{-1} - I_n.$$

(See the proof of Theorem 5.4.5). Now, $X = \left(\tilde{\Phi}^\top\tilde{\Phi}\right)^{-1} - I_n$ is not equal to zero unless $\mathbb{X}_0 = \mathbb{X}$ which is not the case. Therefore, there is a $p > n$ such that $\langle\varphi_p,\varphi_{\hat{k}},\rangle_N \neq 0$. Consequently, $\ker(S^*)$ is trivial, which gives the result. □

Chapter 6

Conclusion

Many people find open-ended novels frustrating. After all the time and effort spent on a story and its characters, one wants to find out what happens and to be left unsure is often a disappointment. Unfortunately, it is the nature of science and scientific communication that both researcher and reader are left with questions. Every scientific publication, be it a research paper or a PhD thesis such as this, is necessarily open-ended since research generally triggers at least as many questions as it can answer. This also holds for the work described in this thesis. Therefore, this last chapter serves to give an overview of this work and, more importantly, it points to what has *not* been achieved and which new questions have arisen.
The organization of this chapter is as follows. We first give an overview and summarize the most important concepts and results of each chapter. Then, we turn to the problem statement and discuss the contributions of this work. The elements of the problem statement that have not been resolved are also indicated. This then automatically leads to discussion of future research questions and we end with general conclusions.

6.1 Overview

The aim of this work has been to develop numerical techniques to extract specific information from process models. Specifically, we considered problems regarding approximation of multi-variable signals and systems. In Chapter 2 we argue that these approximation problems can be phrased as spectral decomposition problems where the notion of spectral content can be (and has been) generalized to different features of a signal. Since we consider multi-dimensional signals and systems, tensors can

be used to solve these low-rank approximation problems. Chapter 2 then proceeds to give the problem statement for this work.

Chapter 3 considers the problem of finding low-rank approximations to tensors. For order-2 tensors, matrices, this problem is well understood, see Appendix A.3. Generalization of these results to higher-order tensors, however, is not straightforward. Finding tensor decompositions that allow suitable approximations after truncation is an active area of research [47], to which this chapter contributes in various ways.

The problem of low-rank approximations to tensors is ill-posed. Therefore, we have considered a different rank concept, referred to as multi-linear or modal rank. We defined a new method to obtain modal rank decompositions to tensors. This method has been referred to as *TSVD*, which is short for Tensor Singular Value Decomposition. We have derived properties of the TSVD and in certain cases we have presented error bounds when the method is used for low-rank approximations to tensors. In Sec. 3.7 we have proposed an adaptation of the TSVD method that may give better approximation results when not all modal directions are approximated. In Sec. 3.8 we have presented a numerical algorithm for the computation of the (dedicated) TSVD. With a small adaptation, this algorithm can also be used to compute successive rank-one approximation to tensors. Finally, in Sec. 3.9, we have included a simulation example which demonstrates the methods proposed in this work and compares them to a well-known existing method.

The concepts that were introduced and discussed in Chapter 3 were used in a system approximation context in Chapter 4. The chapter started with a discussion of the well-known model reduction method of Proper Orthognal Decompositions (POD). We have shown how the low-rank approximations to tensors can be used to define projection spaces in POD. Using these alternative projection spaces leads to changes in the spectral decompositions and Galerkin projection, resulting in an adaptation of the POD method. This adaptation is both a generalization and a restriction. It is a generalization because it allows POD to be used in a scalable fashion for problems with an arbitrary number of dependent and independent variables. On the other hand, it is also a restriction, since the projection spaces used are not ordinary projection spaces, but ones that have a Cartesian product structure. The model reduction method that is obtained by combining the signal and system approximation concepts has been demonstrated on a benchmark example from chemical engineering. This simulation example shows that the method is indeed feasible, and that the performance is comparable to existing methods.

Chapter 5 considered the problem of reconstruction and approximation of multi-dimensional signals, if these signals are sampled with non-uniformly distributed sensors. We considered multi-dimensional signals on a Cartesian domain. The central question of this chapter is that of finding a reconstruction $\hat{w}$ of w from its samples.

We considered a reconstruction map R and have presented conditions for exact reconstruction of w from $\tilde{w}$. In case that exact reconstruction is not possible, we have derived an expression for the reconstruction error.

6.2 Contributions and future research

The focus of this work has been on low-rank approximations to signals and systems. Specifically, the contributions of this work are the following.

- In Chapter 3, we have considered the problem of finding low-rank approximations to tensors. We have defined a new method for the computation of low modal rank approximations to tensors, called *TSVD*. We have derived properties of this method and in certain cases provided error bounds when the method is used for low-rank approximations to tensors. We have also defined an adaptation of the TSVD, that may provide increased accuracy when not all modal directions are approximated. We have derived a numerical algorithm for the computation of the TSVD and analyzed its convergence properties. With a small adaptation, this algorithm can also be used to compute successive rank-one approximations to tensors. The method proposed in this work was compared to existing methods in a simulation example.
  Results in this chapter support earlier findings that indicate that most of the approximation properties of the matrix SVD do not naturally carry over when generalizing to higher-order tensors. Nevertheless, the coordinate-independent framework introduced in Chapter 3 provides additional insight into the problem of low-modal-rank approximations to tensors and underlines the usefulness of the approach in approximation of multi-dimensional signals.

- In Chapter 4 we have considered the problem of finding approximations to multi-dimensional systems. We present an adaptation of the method of Proper Orthogonal Decompositions (POD) for systems whose variables evolve over a Cartesian domain. We have used tensors to compute empirical projection spaces that define the reduced models. This leads to a modified spectral expansion and a more general Galerkin projection. The proposed model reduction method is demonstrated in two numerical examples.
  This adaptation of POD allows multiple dependent and independent variables explicitly to be taken into account. Choosing inner products for dependent and independent variables separately may provide additional degrees of freedom in dealing with scaling problems. Furthermore, the method allows truncation

levels for each independent variable to be chosen separately, which may prove useful in some applications.

- In Chapter 5 we have considered the problem of reconstruction and approximation of multi-dimensional *sampled* signals on Cartesian domains. We have presented conditions for exact reconstruction of such signals from their sampled versions. Whenever exact reconstruction is not possible, we have characterized the reconstruction error together with expressions for alias sensitivities.
  The results described in this chapter allow the Missing Point Estimation method introduced in [3] to be extended to Cartesian domains.

The approach detailed in this thesis has been tested on a small-scale benchmark example in Section 4.4. However, a lot of work still needs to be done and there are still a lot of questions to be answered before this approach can be used in an industrial context. The main steps that need to be taken are the following.

- The signal approximations considered in this work are all obtained from truncated tensor decompositions. The truncation level is chosen such that a specific level of accuracy is obtained. Alternatively, one can define a truncation level and then find the best low multi-linear rank approximation as in [42]. In the matrix case these two approaches would lead to the same result, but in the tensor case these approaches are different. Low multi-linear rank approximation methods were not considered in this work but may prove useful for the computation of projection bases in POD.

- In this thesis we did not focus on the numerical aspects surrounding this work. Various issues need further attention and research in this context. Firstly, the computational load and complexity of the algorithms proposed in Chapter 3 methods need to be investigated further. Secondly, the computational load of the reduced models defined in Chapter 4 needs to be examined. If necessary, numerical techniques such as Missing Point Estimation [3] and Discrete Empirical Interpolation [17] can be incorporated to improve computational efficiency.

- From a system-theoretic point of view, it is important that model reduction methods preserve crucial system properties such as stability and dissipativity. From a physics point of view, it is important that model reduction methods keep conservation laws, such as conservation of mass, intact. These issues have not been addressed in this work and it is a very important aspect of future research to investigate whether it is possible to include mechanisms that include these properties in the reduced model.

- Although the methods presented in this thesis have been tested on small-scale benchmark examples, exhaustive tests on industrial-scale benchmark models still need to be carried out.

- The use of tensors in system theoretic questions of signal and system approximation has resulted in novel insights and novel applications. These involve both algebraic aspects of tensors, as well as numerical tools for the computation of basic algebraic concepts such as rank, eigenvalues or decompositions of tensors. A host of research topics in this direction are foreseeable. As an example, if we look at the more distant future, it would be very interesting to investigate whether the concepts discussed in this thesis can be linked to a research topic from numerical mathematics that is currently receiving a lot of attention. A number of research teams are working towards using tensors to make computation of multi-dimensional functions on discretized grids more efficient, see [11], [36], [61] among others. The methods developed in this area may be combined with the model reduction method introduced in this work, to provide efficient algorithms for control and observer design for multi-variable distributed systems. As an example, [50] discuss Krylov subspace methods for linear systems with a specific tensor structure. It would be worthwhile to investigate whether the structure, or a modified version, of the reduced models introduced in this work allows for specific optimization algorithms that have certain numerical advantages.

6.3 General conclusions

Although a number of future research directions have been indicated in the previous section, this work is a part of developments that lead towards the end goal of further automation and re-design of industrial production processes. In this section we indicate how the work presented in this thesis contributes to the global issues that have been highlighted in Chapter 1.
In Chapters 3 and 4 approximation concepts for multi-dimensional signals and systems have been discussed and developed. These concepts allow the construction of low-complexity models of production processes, as described in Chapter 4. Key feature of these low-complexity models is that they allow extraction of those system trajectories in a way that is suitable for use in real-time simulation and operation of these processes. This means that they allow extraction of the system trajectories that are relevant to process operation. These system trajectories offer insights into the process that may not follow from measurements alone. This information can be used for analysis, optimization and control purposes.

The information contained in relevant system trajectories is especially important for sustainable operation of production processes. It allows the process to be operated in such a way that the use of natural resources is limited while at the same time undesired side-effects are minimized. The use of low-complexity does not end here. They can also be used in a simulation and design context. The trajectories that are obtained may assist in process re-design and lead to a next generation of footprint-free technology. Although these results form only a small element in the evolution towards a more sustainable and equal society, it is a move in the right direction. Many more of such small elements and large breakthroughs are needed, but they all contribute towards building a brighter future. A future where wealth is distributed more equally and our planet is preserved for future generations.

Appendix A

Notation and Technicalities

A.1 Notation

A.1.1 Symbols

The following notation is used throughout this work. Lowercase characters, a, are used to indicate signals and functions. Underlined lowercase characters, $\underline{a}$, denote vector-valued signals and functions. A scalar function of N variables is denoted by $a(x_1, \ldots, x_N)$. A vector-valued function of N independent variables is denoted by $\underline{a}(x_1, \ldots, x_N)$. Throughout $x_1, \ldots, x_N$ will be used to indicate independent variables and may refer to both space and time. Uppercase characters, A, are used to denote operators such as matrices and tensors. Uppercase calligraphic characters, $\mathcal{A}$, are used to indicate vector spaces and function spaces. Real and complex characters are dnoted as $\mathbb{R}$ and $\mathbb{C}$ and will be identified with their corresponding field. The set of integers is denoted by $\mathbb{Z}$. Double-barred characters, $\mathbb{A}$, denote intervals in the set of real or integer numbers. We will denote a projection by Π and the symbol I refers to the identity matrix.

A.1.2 Differentiation

The partial differential operator $\frac{\partial}{\partial x_k}$ will be denoted by ∂_{x_k}. For a vector space $\mathcal{X}$ equipped with an inner product we write $\langle x_1, x_2 \rangle$ to indicate the inner product, $x_1, x_2 \in \mathcal{X}$. $\| \cdot \|$ refers to the norm. $[[\cdot]]$ is used to indicate the elements in a basis-dependent representation of a tensor defined on a Cartesian product $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_N$ of vector spaces. Subscript indices refer to the mode, superscript elements denote the element number. As an example, $f_k^{(\ell_k)}$ denotes the ℓ_k-th function in

the k-th mode the tensor operates on, $\mathcal{X}_k$.

A.1.3 Polynomials

A notation that is used often in this work and requires some explanation is that of matrix-valued polynomials in several indeterminates. Let $r(\xi)$ be a polynomial of order k in the indeterminate ξ with real coefficients $a_0, \ldots, a_k$. That is $r(\xi)$ is given by $r(\xi) = a_0 + a_1\xi + \cdots + a_k\xi^k$. Replacing ξ by the differential operator $\frac{d}{dt}$ gives an order-k differential operator $r(\frac{d}{dt})$ with real coefficients. Then $r(\frac{d}{dt})$ can operate on a scalar function $f : \mathbb{R} \to \mathbb{R}$ that is k times continuously differentiable, yielding $r(\frac{d}{dt})f = a_0 f + a_1 \frac{d}{dt}f + \cdots + a_k \frac{d^k}{dt^k}f$. The shorthand notation that will be used to indicate these polynomials is $r \in \mathbb{R}[\xi]$, where $\mathbb{R}[\xi]$ denotes the set of real-valued polynomials in the indeterminate ξ.
This can be extended to the case of a vector-valued function $f : \mathbb{R} \to \mathbb{R}^n$. In this case, we consider the polynomial operator $R(\xi) \in \mathbb{R}^{\ell \times n}$. The coefficients of R are now matrices $A \in \mathbb{R}^{\ell \times n}$, i.e. $R(\xi) = A_0 + A_1\xi + \ldots + A_k\xi^k$. Again, substituting the differential operator $\frac{d}{dt}$ for ξ, then $R(\frac{d}{dt})$ operates on a function $f : \mathbb{R} \to \mathbb{R}^n$, and yields the function $g = R(\frac{d}{dt})f$.
The final step is to consider polynomials in multiple indeterminates. That is, let ξ now be a multi-indexed indeterminate $\xi = (\xi_1, \ldots, \xi_N)$ and consider a polynomial $R \in \mathbb{R}^{m \times n}[\xi_1, \ldots, \xi_N]$. The coefficients of R are matrices $R_\ell \in \mathbb{R}^{\cdot \times n}$, where ℓ is a multi-index $\ell = (\ell_1, \ldots, \ell_N)$. The generalized indeterminate ξ equals $\xi = (\xi_1, \ldots, \xi_N)$ and we defined $\xi^\ell := \xi_1^{\ell_1} \cdots \xi_N^{\ell_N}$ as the ℓ-th power of xi. The polynomial R is then given by

$$R(\xi_1, \ldots, \xi_N) := R(\xi) = \sum_{0 \leq |\ell| \leq L} R_\ell \xi^\ell = \sum_{0 \leq |\ell| \leq L} R_{\ell_1 \cdots \ell_N} \xi_1^{\ell_1} \cdots \xi_N^{\ell_N} \tag{A.1}$$

where $|\ell| = \sum_{k=1}^N \ell_k$ and $L = \sum_{k=1}^N \max(\ell_k)$. With ξ_k replaced by the partial derivative $\xi_k = \partial_{x_k}$, R defines a polynomial differential operator. To demonstrate how this notation can be derived from a set of PDEs, consider the following example.

Example A.1.1. *Consider the following set of Partial Differential Equations in* $\underline{w} = [w_1(x_1, x_2), w_2(x_1, x_2)]^\top$

$$\alpha_1 \partial_{x_1} w_1 + \alpha_2 \partial_{x_2}^2 w_2 = 0$$
$$\alpha_3 \partial_{x_2} w_1 + \alpha_4 \partial_{x_1 x_2} w_2 = 0.$$

In matrix notation this becomes

$$\begin{bmatrix} \alpha_1\partial_{x_1} & \alpha_2\partial_{x_2}^2 \\ \alpha_3\partial_{x_2} & \alpha_4\partial_{x_1x_2} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = 0. \tag{A.2}$$

The polynomial notation for this set of PDEs is

$$\begin{aligned} R(\xi) &= \sum_{0\leq|\ell|\leq L} R_{\ell_1\ell_2}\xi_1^{\ell_1}\xi_2^{\ell_2} \\ &= R_{10}\xi_1 + R_{01}\xi_2 + R_{11}\xi_1\xi_2 + R_{02}\xi_2^2. \end{aligned}$$

where $L = \max(\ell_1) + \max \ell_2 = 2$*. The coefficient matrices are given by*

$$R_{10} = \begin{bmatrix} \alpha_1 & 0 \\ 0 & 0 \end{bmatrix}; R_{01} = \begin{bmatrix} 0 & 0 \\ \alpha_3 & 0 \end{bmatrix}; R_{11} = \begin{bmatrix} 0 & 0 \\ 0 & \alpha_4 \end{bmatrix}; R_{02} = \begin{bmatrix} 0 & \alpha_2 \\ 0 & 0 \end{bmatrix}.$$

The other coefficient matrices are equal to zero. Note that the solution set of (A.2) *is linear.*

A.2 Discrete-time systems

Let $\mathbb{X} = \mathbb{X}_1 \times \cdots \times \mathbb{X}_N$, where $\mathbb{X}_k = \{p_k^{(\ell_k)} \mid \ell_k = 1, \ldots, L_k\}$ is a finite discrete grid of points in mode k and ordered according to $p_k^{(1)} \leq p_k^{(2)} \leq \cdots \leq p_k^{(L_k)}$. Consider a signal $w : \mathbb{X} \to \mathbb{R}$. Let ς_k be the forward shift operator acting on the spatial discretization in the kth mode as defined below.

Definition A.2.1. *The* forward *shift operator acting on the spatial discretization in the* k*th mode,* ς_k*, is defined as*

$$\varsigma_k w(p_1^{(\ell_1)}, \ldots, p_k^{(\ell_k)}, \ldots, p_N^{(\ell_N)}) = \begin{cases} w(p_1^{(\ell_1)}, \ldots, p_k^{(\ell_k+1)}, \ldots, p_N^{(\ell_N)}) & \ell_k < L_k \\ 0 & \ell_k = L_k \end{cases} \tag{A.3}$$

where $w : \mathbb{X} \to \mathbb{R}$ *with* $\mathbb{X} = \Pi_{k=1}^N \mathbb{X}_k$*. For infinite countable discrete grids, one or more dimensions* L_k *are infinite and we define*

$$\varsigma_k w(p_1^{(\ell_1)}, \ldots, p_k^{(\ell_k)}, \ldots, p_N^{(\ell_N)}) = w(p_1^{(\ell_1)}, \ldots, p_k^{(\ell_k+1)}, \ldots, p_N^{(\ell_N)}) \tag{A.4}$$

The backward *shift operator acting on the spatial discretization in the kth mode, ς_k^{-1}, is defined as*

$$\varsigma_k^{-1} w(p_1^{(\ell_1)}, \ldots, p_k^{(\ell_k)}, \ldots, p_N^{(\ell_N)}) = \begin{cases} w(p_1^{(\ell_1)}, \ldots, p_k^{(\ell_k - 1)}, \ldots, p_N^{(\ell_N)}) & \ell_k > 1 \\ 0 & \ell_k = 1 \end{cases}. \tag{A.5}$$

Definition A.2.2 (Discrete-time lumped system)**.** *A discrete-time lumped dynamical system Σ is defined as a triple*

$$\Sigma = (\mathbb{T}, \mathbb{W}, \mathcal{B}). \tag{A.6}$$

In this triple, $\mathbb{T} \subseteq \mathbb{Z}$ is the time axis, $\mathbb{W}$ is the signal space and $\mathcal{B}$ is a subset of the collection of maps from $\mathbb{T}$ to $\mathbb{W}$.

As for the continuous case, we are especially interested in those discrete-time lumped systems that admit a representation by means of a (linear) set of Ordinary Difference Equations. Let $D \in \mathbb{R}^{m \times n}[\xi, \eta]$ be a polynomial in two indeterminates and consider the following difference equation

$$D(\varsigma_1, \varsigma_1^{-1}) w = 0. \tag{A.7}$$

This defines the discrete-time lumped system $\Sigma = (\mathbb{T}, \mathbb{W}, \mathcal{B})$ with time set $\mathbb{T} = \{x_1, \ldots, x_{L_1}\}$ or $\mathbb{T} = \{x_k \mid k \in \mathbb{Z}\}$ with $x_k = kT_{\text{sample}}$, the behavior is given by

$$\mathcal{B} = \left\{ w \in \mathbb{W}^{\mathbb{T}} \mid D(\varsigma_1, \varsigma_1^{-1}) w = 0 \right\}. \tag{A.8}$$

The signal w may be scalar- or vector-valued. In the scalar case we have $\mathbb{W} = \mathbb{R}$, whereas in the vector-valued case $\mathbb{W} = \mathbb{R}^n$.

Definition A.2.3 (Distributed dynamical system on a discrete domain)**.** *A distributed dynamical system Σ on a discrete domain is defined as the triple*

$$\Sigma = (\mathbb{X}, \mathbb{W}, \mathcal{B}). \tag{A.9}$$

In this triple, $\mathbb{X} \subseteq \mathbb{Z}^N$ is the set of independent variables, $\mathbb{W}$ is the signal space and $\mathcal{B}$ is a subset of $\mathbb{W}^{\mathbb{X}}$ called the behavior of the system.

As for the continuous case, we are especially interested in those discrete-time lumped systems that admit a representation by means of a (linear) set of Partial Difference Equations. Let $D \in \mathbb{R}^{m\times n}[\xi_1,\dots,\xi_N,\eta_1,\dots,\eta_N]$ be a polynomial in $2N$ indeterminates and consider the following difference equation

$$D(\varsigma_1,\varsigma_1^{-1},\dots,\varsigma_N,\varsigma_N^{-1})w = 0. \tag{A.10}$$

This defines the discrete-time lumped system $\Sigma = (\mathbb{T},\mathbb{W},\mathcal{B})$ with $\mathbb{T} = \mathbb{X}_1\times\cdots\times\mathbb{X}_N$ and $\mathbb{X}_k = \{x_k^{(1)},\dots,x_k^{(L_k)}\}$ or $\mathbb{X}_k = \{x_k^{(m)} \mid m \in \mathbb{Z}\}$ with $x_k^m = mT_{\text{sample}}$, the behavior is given by

$$\mathcal{B} = \left\{ w \in \mathbb{W}^{\mathbb{X}} \mid D(\varsigma_1,\varsigma_1^{-1},\dots,\varsigma_N,\varsigma_N^{-1})w = 0 \right\} \tag{A.11}$$

The signal w may be scalar- or vector-valued. In the scalar case we have $\mathbb{W} = \mathbb{R}$, whereas in the vector-valued case $\mathbb{W} = \mathbb{R}^n$.

A.3 Optimal rank approximation to matrices

Section 3.2 introduced tensors and some of their properties. This appendix is devoted to a special class of tensors, namely order-2 tensors on finite domains, commonly referred to as matrices. As we will demonstrate , matrices have several special properties. We will introduce concepts such as matrix rank and the Singular Value Decomposition. These concepts are well-known and can be found in many books on matrices, such as [33]. These concepts form the background dor the tensor decomposition concepts.

Consider a tensor $W : \mathcal{X}_1\times\mathcal{X}_2 \to \mathbb{R}$, where $\mathcal{X}_i = \mathbb{R}^{L_i}$ is equipped with the standard Euclidean inner product for $i = 1,2$. The array of coefficients $[[w_{\ell_1\ell_2}]]$ obtained by operating W on the standard bases for $\mathbb{R}^{L_i},\ i = 1,2$ is an object $[[w_{\ell_1\ell_2}]] \in \mathbb{R}^{L_1\times L_2}$. This object is what is usually referred to as a matrix. In other words, the elements $w_{\ell_1\ell_2}$ of a matrix $[[w_{\ell_1\ell_2}]] \in \mathbb{R}^{L_1\times L_2}$ are the coefficients of the representation of W with respect to the standard bases for $\mathbb{R}^{L_1}$ and $\mathbb{R}^{L_2}$, i.e.

$$W = \sum_{\ell_1=1}^{L_1}\sum_{\ell_2=1}^{L_2} w_{\ell_1\ell_2} e_1^{(\ell_1)} \otimes e_2^{(\ell_2)}$$

Conversely, the matrix $A := [[w_{\ell_1\ell_2}]] \in \mathbb{R}^{L_1\times L_2}$ defines the tensor $W : \mathbb{R}^{L_1}\times\mathbb{R}^{L_2} \to \mathbb{R}$ according to

$$W(x_1,x_2) = x_1^\top A x_2 = \langle x_1, Ax_2\rangle = \langle A^\top x_1, x_2\rangle.$$

In the remainder of this chapter we will use the notation $A := [[w_{\ell_1\ell_2}]] \in \mathbb{R}^{L_1 \times L_2}$ so as not to confuse the tensor W with its representation.
An important concept is that of matrix rank, which is defined as follows

Definition A.3.1 (Matrix Rank)**.** *Regarding matrix rank, we can define the column-rank and the row-rank.*

1. *Let $A = [a_1 \cdots a_{L_2}] \in \mathbb{R}^{L_1 \times L_2}$ with $a_{\ell_2} \in \mathbb{R}^{L_1}$, $\ell_2 = 1, \ldots, L_2$. The column rank of A is defined as*

$$\textit{col-rank}(A) := \dim(\operatorname{span}\{a_1 \ldots a_{L_2}\}) \tag{A.12}$$

2. *Let $A = [a_1 \cdots a_{L_1}]^\top \in \mathbb{R}^{L_1 \times L_2}$ with $a_{\ell_1} \in \mathbb{R}^{L_2}$, $\ell_1 = 1, \ldots, L_1$. The row rank of A is defined as*

$$\textit{row-rank}(A) := \dim \operatorname{span}\{a_1^\top \ldots a_{L_1}^\top\} \tag{A.13}$$

Alternatively, the row and column rank of a matrix can be defined using the following kernels. Let

$$\ker_1(W) := \left\{x_1 \in \mathbb{R}L_1 \mid W(x_1, x_2) = 0,\ \forall x_2 \in \mathbb{R}^{L_2}\right\}$$
$$\ker_2(W) := \left\{x_2 \in \mathbb{R}L_2 \mid W(x_1, x_2) = 0,\ \forall x_1 \in \mathbb{R}^{L_1}\right\}.$$

Then, the row-rank and column-rank of A can be defined as

$$\text{row-rank}(A) := L_1 - \ker_1(W)$$
$$\text{column-rank}(A) := L_2 - \ker_2(W)..$$

The following result is very well known and states that the row- and column-ranks of a matrix are always equal. The proof can be found in [33] for instance.

Theorem A.3.2. *Consider a matrix $A \in \mathbb{R}^{L_1 \times L_2}$, the following holds*

$$\textit{col-rank}(A) = \textit{row-rank}(A) =: \operatorname{rank}(A). \tag{A.14}$$

The Singular Value Decomposition of matrices will play an important role in the remainder of this chapter. Its definition can be found in many books on matrices and linear algebra, such as [33]. First, we need to define unitary matrices.

Definition A.3.3 (Unitary matrix)**.** *A matrix $A \in \mathbb{R}^{L_1 \times L_1}$ is said to be a unitary matrix if*

$$A^\top A = I$$

Theorem A.3.4 (Singular Value Decomposition (SVD))**.** *Let A be a real matrix of dimension L_1-by-L_2. Then there exist orthogonal matrices*

$$U = [u_1 \cdots u_{L_1}] \in \mathbb{R}^{L_1 \times L_1}; \quad V = [v_1 \cdots v_{L_2}] \in \mathbb{R}^{L_2 \times L_2}$$

such that

$$U^\top A V = \Sigma = \operatorname{diag}(\overline{\Sigma}, 0) \ \in \mathbb{R}L_1 \times Ł_2 \tag{A.15}$$

where $\overline{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_p) \in \mathbb{R}^{p \times p}$ with $\sigma_1 \geq \cdots \geq \sigma_p > 0$ and $p = \operatorname{rank}(A)$. The numbers $\sigma_1, \ldots, \sigma_p$ are called the singular values *of A, the vectors $u_1, \ldots, u_{L_1}$ are called the* left singular vectors *of A and the vectors $v_1, \ldots, v_{L_2}$ are called the* right singular vectors *of A.*

In other words, $A \in \mathbb{R}^{L_1 \times L_2}$ can be decomposed as follows

$$A = U \Sigma V^\top$$

where, $U \in \mathbb{R}^{L_1 \times L_1}$ and $V \in \mathbb{R}^{L_2 \times L_2}$ are orthogonal matrices and Σ is an $L_1 \times L_2$ diagonal matrix with non-zero elements $\sigma_1, \ldots, \sigma_p$ on its main diagonal.
The SVD of a matrix can be interpreted in different ways. We now give three alternative interpretations, these will prove useful later on in this chapter.

1. The Singular Value Decomposition of a matrix A of rank p is a dyadic expansion, i.e. it is an expansion of A in rank-one matrices of the form

   $$A = \sum_{i=1}^{p} \sigma_i u_i v_i^\top.$$

   The orthogonality properties are now expressed in terms of the vectors, i.e. $\langle u_i, u_j \rangle = \delta_{ij}$ and $\langle v_i, v_j \rangle = \delta_{ij}$. Again, the sigma's are in non-increasing order, i.e. $\sigma_1 \geq \cdots \geq \sigma_p > 0$.

2. The Singular Value Decomposition of a matrix can be obtained through successive rank-one approximations of the matrix. Given a matrix $A \in \mathbb{R}^{L_1 \times L_2}$, let $U_1 \in \mathbb{R}^{L_1 \times L_2}$ be a rank-one matrix that minimizes the norm

   $$\|A - U_1\|_F. \tag{A.16}$$

It is straightforward to show that the solution to this problem is given by $U_1 = \sigma_1 u_1 v_1^\top$. Now, define for successive values $k = 1, \ldots, p$, the error $A_k := A - U_1 - \cdots - U_k$ and find the best rank-one matrix U_{k+1} in the sense that $\|A_k - U_{k+1}\|_F$ is minimal. Then $U_{p+1} = 0$ and $U_k = \sigma_k u_k v_k^\top$ and we infer that

$$A = \sum_{i=1}^{p} \sigma_i u_i v_i^\top.$$

3. The SVD can also be obtained through the following maximization problem

$$\max_{\substack{u,v \\ \|u\|=1, \|v\|=1}} |\langle Av, u\rangle|$$

The vectors that yield the maximum are u_1 and v_1 and the maximum is given by $\langle Av_1, u_1\rangle = \sigma_1$. This maximization can be repeated, with the additional constraints that $u \perp u_1$ and $v \perp v_1$ for $k = 2, \ldots, p$ by setting

$$\max_{\substack{u,v \\ \|u\|=1, \|v\|=1 \\ u \perp \operatorname{span}\{u_1, \ldots, u_{k-1}\}, v \perp \operatorname{span}\{v_1, \ldots, v_{k-1}\}}} |\langle Av, u\rangle|.$$

The vectors that yield the maximum are u_k and v_k and the maximum is given by $\langle Av_k, u_k\rangle = \sigma_k$. Again, we find the decomposition

$$A = \sum_{i=1}^{p} \sigma_i u_i v_i^\top.$$

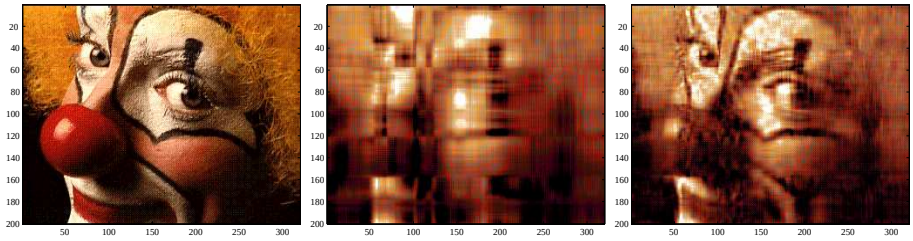
The application of the SVD that is most relevant to this work is the application to optimal rank approximation of matrices. The optimal rank approximation problem of matrices can be formulated as follows.

Problem A.3.5. *Given a matrix $A \in \mathbb{R}^{L_1 \times L_2}$, find a matrix A_k of rank k such that*

1. *$\|A - A_k\|_F$ is minimized.*
2. *$\|A - A_k\|_{ind}$ is minimized*

The solution to the optimal rank approximation problem is as follows. Let $U\Sigma V^\top$ be the SVD of A. The optimal rank-k approximation A_k can now be defined as $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^\top$ with the error equal to

$$\|A - A_k\|_{\text{ind}}^2 = \sigma_{k+1}^2$$
$$\|A - A_k\|_{\text{F}}^2 = \sum_{i=k+1}^{p} \sigma_i^2.$$



(a) Original Image (b) Rank-5 Approximation (c) Rank-15 Approximation

Figure A.1: Optimal rank approximation of the clown image. The original image is of size 200×300 and the matrix describing it has rank 200. In the middle and on the right two rank approximations of the original image are shown.

This error is minimal for both problems. This solution is unique in case the Frobenius norm is used. The solution is *not* unique if the problem is stated in the induced norm. Indeed, if we set $A_k = \sum_{i=1}^{k}(\sigma_i - \rho_i)u_i v_i^\top$ where $\rho_1, \ldots, \rho_k$ are arbitrary real numbers such that, for $i = 1, \ldots, k$

$$0 < \sigma_i - \rho_i < \sigma_{k+1}$$

then $\|A - A_k\|_{\text{ind}}$ will remain equal to σ_{k+1}, yet the approximant matrix A_k is different. An imaging example of optimal rank approximation to matrices is shown in Figure A.1. This shows an image which can be represented by a 200×300 matrix of rank 200. The middle and right of Fig. A.1 show rank approximations of the image. These low-rank approximations don't capture all detail of the original, yet it is clear that one is looking at a (distorted) image of a clown.
Optimal rank approximation to matrices is a powerful tool that has found widespread use. Generalization of this property to the more general case of order-N tensors, $N > 2$, is not straightforward as is discussed from Sec. 3.3 onward

A.4 A Useful Lemma

The following lemma proves useful and relates tensor evaluations with tensorial inner products.

Lemma A.4.1. *Let $W \in \mathcal{T}_N$, $W : \mathcal{X}_1 \times \cdots \times \mathcal{X}_N \to \mathbb{R}$, with $\mathcal{X}_n$ inner product spaces, possibly infinite dimensional, and $x_n \in \mathcal{X}_n$ for $n = 1, \ldots, N$. Then*

1.

$$W(x_1, \cdots, x_N) = \langle W, x_1 \otimes \cdots \otimes x_N \rangle.$$

2.

$$W(x_1, \cdots, x_N) = \langle x_N, \cdots \langle x_2, \langle x_1, W \rangle_1 \rangle_2 \cdots \rangle_N.$$

Proof. Proof of Lemma A.4.1

1. Let $\{\xi_n^{(\ell_n)}\}_{\ell_n=1}^{\infty}$ be an orthonormal basis for $\mathcal{X}_n$, $n = 1, \ldots, N$. W can be represented with respect to these bases as $W = \sum_{\ell_1} \cdots \sum_{\ell_N} w_{\ell_1 \cdots \ell_N} \xi_1^{(\ell_1)} \otimes \cdots \otimes \xi_N^{(\ell_N)}$. The tensor evaluation can be written as $W(x_1, \cdots, x_N) = \sum_{\ell_1} \cdots \sum_{\ell_N} w_{\ell_1 \cdots \ell_N} \langle x_1, \xi_1^{\ell_1} \rangle \cdots \langle x_N, \xi_N^{\ell_N} \rangle$. Let $U := x_1 \otimes \cdots \otimes x_N$. U can be represented as $U = \sum_{\ell_1} \cdots \sum_{\ell_N} u_{\ell_1 \cdots \ell_N} \xi_1^{(\ell_1)} \otimes \cdots \otimes \xi_N^{(\ell_N)}$ with $u_{\ell_1 \cdots \ell_N} = \prod_{i=1}^{N} \langle x_i, \xi_i^{(\ell_i)} \rangle_i$. Then,

$$\begin{aligned}
\langle W, U \rangle &= \sum_{k_1} \cdots \sum_{k_N} \sum_{\ell_1} \cdots \sum_{\ell_N} w_{k_1 \cdots k_N} u_{\ell_1 \cdots \ell_N} \\
&\qquad \cdot \underbrace{\langle \xi_1^{k_1}, \xi_1^{\ell_1} \rangle}_{0 \text{ unless } k_1 = \ell_1} \cdots \langle \xi_N^{k_N}, \xi_N^{\ell_N} \rangle \\
&= \sum_{\ell_1} \cdots \sum_{\ell_N} w_{\ell_1 \cdots \ell_N} u_{\ell_1 \cdots \ell_N} \\
&= \sum_{\ell_1} \cdots \sum_{\ell_N} w_{\ell_1 \cdots \ell_N} \langle \xi_1^{\ell_1}, x_1 \rangle \cdots \langle \xi_N^{\ell_N}, x_N \rangle
\end{aligned}$$

   which is the tensor evaluation.

2. To prove the second statement, we first show that $\langle x_1, W(\cdot, v_2, \ldots, v_N) \rangle_1 = W(x_1, v_2, \ldots, v_N)$ for some $v_n \in \mathcal{X}_n$, $n = 2, \ldots, N$. Let $\{\xi_n^{(\ell_n)}\}_{\ell_n=1}^{\infty}$ be an orthonormal basis for $\mathcal{X}_n$, $n = 1, \ldots, N$. W can be represented with respect to these bases as $W = \sum_{\ell_1} \cdots \sum_{\ell_N} w_{\ell_1 \cdots \ell_N} \xi_1^{(\ell_1)} \otimes \cdots \otimes \xi_N^{(\ell_N)}$. Then

$$W(x_1, v_2, \ldots, v_N) = \sum_{\ell_1} \cdots \sum_{\ell_N} w_{\ell_1 \cdots \ell_N} \langle \xi_1^{(\ell_1)}, x_1 \rangle \prod_{k=2}^{N} \langle \xi_k^{(\ell_k)}, v_k \rangle.$$

On the other hand, we can write x_1 as $x_1 = \sum_{k=1}^{\infty} \langle x_1, \xi_1^{(k)} \rangle \xi_1^{(k)}$. Then

$$\begin{aligned}
&\langle x_1, W(\cdot, v_2, \ldots, v_N) \rangle_1 \\
&\qquad = \sum_{k_1} \langle x_1, \xi_1^{(k_1)} \rangle_1 W(\xi_1^{(k_1)}, v_2, \ldots, v_N) \\
&\quad = \sum_{k_1} \langle x_1, \xi_1^{(k_1)} \rangle_1 \sum_{\ell_2} \cdots \sum_{\ell_N} w_{k_1 \ell_2 \cdots \ell_N} \prod_{k=2}^{N} \langle \xi_k^{(\ell_k)}, v_k \rangle \\
&\qquad = W(x_1, v_2, \ldots, v_N)
\end{aligned}$$

Thus, we have that $\langle x_1, W(\cdot, v_2, \ldots, v_N) \rangle_1 = W(x_1, v_2, \ldots, v_N)$. Since tensors are multilinear functionals, this completes the proof.

□

Bibliography

[1] T. C. Antoulas, *Approximation of large-scale dynamical systems*, Society for Industrial and Applied Mathematics, 2005.

[2] P. Astrid, *Reduction of process simulation models: a proper orthogonal decomposition approach*, Ph.D. thesis, Eindhoven University of Technology, The Netherlands, 2004.

[3] P. Astrid, S. Weiland, K. Willcox, and T. Backx, *Missing point estimation in models described by proper orthogonal decomposition*, IEEE Transactions on Automatic Control **53** (2008), no. 10, 2237–2251.

[4] B. W. Bader and T. G. Kolda, *Algorithm 862: Matlab tensor classes for fast algorithm prototyping*, ACM Transactions on Mathematical Software (2006).

[5] F. van Belzen and S. Weiland, *Reconstruction and approximation of multidimensional signals described by proper orthogonal decompositions*, IEEE Transactions on Signal Processing **56** (2008), no. 2, 576 –587.

[6] ———, *Efficient simulation of large-scale dynamical systems using tensor decompositions*, Scientific Computing in Electrical Engineering SCEE 2008 (Janne Roos and Luis R.J. Costa, eds.), Mathematics in Industry, vol. 14, Springer Berlin Heidelberg, 2010, pp. 413–420.

[7] ———, *A tensor decomposition approach to data compression and approximation of nd systems*, Multidimensional Systems and Signal Processing (2010), 1–28, 10.1007/s11045-010-0144-x.

[8] F. van Belzen, S. Weiland, and L. Ozkan, *Model reduction of multi-variable distributed systems through empirical projection spaces*, Proceedings of the Joint 48th IEEE Conference on Decision and Control and 28th Chinese Control Conference, IEEE, 2009, pp. 5351–5356.

[9] G Berkooz and E.S. Titi, *Galerkin projections and the proper orthogonal decomposition for equivariant equations*, Physics Letters A. **174** (1993), 94–102.

[10] D. P. Bertsekas, *Constrained optimization and lagrange multiplier methods*, Academic Press New York, 1982.

[11] G. Beylkin and M. J. Mohlenkamp, *Algorithms for numerical analysis in high dimensions*, SIAM Journal on Scientific Computing **26** (2005), no. 6, 2133–2159.

[12] R. B. Bird, W. E. Stewart, and E. N. Lightfoor, *Transport phenomena*, Wiley, 2002.

[13] R. Bos, X. Bombois, and P. Van den Hof, *Accelerating large-scale nonlinear models for monitoring and control using spatial and temporal correlations*, Proceedings of the American Control Conference (Boston), vol. 4, June 30 - July 2 2004, pp. 3705–3710.

[14] C. Canuto, M.Y. Hussaini, A. Quarteroni, and T.A. Zang, *Spectral methods in fluid dynamics*, Springer, 1988.

[15] J. Douglas Carroll and Jih-Jie Chang, *Analysis of individual differences in multidimensional scaling via an n-way generalization of eckart-young decomposition*, Psychometrika **35** (1970), no. 3, 283–319.

[16] S. N. Chapman, *The fundamentals of production planning and control*, Pearson Prentice Hall, 2006.

[17] S. Chaturantabut and D.C. Sorensen, *Discrete empirical interpolation*, Proceedings of the Joint 48th IEEE Conference on Decision and Control and 28th Chinese Control Conference, IEEE, 2009, pp. 4316–4321.

[18] J. Chen and Y. Saad, *On the tensor svd and the optimal low rank orthogonal approximation of tensors*, SIAM Journal on Matrix Analysis and Applications **30** (2009), no. 4, 1709–1734.

[19] S. Chen and M. Ravallion, *The developing world is poorer that we thought, but no less successful in the fight against poverty*, World Bank, August 2008.

[20] P. Comon, *Mathematics in signal processing*, vol. V, ch. Tensor Decompositions: State of the Art and Applications, Oxford University Press, 2001.

[21] Comsol, *Flow past a cylinder*, `http://www.comsol.com/showroom/gallery/97/`.

[22] R. Costantini, L. Sbaiz, and S. Suesstrunk, *Higher order svd analysis for dynamic texture synthesis*, IEEE Transactions on Image Processing **17** (2008), no. 1, 42–52.

[23] L. De Lathauwer, *Signal processing based on multi-linear algebra*, Ph.D. thesis, K.U. Leuven, Belgium, 1997.

[24] L. de Lathauwer, B. de Moor, and J. Vandewalle, *A multilinear singular value decomposition*, SIAM Journal on Matrix Analysis and Applications **21** (2000), no. 4, 1253–1278.

[25] ———, *On the best rank-1 and rank-$(r_1, r_2, \ldots, r_n)$ approximation of higher-order tensors*, SIAM Journal on Matrix Analysis and Applications **21** (2000), no. 4, 1324–1342.

[26] V. de Silva and L.-H. Lim, *Tensor rank and the ill-posedness of the best low-rank approximation problem*, SIAM Journal on Matrix Analysis and Applications **30** (2008), no. 3, 1084–1127.

[27] C. Eckart and G. Young, *The approximation of one matrix by another of lower rank*, Psychometrika **1** (1936), no. 3, 211–218.

[28] Y.C. Eldar, *Sampling with arbitrary sampling and reconstruction spaces and oblique dual frame vectors*, The Journal of Fourier Analysis and Applications **9** (2003), 77–96.

[29] Y.C. Eldar and T.G. Dvorkind, *A minimum squared-error framework for generalized sampling*, IEEE Transactions on Signal Processing **54** (2006), 2155–2167.

[30] R Everson and L. Sirovich, *The Karhunen-Loève procedure for gappy data*, Journal Opt. Soc. Am. **12** (1995), 1657–1664.

[31] R. Fletcher, *Practical methods of optimization*, vol. Vol. 2 Constrained optimization, John Wiley and Sons, 1981.

[32] D Gabor, *Theory of communication*, Journal IEE (London) **93 part III** (1946), 429–457.

[33] G. H. Golub and C. F. Van Loan, *Matrix computations, third edition*, The John Hopkins University Press, 1996.

[34] J. de Graaf, *Singular values of tensors are lagrange multipliers*, CASA Report No. 11-02, Eindhoven University of Technology, 2011.

[35] B.-Y. Guo, *Spectral methods and their applications*, World Scientific, 1998.

[36] W. Hackbusch, B. N. Khoromskij, S. Sauter, and E. E. Tyrtyshnikov, *Use of tensors in elliptic eigenvalue problems*, This is a preprint.

[37] R. A. Harshman, *Foundation of the prafac procedure: model and conditions for an explanatory multi-mode factor analysis*, UCLA Working Papers in Phonetics **16** (1970), no. 1-84.

[38] F.L. Hitchcock, *The expression of a tensor or a polyadic as a sum of products*, J.Math. Phys. **7** (1927), 164–189.

[39] P. Holmes, J.L. Lumley, and G. Berkooz, *Turbulence, coherent sructures, dynamical systems and symmetry*, Cambridge University Press, U.K, 1997.

[40] K. A. Hoo and D. Zheng, *Low-order control-relevant models for a class of distributed parameter systems*, Chemical Engineering Science **56** (2001), no. 23, 6683 – 6710.

[41] L. Huisman, *Control of glass melting processes based on reduced cfd models*, Ph.D. thesis, Eindhoven University of Technology, The Netherlands, 2005.

[42] M. Ishteva, P. A Absil, S. Van Huffel, and L. De Lathauwer, *Tucker compression and local optima*, Chemometrics and intelligent laboratory systems (2010).

[43] C.R. Johnson, *Positive definite matrices*, American Mathematical Monthly **77** (1970), no. 3, 259–264.

[44] I. T. Jolliffe, *Principal component analysis*, Springer, Berlin, 1986.

[45] B. N. Khoromskij and V. Khoromskaia, *Multigrid accelerated tensor approximation of function related multidimensional arrays*, SIAM Journal on Scientific Computing **31** (2009), no. 4, 3002–3026.

[46] T. G. Kolda, *Orthogonal tensor decompositions*, SIAM Journal on Matrix Analysis and Applications **23** (2001), no. 1, 243–255.

[47] T. G. Kolda and B. W. Bader, *Tensor decompositions and applications*, SIAM Review **51** (2009), no. 3, 455–500.

[48] Tamara Kolda, *A counterexample to the possibility of an extension of the eckart-young low-rank approximation theorem for the orthogonal rank tensor decomposition*, SIAM Journal on Matrix Analysis and Applications **24** (2003), no. 3, 762–767.

[49] V. Kotelnikov, *On the carrying capacity of the ether and wire in telecommunications, material for the first all-union conference on questions of communications*, Izd. Red. Upr. Svyazi RKKA, Moscow (1933).

[50] D. Kressner and C. Tobler, *Krylov subspace methods for linear systems with tensor product structure*, SIAM Journal on Matrix Analysis and Applications **31** (2010), no. 4, 1688–1714.

[51] E. Kreyszig, *Introductory functional analysis with applications*, Wiley, 1989.

[52] K. Kunisch and S. Volkwein, *Galerkin proper orthogonal decomposition methods for a general equation in fluid dynamics*, SIAM Journal of Numerical Analysis **40** (2002), 492–515.

[53] L. de Lathauwer, *Decompositions of a higher-order tensor in block terms - part ii: Definitions and uniqueness*, SIAM Journal on Matrix Analysis and Applications **30** (2008), no. 3, 1033–1066.

[54] D. Leibovici and R. Sabatier, *A singular value decomposition of a k-way array for a principal component analysis of multiway data, ppta-k*, Linear Algebra and its Applications **269** (1998), 307–329.

[55] L.-H. Lim, *What's possible and what's not possible in tensor decompositions - a freshman's view*, Workshop on Tensor Decompositions, July 2004.

[56] Kirby. M., *Geometric data analysis, an empirical approach to dimensionality reduction and the study of patterns*, John Wiley, New York, 2001.

[57] R. J. Marks, *Advanced topics in shannon sampling and interpolation theory*, Springer, 1993.

[58] United Nations Department of Public Information, *UN summit concludes with adoption of global action plan to achieve development goals by 2015*, Press Release, 2010.

[59] J.M. Ortega and W.C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*, Academic Press, New York, 1970.

[60] I. V. Oseledets and E. E Tyrtyshnikov, *Breaking the curse of dimensionality, or how to use svd in many dimensions*, SIAM Journal on Scientific Computing **31** (2009), no. 5, 3744–3759.

[61] I. V. Oseledets and E. E. Tyrtyshnikov, *Recursive decomposition of multidimensional tensors*, Doklady Mathematics **80** (2009), no. 1, 460–462.

[62] ———, *TT-cross approximation for multidimensional arrays*, Linear Algebra and its Applications **432** (2010), 70–88.

[63] A. Papoulis, *Signal analysis*, McGraw-Hill, New York, 1977.

[64] J.R. Partington, *Interpolation, identification and sampling*, Oxford, Clarendon Press, 1997.

[65] The Associated Press, *Apple: ipad sales top 2 million since launch*, `http://www.usatoday.com/`, June 2010.

[66] M. Rewienski, *A trajectory piecewise-linear approach to model order reduction of nonlinear dynamical systems*, Ph.D. thesis, Dept. of Electrical Engineering and Computer Science, MIT, June 2003.

[67] M. Rewienski and J. White, *A trajectory piecewise-linear approach to model order reduction and fast simulation of nonlinear circuits and micromachined devices*, Proceedings of the International Conference on Computer-Aided Design (2001), 252–257.

[68] C. W. Rowley, T. Colonius, and R. M. Murray, *Model reduction for compressible flows using pod and galerkin projection*, Physica D: Nonlinear Phenomena **189** (2004), no. 1-2, 115 – 129.

[69] C.W. Rowley and J.E. Marsden, *Reconstruction equations and the Karhunen-Loève expansion for systems with symmetry*, Journal of Physica D. **142** (2000), 1–19.

[70] C.E. Shannon, *A mathematical theory of communication*, Bell Systems Tech. Journal **27** (1948), 379–423.

[71] L Sirovich, *Turbulence and the dynamics of coherent structures, parts i-iii*, Q. Applied Mathematics XLV **3** (1987), 561–590.

[72] I. Y. Smets, D. Dochain, and J. F. Van Impe, *Optimal temperature control of a steady-state exothermic plug-flow reactor*, AIChE Journal **48** (2002), no. 2, 279–286.

[73] W. Splettstosser, *Sampling approximation of continuous functions with multidimensional domain*, IEEE Transactions on Information Theory **28** (1982), 809–814.

[74] V. Thomee, *Galerkin finite element methods for parabolic problems*, Springer, Berlin, 1997.

[75] L. R. Tucker, *Some mathematical notes on three-mode factor analysis*, Psychometrika **31** (1966), no. 3, 279–311.

[76] M. A. O. Vasilescu and D. Terzopoulos, *Multilinear subspace analysis of image ensembles*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2003.

[77] ———, *Multilinear independent component analysis*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2005.

[78] Compendium voor de leefomgeving, *Energieverbruik per doelgroep*, `http://www.compendiumvoordeleefomgeving.nl/`, July 2010.

[79] Centraal Bureau voor de Statistiek, *De Nederlandse economie 2009*, Centraal Bureau voor de statistiek, 2010.

[80] S. K. Wattamwar, *Identification of low order models for large scale processes*, Ph.D. thesis, Eindhoven University of Technology, The Netherlands, 2010.

[81] S. Weiland and F. van Belzen, *Singular value decompositions and low rank approximations of tensors*, IEEE Transactions on Signal Processing **58** (2010), no. 3, 1171–1182.

[82] E.T. Whittaker, *On the functions which are represented by the expansion of the interpolation theory*, Proceedings Royal Society Edinburgh **35** (1915), 181–194.

[83] K. Willcox, *Unsteady flow sensing and estimation via the gappy proper orthogonal decomposition*, Computers and Fluids **35** (2006), no. 2, 208–226.

[84] A. I. Zayed, *Advances in Shannon's sampling theory*, CRC Press, 1993.

[85] T. Zhang and G. H. Golub, *Rank-one approximation to high order tensors*, SIAM Journal on Matrix Analysis and Applications **23** (2001), no. 2, 534–550.

Summary

Approximation of multi-variable signals and systems: a tensor decomposition approach

Signals that evolve over multiple variables or indices occur in all fields of science and engineering. Measurements of the distribution of temperature across the globe during a certain period of time are an example of such a signal. Multi-variable systems describe the evolution of signals over a spatial-temporal domain. The mathematical equations involved in such a description are called a model and this model dictates which values the signals can obtain as a function of time and space. In an industrial production setting, such mathematical models may be used to monitor the process or determine the control action required to reach a certain set-point. Since their evolution is over both space and time, multi-variable systems are described by Partial Differential Equations (PDEs).

Generally, it is not the signals or systems themselves one is interested in, but the information they carry. The main numerical tools to extract system trajectories from the PDE description are Finite Element (FE) methods. FE models allow simulation of the model via a discretization scheme. The main problem with FE models is their complexity, which leads to large simulation time, making them not suitable for applications such as on-line monitoring of the process or model-based control design. Model reduction techniques aim to derive low-complexity replacement models from complex process models, in the setting of this work, from FE models. The approximations are achieved by projection on lower-dimensional subspaces of the signals and their dynamic laws. This work considers the computation of empirical projection spaces for signals and systems evolving over multi-dimensional domains. Formally, signal approximation may be viewed as a low-rank approximation problem. Whenever the signal under consideration is a function of multiple variables, low-rank approximations can be obtained via multi-linear functionals, tensors. It has been explained in this work that approximation of multi-variable *systems* also boils down to low-rank approximation problems.

The first problem under consideration was that of finding low-rank approximations to tensors. For order-2 tensors, matrices, this problem is well understood. Generalization of these results to higher-order tensors is not straightforward. Finding tensor decompositions that allow suitable approximations after truncation is an active area of research. In this work a concept of rank for tensors, referred to as multi-linear or modal

rank, has been considered. A new method has been defined to obtain modal rank decompositions to tensors, referred to as *Tensor Singular Value Decomposition (TSVD)*. Properties of the TSVD that reflect its sparsity structure have been derived and low-rank approximation error bounds have been obtained for certain specific cases. An adaptation of the TSVD method has been proposed that may give better approximation results when not all modal directions are approximated. A numerical algorithm has been presented for the computation of the (dedicated) TSVD, which with a small adaptation can also be used to compute successive rank-one approximation to tensors. Finally, a simulation example has been included which demonstrates the methods proposed in this work and compares them to a well-known existing method.

The concepts that were introduced and discussed with regard to signal approximation have been used in a system approximation context.We have considered the well-known model reduction method of Proper Orthogonal Decompositions (POD). We have shown how the basis functions inferred from the TSVD can be used to define projection spaces in POD. This adaptation is both a generalization and a restriction. It is a generalization because it allows POD to be used in a scalable fashion for problems with an arbitrary number of dependent and independent variables. However, it is also a restriction, since the projection spaces require a Cartesian product structure of the domain. The model reduction method that is thus obtained has been demonstrated on a benchmark example from chemical engineering. This application shows that the method is indeed feasible, and that the accuracy is comparable to existing methods for this example.

In the final part of the thesis the problem of reconstruction and approximation of multi-dimensional signals was considered. Specifically, the problem of sampling and signal reconstruction for multi-variable signals with non-uniformly distributed sensors on a Cartesian domain has been considered. The central question of this chapter was that of finding a reconstruction of the original signal from its samples. A specific reconstruction map has been examined and conditions for exact reconstruction have been presented. In case that exact reconstruction was not possible, we have derived an expression for the reconstruction error.

Dankwoord

Mijn promotie-periode van vier jaar was in mijn ervaring zo voorbij. Als ik terugkijk, zie ik dat er in vier jaar veel gebeurd is. Ik heb met veel mensen samengewerkt en ben vele ervaringen rijker. Deze plek in mijn proefschrift wil ik graag gebruiken om iedereen die aan deze vier jaar heeft bijgedragen, in welke vorm dan ook, te bedanken.
Siep, we werken nu samen sinds januari 2006 toen ik begon met afstuderen. Ik heb in deze lange periode veel van je geleerd en ik wil je bedanken voor je enthousiasme en je vertrouwen in ons onderzoek. Zonder jou had dit proefschrift er heel anders uit gezien! Ik vind het heel leuk dat je precies op tijd voltijd-hoogleraar bent geworden om mijn eerste promotor te kunnen zijn.
Paul, Ton, bedankt voor jullie coaching en jullie vertrouwen in mij. Ik heb in onze discussies veel geleerd over de wereld binnen en buiten de universiteit. Mijn dank hiervoor is groot.
Mijn onderzoek maakte deel uit van het STW project 'Model Reduction and Control Design for Large-Scale Dynamical Systems'. Ik wil iedereen die betrokken is bij dit project bedanken, met name Mark en Thomas voor de samenwerking en de leden van de gebruikerscommissie voor hun bijdrage aan mijn onderzoek. Mijn dank gaat ook uit naar de studenten met wie ik in het kader van dit project heb samengewerkt, met name Jaron, Lykele, Gerben, Rik en Lennart.
Jan de Graaf, Paul van den Hof en Lieven de Lathauwer, bedankt voor jullie deelname in de kerncommissie en het extra inzicht dat bij mij ontstaan is naar aanleiding van jullie opmerkingen.
Gedurende mijn promotieperiode had ik het voorrecht om een aantal maanden een kleiner onderzoeksproject uit te voeren binnen ASML. Ik wil Wim Coene en de onderzoeksgroep bij ASML bedanken voor de goede samenwerking, ik heb het bij jullie erg naar mijn zin gehad!
Ik heb mijn promotie uitgevoerd als lid van de vakgroep 'Control Systems'. Bij deze wil ik iedereen in de groep bedanken voor de goede samenwerking, de deelname aan de door mij georganiseerde lunch seminars en de gezelligheid tijdens conferenties. Een speciaal woord van dank gaat naar mijn kamergenoten John en Leyla.
Als laatste wil ik mijn familie en vrienden bedanken, voor alles. Ik voel me heel bevoorrecht om jullie in mijn leven te hebben en ik hoop mijn dankbaarheid hiervoor nog vaak richting jullie uit te spreken.

Curriculum Vitae

Femke van Belzen was born on December 15th, 1982 in Stuttgart, Germany. She completed her secondary education at Carolus Borromeus College in Helmond in 2001 and went on to study at Eindhoven University of Technology. She completed her Bachelors (2005) and Masters (2006, with honors) degree in Electrical Engineering from TU/e. Her Masters Thesis was titled *Model reduction through Proper Orthogonal Decompositions for multidimensional systems*. The research for this thesis was carried out in the *Control Systems* group of the department of Electrical Engineering at the TU/e. She continued her research on this topic in the same group as a PhD student in 2007. The results of this research are covered in this thesis titled *Approximation of multi-variable signals and systems: a tensor decomposition approach*.