SYSTEM IDENTIFICATION IN HIGHLY NON-INFORMATIVE ENVIRONMENT

RENDSZERIDENTIFIKÁCIÓ ERŐSEN INFORMÁCIÓHIÁNYOS KÖRNYEZETBEN

Ph.D. Thesis
Sándor Kolumbán
January 2016

Advisors:
Prof. István Vajk
Prof. Johan Schoukens
This thesis is submitted in fulfillment of the requirements for the PhD degree at Budapest University of Technology and Economics and for the Doctor in Engineering degree at Vrije Universiteit Brussel by Sándor Kolumbán.

**Advisors**

Prof. István Vajk
Prof. Johan Schoukens

Prof. Marco Campi
Prof. András Edelmayer
Prof. Paul Van den Hof
Prof. János Levendovszky

**Opponents**

University Of Brescia
Hungarian Academy of Sciences
Eindhoven University of Technology
Budapest University of Technology and Economics

**Members of the jury in Brussels**

Prof. Patrick Guillaume (president)
Prof. Johan Deconinck (vice-president)
Prof. Gerd Vandersteener (secretary)

**Members of the jury in Budapest**

Prof. János Levendovszky (president)
Dr. Bálint Kiss (secretary)
Prof. András Edelmayer
Prof. Marco Campi
Prof. Patrick Guillaume
Prof. Zombory László

**SYSTEM IDENTIFICATION IN HIGHLY NON-INFORMATIVE ENVIRONMENT**

**RENSZERIDENTIFIKÁCIÓ ERŐSEN INFORMÁCIÓHIÁNYOS KÖRNYEZETBEN**
Abstract

System identification has gone a long way in the past century but there are still problems to be solved and new directions waiting to be discovered. The mainstream identification literature contains results that rely on practically infinite amount of data (either measurement data or prior knowledge), which is an acceptable compromise for many applications, but not for all.

The focus in this thesis is put on obtaining statistically reliable uncertainty estimates when the identification task is carried out in a non-informative environment, meaning that the number of observations and that of the estimated parameters is barely enough to make the estimation problem meaningful. As this problem is tackled, a series of questions emerge that serve as the subject of the thesis.

First, a recent method (Sign-Perturbed Sums, SPS) for constructing exact confidence regions is described from a new point of view by defining a generic family of such methods (Data Perturbation methods, DP methods). Exact confidence level hypothesis tests can be constructed in the framework of DP methods where some of the elements of a DP method are determined by the assumed characteristics of the noise that contaminates the measurements. The other elements need to be defined to suit the model structure that is estimated. It is shown that the SPS method is a DP method belonging to symmetrically distributed noise sources. Other DP methods are also given and the structure of the corresponding confidence sets is examined.

DP methods for parameter estimates of linear dynamical systems rely on the solution of polynomial optimization problems. The second part of the thesis deals with how the optimum of such problems can be obtained in light of recent developments in the optimization community.

As computing exact confidence regions turns out to be computationally taxing, it needs to be quantified when can results, asymptotic in the sample count, be trusted. The last part of the thesis describes a method that estimates the uncertainty error introduced by asymptotic results. Based on these estimates it can be decided whether the asymptotic confidence regions are to be trusted or some sort of compensation is needed.
Összefoglaló

A rendszeridentifikáció hosszú utat járt be az elmúlt században de még mindig vannak nyitott kérdések és új irányvonalak, amik felderítésre várnak. Az irodalmi fősodor olyan eredményeket tartalmaz amelyek gyakorlatilag végében mennyiségi információt tételeznek fel (adatmennyiségben vagy előzetes feltételezésként), ami elfogadható kompromisszum bizonyos alkalmazásoknál, de nem minden esetben.

Ebben az értekezésben statisztikailag megbízható bizonytalanság-becslesek előállítására került a hangsúly, amikor az identifikációs feladat erősen információhiányos környezetből származik. Ezalatt azt értjük, hogy a mintaszám és a becsülső paraméterek száma éppen csak lehetővé teszi használható becslés előállítását. A probléma megközelítése során különböző kérdések merülnek fel és a rájuk adott válaszok képezik a tézis tartalmát.

Először egy viszonylag friss konfidenciaalmaz szerkesztési eljárás (Sign-Perturbed Sums, SPS) kerül bemutatásra új nézőpontból, egy általánosabb eljárásosaládot (aDat Perturbációs eljárások, DP eljárások) definiálva. A DP eljárások egzakt konfidenciaszintű hipotézisvizsgálatok. Ezen eljárások néhány elemét a méréseket terhelő zaj feltételezett tulajdonságai határozzák meg, míg másokat úgy kell megválasztani, hogy illeszkedjenek a becsült modellelsalához. Megmutatjuk, hogy az SPS eljárás is egy DP eljárás, ahol szimmetrikus eloszlású zajforrást tételezünk fel. További DP eljárásokat is mutatunk és megvizsgáljuk az eljárásokhoz tartozó konfidenciaalmazok struktúráját is.

A lineáris dinamikus rendszerek paraméterbecsléséhez tartozó DP eljárások építenek egy polinom optimalizálási feladat megoldására. A dolgozat második részében azt vizsgáljuk, hogy a numerikus optimalizálás új fejleményeinek fényében miként lehet számítani az ilyen feladatok optimumát.

Az egzakt konfidenciaalmazok meghatározása számításigényes feladatnak bizonyult, ezért valamilyen módon mérhetővé kell tenni, hogy az aszimptotikus eredmények mikor megbízhatóak. A dolgozat utolsó részében azt vizsgáljuk, hogy a numerikus optimalizálás új fejleményeinek fényében miként lehet számítani az ilyen feladatok optimumát.

Az egzakt konfidenciaalmazok meghatározása számításigényes feladatnak bizonyult, ezért valamilyen módon mérhetővé kell tenni, hogy az aszimptotikus eredmények mikor megbízhatóak. A dolgozat utolsó része egy olyan eljárást mutat be, amivel becsülső az aszimptotikus eredmények használatából fakadó bizonytalansági hiba. Ezen becslésre alapozva eldönthető, hogy egy aszimptotikus konfidenciaalmaz megbízható, vagy valamilyen további korrekciót igényel.
Acknowledgments

The content of this dissertation is the result of work on the topic mentioned in the title over an extended period of time. It demanded a lot from me to finish this dissertation but it would not have been possible without the contribution and support of many people, to whom I am most grateful.

Firstly, I am indebted to my supervisors István Vajk and Johan Schoukens. The personality of István Vajk and the subjects he was interested in made me choose him as an adviser for my PhD work. His initial topic proposal, deep insights, the long and sometimes late consultations and his way above average standards helped me a great deal along the way. I met Johan Schoukens at a summer school and it turned out that we could collaborate nicely. My short visits to Brussels always proved to be a major step along the way. Mostly because guests are left to work without any disturbance, but also because he could always fit in a short discussion whenever I felt stuck.

Looking at the contents of this thesis I know that I would not have been capable of completing it without the dedicated work of some of my teachers whom I must name here in chronological order: Levente Bakos, Emőke Szélyes, Sándor Török, Géza Dávid, András Vetier, Márton Balázs and Bálint Tóth.

I would like to thank the community of the BME AUT for all the memories, Hassan Charaf for providing the appropriate circumstances and Jenő Hetthéssy for all the soul lifting conversations we had.

I owe a great deal of gratitude to my ‘comrade’ Ákos Dudás, we sure had to push through allot together in the past years. Helping each other and knowing that I am not alone with my troubles made a difference for sure.

Last but not least, I am grateful to my parents Imola and Gábor for their unconditional love and support.
Köszönetnyilvánítás

Az értekezés tartalma a címben említett témának szentelt munka eredménye, ami egy hosszabb peridódust ölelt fel. Tőlem is sokat követelt az értekezés elkészítése, de sokak hozzájárulása és támogatása nélkül, akiknek ezért kifejezetten hálás vagyok, lehetetlen lett volna.

Először is köszöntettel tartozom témavezetőimnek, Vajk Istvánnak és Johan Schoukensnek. Vajk István személyiségé és a számára érdekes tématerületek miatt válaszottam öt PhD munkám témavezetőjének. Az elsődleges témaválasztása, mélyrehatoló meglátásai, a hosszú és nem ritkán későig tartó konzultációk valamint a jóval átlag feletti igényessége sokat segített az út folyamán. Johan Schoukens-el egy nyári iskola során találkoztam és úgy alakult, hogy jól tudtunk együtt dolgozni. A brüsszeli rövid látogatásaim mindig jelentős előrelépésnek bizonyultak. Többnyire azért, mert a vendégek nyugodtan dolgozhatnak mindenféle zavaró hatás nélkül, de azért is mert egy rövid eszmecsere mindig belefért neki, ha elakadtam.

A dolgozatra tekintve tudom, hogy nem lettem volna képes annak megalkotására néhány tanárom odaadó munkája nélkül, akiket mindenképp meg kell említenem itt, időrendi sorrendben: Bakos Levente, Szélyes Emőke, Török Sándor, Dávid Géza, Vetier András, Balázs Márton és Tóth Bálint.

Meg szeretném köszönni a BME AUT közösségének a közös emlékeket, Charaf Hassannak a megfelelő körülmények megteremtését és Hetthéssy Jenőnek a lélekemelő beszélgetéseinket.

Nagyon sokkal tartozom ‘bajtársamnak’ Dudás Ákosnak, sok mindennel küzdöttünk meg közösen az elmúlt években. Egymás segítése és a tudat, hogy nem vagyok egyedi a problémáimmal néha határozottan segített.

Végül, de nem utolsó sorban, hálás vagyok szüleimnek Imolának és Gábornak a feltételnélküli szeretetükért és támogatásukért.
Declaration

The content of this thesis is a product of the author’s original work except where explicitly stated otherwise.

This thesis is submitted in fulfillment of the requirements for the PhD degree at Budapest University of Technology and Economics and for the Doctor in Engineering (Doctor in de Ingenieurswetenschappen) degree at Vrije Universiteit Brussel.

Nyilatkozat*

Ahulírott Kolumbán Sándor kijelentem, hogy ezt a doktori értekezést magam készítettem, és abban csak a megadott forrásokat használtam fel. Minden olyan részt, amelyet szó szerint, vagy azonos tartalomban, de átfogalmazva más forrásból átvettem, egyértelműen, a forrás megadásával megjelöltem.

Ez az értekezés a Budapesti Műszaki és Gazdaságtudományi Egyetem által adományozható PhD fokozat, valamint a Vrije Universiteit Brussel által adományozható Doctor in Engineering (Doctor in de Ingenieurswetenschappen) fokozat megszerzési feltételeinek teljesítéséért kerül benyújtásra.

Budapest, 2015.

(Sándor Kolumbán)

*A bűrálatok és a védésről készült jegyzőkönyv a későbbiekben a Dékáni Hivatalban elérhetőek.
Table of Contents

List of Figures  ix

List of Tables  x

List of Symbols and Abbreviations  xi

Chapter 1
Introduction  1
1.1 Thesis Outline  1
1.2 Basics of System Identification  4
  1.2.1 Definition of the Model Class  6
  1.2.2 Obtaining Estimates  9
  1.2.3 Evaluating the Uncertainty of the Estimates  10
1.3 Motivating Example and Problems  12
1.4 Highly Non-Informative Environment  14

Chapter 2
Data Perturbation Methods  16
2.1 Introduction to Distribution Free Statistics  17
2.2 Example  20
  2.2.1 Problem Data  20
  2.2.2 Membership Tests  22
    2.2.2.1 Membership Test of $\theta_0$  22
    2.2.2.2 Membership Test of $\theta_{LS}$  24
    2.2.2.3 Membership Test of Faraway Models  25
  2.2.3 The Confidence Region  26
  2.2.4 Conclusions  26
2.3 Definition of Data Perturbation Methods  27
  2.3.1 Probability Prerequisites  30
  2.3.2 The General Framework  34
2.3.3 SPS as a DP Method ........................................... 38
2.3.4 Transformation Invariant Noise Sources .................. 39
  2.3.4.1 Symmetric Noise Sequences ............................. 39
  2.3.4.2 Exchangeable Noise Sequences ......................... 40
  2.3.4.3 Power Defined Noise Sequences ......................... 41
2.3.5 About Performance Measures in General ................. 42
2.3.6 Evaluating the Risk of a Model ............................ 45
2.4 DP Methods for Linear Regression Problems ................. 46
  2.4.1 Performance Measure for Noise Distributions Invariant Under Orthogonal Transformation ................ 47
  2.4.2 An Instructive Example for a Performance Measure .... 52
  2.4.3 The Price of Assumptions .................................. 56
2.5 DP Methods for Linear Dynamical Systems .................. 60
  2.5.1 Different Interpretations of the Performance Measure .. 61
  2.5.2 Gradient Based Performance Measure ...................... 64
  2.5.3 The Output Error Model ................................... 67
  2.5.4 General Noise Models ..................................... 71
  2.5.5 Computational Complexity .................................. 72
2.6 Place in the Identification Workflow ........................ 77
2.7 Remarks and Contribution Summary ........................... 79

Chapter 3

Globally Optimal Estimation of Polynomial Models 82
3.1 Lasserre Hierarchies for Polynomial Optimization Problems .......... 84
  3.1.1 Moment and Localizing Matrices .......................... 85
  3.1.2 Reformulation as SDP ...................................... 87
  3.1.3 Convergence of the Hierarchy .............................. 87
  3.1.4 Sparsity .................................................... 88
3.2 The Identification Problem as POP ............................ 90
  3.2.1 Parameterization ........................................... 90
    3.2.1.1 Monic Parameterization .............................. 90
    3.2.1.2 Norm Constrained Parameterization ................ 91
  3.2.2 Reformulations of the LS Optimization Problem ............ 92
    3.2.2.1 Unconstrained Formulation .......................... 92
    3.2.2.2 Constrained Formulation ............................ 93
3.3 General Method and the Identification POPs .................... 94
  3.3.1 Convergence of Relaxations ............................... 94
    3.3.1.1 Additional Constraints for Convergence .............. 94
    3.3.1.2 Rate of Convergence ................................. 96
    3.3.1.3 Extracting Optimizers ............................... 96
  3.3.2 Size of the SDPs ........................................... 97
    3.3.2.1 Unconstrained Formulation .......................... 97
    3.3.2.2 Constrained Formulation ............................ 99
Chapter 4
Effect of Finite Measurement Length on Uncertainty Estimation

4.1 Illustration of Finite Sample Effects
4.2 Finite Sample Variance Loss
4.3 Approximation of Expected Variance Loss in LTI Systems
   4.3.1 Approximation Algorithm
   4.3.2 Concrete Approximation Example
      4.3.2.1 A Closer Look on the Approximation Algorithm
      4.3.2.2 Approximating the Wishart Parameters
      4.3.2.3 Deducing the Expected Variance Loss
   4.3.3 Properties of the Approximation
   4.3.4 Initial Conditions and Variance Loss
   4.3.5 Filtered Noise Input
   4.3.6 Adjusting Asymptotic Confidence Sets for Variance Loss
4.4 Conclusions

Chapter 5
Concluding Remarks

Appendix A
Statement of contribution
   A.1 Preliminaries and objectives
   A.2 New scientific results

Appendix B
Variance of Filtered White Noise Output

Bibliography
## List of Figures

1.1 Dependence graph between chapters ........................................ 2  
1.2 Schematic diagram of systems and signals ............................... 5  
1.3 Generalized Box-Jenkins model structure ................................. 9  
1.4 Finite sample identification example ...................................... 13  

2.1 Example data set ................................................................... 21  
2.2 Example confidence set $C_{2/3}^{\Pi}$ .................................... 26  
2.3 Shapes of power functions .................................................... 43  
2.4 Confidence regions corresponding to different DP setups and the  Gaussian confidence region .............................................. 51  
2.5 Comparison of confidence regions belonging to different performance  
measures .................................................................................. 55  
2.6 Power functions for different assumptions with different input ...... 57  
2.7 Difference of power functions for different assumptions with different  
input ......................................................................................... 59  
2.8 Not connected confidence region from gradient based DP method . 66  
2.9 Membership tests for $\theta$ around $\hat{\theta} = \hat{\theta}^{(1)}$ .................. 70  
2.10 Justification of the principal connected component heuristic ....... 75  

3.1 Chordal extension of the CSP graph of the equality constrained  
identification POP .................................................................... 102  
3.2 Difference between cost function values obtained using oe and $S_3$ . 104  

4.1 Normalized variances of parameters against sample count. ........ 109  
4.2 Squared error of $\hat{f}_1$ for different sample counts $N$ ............... 112  
4.3 Empirical and approximate expected variance loss of parameters. . 115  
4.4 The variance of $\hat{f}_1$ for filtered white noise input against different  
filtering frequencies and sample counts $N/T_0$ ........................... 122  
4.5 The estimated expected variance loss $\bar{\alpha}_N(\hat{f}_1)$ for filtered white noise  
input against different filtering frequencies and sample counts $N/T_0$ . 123
List of Tables

2.1 Sample data .................................................. 21
2.2 Perturbed data sets corresponding to $\theta = \theta_0$ ......................... 23
2.3 Perturbed data sets corresponding to $\theta = \theta_{LS}$ ......................... 24
2.4 Perturbed data sets corresponding to $\theta \to \infty$ ......................... 25
2.5 Pros and cons of different performance measures ....................... 64
List of Symbols and Abbreviations

SDP  SemiDefinite Programming, Section 1.1, p. 3

POP  Polynomial Optimization Problem, Section 1.1, p. 3

$u$ Input signal, Section 1.2, p. 6

$y$ Output signal, Section 1.2, p. 6

$z$ Forward shift operator, Section 1.2, p. 6

LTI Linear Time Invariant, Section 1.2, p. 7

$N$ Sample count, Section 1.2, p. 7

$δ[k]$ The discrete unit impulse at $k = 0$, Section 1.2, p. 7

$T(z)$ A transient model, Section 1.2, p. 7

$\mathbb{R}$ The set of real numbers, Section 1.2, p. 8

$I_n$ $n$ dimensional identity matrix, Section 1.2, p. 8

$Y$ $N$ dimensional vector of observed outputs, Section 1.2, p. 8

$U$ $N$ dimensional vector of observed inputs, Section 1.2, p. 8

$Γ$ Extended observability matrix, Section 1.2, p. 8

$H$ Lower triangular matrix of the Markov parameters, Section 1.2, p. 8

SISO Single-Input Single-Output, Section 1.2, p. 8

$G(z)$ Plant model, Section 1.2, p. 9

$H(z)$ Noise model, Section 1.2, p. 9
\( n_{\mathcal{X}} \)  Degree of polynomial \( \mathcal{X} \), Section 1.2, p. 9
\( \theta \)  Vector of unknown parameters, Section 1.2, p. 9
\( n_{\theta} \)  Dimension of the parameter vector \( \theta \), Section 1.2, p. 9
\( \theta_0 \)  Vector of nominal parameters, Section 1.2, p. 9
\( \hat{\theta} \)  Estimate for \( \theta \), Section 1.2, p. 9
\( V(\cdot) \)  Cost function for identification, Section 1.2, p. 10
PEM  Prediction Error Method, Section 1.2, p. 10
LS  Least Squares, Section 1.2, p. 10
ML  Maximum Likelihood, Section 1.2, p. 10
\( \mathcal{N}(m, \Sigma) \)  Gaussian distribution with mean \( m \) and covariance \( \Sigma \), Section 1.2, p. 11
i.i.d.  Independent and Identically Distributed, Section 1.2, p. 12
\([n]\)  The set of numbers \([n] = \{1, \ldots, n\}\), Section 2.1, p. 17
SPS  Sign-perturbed Sums, Section 2.1, p. 19
\( \mathcal{U}_G \)  Uniform distribution over \( G \), Section 2.3.1, p. 30
\( \mathbb{P}(A) \)  Probability of an event \( A \).
FIR  Finite Impulse Response, Section 2.5.3, p. 68
LMI  Linear Matrix Inequality, Section 3.1, p. 88
SOS  Sum of Squares, Section 3.3, p. 95
Chapter 1

Introduction

In this first chapter, we provide an overview of the contents of the dissertation in Section 1.1 and we give a short introduction to system identification. The classical problems of the field are introduced in Chapter 1.2 along with the basic notation that we are going to use. To provide motivation for what is going to be presented in the rest of the dissertation, an example is given in Chapter 1.3, emphasizing those issues that we are going to pursue.

1.1 Thesis Outline

The dissertation is organized into four chapters, each of these having a well defined scope. The current Chapter 1 contains the introductory material needed for the following chapters; its content serves as a starting ground for the rest of the material. All following chapters depend on notation and notions introduced here, so this should be read first. However, readers familiar with system identification may skip this part and directly go to any other chapter of interest, as most notation follow the standards in the literature.

The remaining three chapters, although relying on Chapter 1, can be read independently of each other. References to earlier chapters are restricted to drawing motivation from problems raised previously, but they do not impact the consistency of the text. The dependence of chapters on each other is given in Figure 1.1 with solid arrows corresponding to strong, while dashed lines denoting only minor dependencies.
In Section 1.2 the basic problems of system identification are introduced. We start from single input single output (SISO) systems described by a linear differential equation with constant coefficients and we make our way to the description of discrete systems. The difference equation and the state space description of these systems are introduced and the generalized Box-Jenkins model structure is presented, which is going to be used in the sequel. Some of the classical results of system identification are presented afterwards, that can serve as a starting point for our work. Using a simple toy example, Section 1.3 illustrates what are some of the possible shortcomings of the classical identification approach. The goal of this section is to pinpoint those problems that we are going to address throughout the rest of the dissertation.

The problem that is going to be singled out is that when the distribution of the noise that contaminates our measurements is unknown and the amount of available data samples is finite, then uncertainty estimates delivered by the classical asymptotic theory can be unreliable.

Chapter 2 addresses the first of these questions, namely, how can the strong assumption about exact knowledge of the noise distribution be relaxed. At the beginning of the chapter, Section 2.1 presents the idea of distribution free statistical hypothesis testing along with a short overview of different methods that were invented in this spirit. One of these methods is the method of Sign-Perturbed Sums (SPS), that served as an inspiration to our work. Section 2.3 describes a framework of distribution free methods, that we call data perturbation (DP) methods. Before giving the precise description of DP methods Section 2.2 contains a detailed example to illustrate the steps and capabilities of these methods. The general definition is given in Section 2.3.2 in a way that different modules of this
framework can be adjusted to the problem at hand. We will use the SPS method as a showcase in Section 2.3.3 to illustrate how a DP method is constructed.

There are two major components in a DP method. One of these depends heavily on the joint distribution of the contaminating noise. To help grasp the condition that defines what kind of noise distributions allow the application of DP methods, we provide examples of such distributions in Section 2.3.4. The other important component is the so called performance measure that is needed in a DP method. Section 2.3.5 contains general discussions about these performance measures.

At this point we are capable of creating distribution free hypothesis tests with exact confidence levels and we shift our focus towards creating these methods such that the confidence regions belonging to these tests have desirable properties. We start with analyzing the confidence region of DP methods constructed for linear regression problems in Section 2.4, where we present sufficient and necessary conditions for having bounded and connected confidence regions. Section 2.5 describes DP methods for linear dynamical system models. The analysis for this model structure is more complicated so we start with examining the case of output error models first in Section 2.5.3 then we extend the analysis to general Box-Jenkins models in Section 2.5.4. The computational complexity of the presented methods is elaborated in Sections 2.5.5, followed by some general comments about the presented methodology in Section 2.7.

As it will be shown in Section 1.2, the identification problem can, in many cases, be interpreted as an optimization problem with multiple local optima. Obtaining the globally optimal solution of such problems is interesting in itself, but so are the confidence region characterizations, defined in Chapter 2, that rely on obtaining these values. In Chapter 3 we discuss the possibilities offered by recent advances in numerical optimization. Section 3.1 describes the Lasserre hierarchy, which is a sequence of semidefinite programs (SDP) of growing size, whose optimal value approximates that of an original polynomial optimization problem (POP). Section 3.2 presents the identification problem in the form of different POPs. The computational complexity of the “naive” approach is discussed in Section 3.3. The size of the corresponding SPDs turns out to be exponential in every complexity parameter of the identification problem. The sparsity of the identification POP is exploited in Section 3.4 to reduce the size of these SDPs. This enables the reduc-
We will see in Section 2.5.5 that computing exact confidence regions is computationally taxing; also, the description of such confidence regions is hard to work with afterwards. Asymptotic confidence regions have the advantage of being efficiently calculable and having a compact description, so they should be used whenever they are reliable. The scope of Chapter 4 is to provide means to quantify the reliability of the asymptotic results. We pinpoint those imperfections of the asymptotic results that we want to measure is Section 4.1. The notion of finite sample variance loss for general finite sample estimation problems is defined in Section 4.2. Based on this measure of inaccuracy we suggest a re-scaling step in Section 4.3 that can be applied to the asymptotic ellipsoids to improve the accuracy of their confidence level.

Chapter 5 briefly overviews the findings of this thesis, while Appendix A contains a detailed extract of this thesis, explicitly stating its scientific contribution.

Throughout the dissertation, the notation follows some typesetting guidelines to help the interpretation of mathematical expressions. Scalar values are usually denoted by Latin or Greek lower case letters ($c, a_i, \alpha$), vectors are denoted with boldface lowercase letters ($\mathbf{\theta}, \mathbf{x}_i$) or upper case letters ($Y, U$), matrices are denoted using boldface upper case letters ($\mathbf{X}$). The difference between a random variable and an actual realization is not reflected by the notation; it is either evident from the context or it is explicitly stated when the notation is introduced. Calligraphic capital letters ($\mathcal{A}$) will denote both sets and polynomials. Exceptions from these rules are always made explicit.

References to my own work are given in IEEE style, like [1], while references to other authors works are given in a more verbose style, like [Ljung, 1999].

1.2 Basics of System Identification

We want to understand the world around us and sometimes we would also like to influence certain properties of it. Usually it is not possible to directly affect the properties that we want to control but we can take certain actions, and these actions might have the desired effect on the property of interest through causal
relationships. A simple example would be when we want to prepare a cup of tea, we want to have the temperature of the water in a teapot to be close to boiling temperature. We cannot set the temperature of the water as it is, but placing it on the stove and manipulating the intensity of the flame under it, we know that eventually we will be able to get the water to the right temperature to enjoy our cup of tea. The usual model to describe such systems is given in Figure 1.2, where we acknowledge that: there are observed outputs that are of interest to us (the temperature of the water); there are observed inputs (the knobs of the stove) that we have access to and we can use to influence these outputs; there can be other unobserved inputs that influence the system but we do not know their exact value (temperature in the room, ventilation); and most importantly the system dynamics itself that describes how all these inputs govern the outputs exactly.

The now classical, books \cite{SoderstromStoica1989}, \cite{Ljung1999} and \cite{PintelonSchoukens2012} give a more than adequate description to how estimates of dynamical systems can be obtained based on sampled data. Here we are presenting shortly the main concepts for the sake of introduction and notation.

The aim of system identification is to provide an approximate description of the mapping between the observable inputs and outputs based on measured data. This mapping can never be perfect for many reasons, but we aim for the achievable best description. The obtained models for the system can be used in many ways: amongst others controlling the outputs by appropriately changing the inputs, delivering predictions for future values of the outputs or detecting changes in the system dynamics over time.
1.2.1 Definition of the Model Class

We are mostly interested in systems that are dynamic in time. This means that at time \( t \) the values of the outputs \( y(t) \) are depending not just on the current values of the input \( u(t) \), but it is also important what happened in the past. Getting back to the teapot example, the temperature of the water is not directly determined by the flame intensity, but it depends on the time profile of that. An elementary group of such dynamical systems is formed by those whose dynamics can be given by a linear differential equation with constant coefficients

\[
y^{(n)}(t) + \theta_1 y^{(n-1)}(t) + \cdots + \theta_{n-1} y^{(1)}(t) + \theta_n y(t) = \\
\theta_{n+1} u^{(m-1)}(t) + \cdots \theta_{n+m-1} u^{(1)}(t) + \theta_{n+m} u(t)
\]  

(1.1)

where \( .^{(k)} \) denotes the \( k \)-th time derivative of a signal. The goal of system identification is to find estimates of the unknown parameter vector \( \theta = [\theta_1 \cdots \theta_{n+m}]^T \) based on measured values of \( y(t) \) and \( u(t) \). In the setup, that we are going to be concerned with, only samples of the signals are available from discrete time instances. Assuming equidistant sampling of the continuous signals, with sampling time \( T_s \), we have \( u[k] = u(kT_s) \), and \( y[k] = y(kT_s) \) similarly. After further assumptions on the inter-sample behavior of the signals it is possible to obtain mathematical models that describe the behavior of the sampled signals as well

\[
y[k] + \theta_1 y[k - 1] + \cdots + \theta_{n-1} y[k - n + 1] + \theta_n y[k - n] = \\
\theta_{n+1} u[k] + \cdots \theta_{n+m-1} u[k - m + 2] + \theta_{n+m} u[k - m + 1]
\]  

(1.2)

with a different meaning of the parameter vector \( \theta \). By defining the forward shift operator \( z \) acting on discrete time signals as \( z^{-1} u[k] = u[k - 1] \), the difference equation (1.2) can be written using polynomials of \( z \) as

\[
\mathcal{D}(\theta, z)y[k] = \mathcal{N}(\theta, z)u[k]
\]  

(1.3)

with the polynomials

\[
\mathcal{D}(\theta, z) = 1 + \theta_1 z^{-1} + \cdots + \theta_{n-1} z^{-n+1} + \theta_n z^{-n}
\]  

(1.4)
\[ N(\theta, z) = \theta_{n+1} + \cdots \theta_{n+m-1}z^{-m+2} + \theta_{n+m}z^{-m+1} \] (1.5)

To ease the readability of the notation, different parts of the parameter vector \( \theta \) will be noted in a way that reflects the structure of the parameter vector. By defining the notation
\[ D(\theta, z) = 1 + \theta_1 z^{-1} + \cdots + \theta_n z^{-n} = 1 + d_1 z^{-1} + \cdots + d_n z^{-n} \] (1.6)
\[ N(\theta, z) = \theta_{n+1} + \cdots \theta_{n+m}z^{-m+1} = n_0 + \cdots + n_{m-1}z^{-m+1} \] (1.7)

the parameter vector becomes \( \theta = [d_1, \ldots, d_n, n_0, \ldots, n_{m-1}] \). The transfer function of such a model is defined as
\[ y[k] = H(\theta, z)u[k] \quad H(\theta, z) = \frac{N(\theta, z)}{D(\theta, z)} \] (1.8)

The dependence of the model structure \( H \), or the polynomials \( N \) and \( D \) on the model parameters \( \theta \) will no longer be explicitly noted to shorten notation, except when it contributes to understanding the expressions.

The impulse transfer function (or the difference equation) describes the relation between the input and output of a discrete time linear time invariant (LTI) system when enough past data is available. However, in practical situations, the measured samples start at a given time instant, when the past is not known. In order to give the relation between inputs and outputs that are not infinite in both directions of time, the initial conditions need to be introduced. If the input and output is available for the time interval \( k = 1, \ldots, N \), where \( N \) denotes the sample count, then the measurement can be described with the equation
\[ y[k] = T(z) \delta[k] + H(z)u[k] \quad k = 1, \ldots, N \] (1.9)
where \( y[k] = u[k] = 0 \) for \( k \leq 0 \), \( T(z) \) is the transient depending on the initial condition of the system at time \( k = 0 \) and \( \delta[k] \) is the unit impulse at \( k = 0 \). The transient model is of the form
\[ T(z) = \frac{X_0(z)}{D(z)} \quad X_0(z) = x_0,1q^{-1} + \cdots + x_{0,n}q^{-n} \] (1.10)
where the coefficients of $X_0$ depend on the model parameters in $H$ and the past of the system up until time $k = 0$.

Another typical representation of such a system can be given with its state space representation, where the past of the system up until time $k$ is encoded in a state vector $x[k]$ and the evolution of this state variable and output is determined as

\[
\begin{align*}
x[k + 1] &= Ax[k] + bu[k] \\
y[k] &= c^T x[k] + du[k]
\end{align*}
\]  

(1.11)

(1.12)

where $x, b, c \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$ with $\mathbb{R}$ denoting the set of real numbers. A given state space description $(A, b, c^T, d)$ represents a system $P(z) = c^T (zI_n - A)^{-1}b + d$

where $I_n$ is the $n$ dimensional identity matrix. As the number of parameters in the state space description is larger than that in the impulse transfer function, there are many state space descriptions representing the same system. They are identical up to a full rank linear transformation of the their state vectors $x$. The full description of a data record of length $N$, similar to (1.9) can be given as

\[
Y = \Gamma x[0] + HU
\]

(1.13)

where $Y = (y[k])_{k=1}^N \in \mathbb{R}^N, U = (u[k])_{k=1}^N \in \mathbb{R}^N, x[0]$ is the initial condition of the system, $\Gamma$ is the extended observability matrix

\[
\Gamma = [c (c^T A)^T (c^T A^2)^T \cdots (c^T A^{N-1})^T]^T
\]

(1.14)

and $H$ is the lower triangular Toeplitz matrix with first column

\[
H(:, 1) = [d (c^T b)^T (c^T Ab)^T \cdots (c^T A^{N-2}b)^T]^T
\]

(1.15)

We are going to focus on models that are interconnections of such single-input single-output (SISO) LTI systems in the form given in Figure 1.3. The plant model between the observable input $u$ and output $y$ is defined with the impulse transfer
function \( G(z) = \frac{B(z)}{F(z)A(z)} \), while the noise model between the unobservable input \( e \) and output \( y \) is defined as \( H(z) = \frac{C(z)}{D(z)A(z)} \). Throughout the dissertation it is going to be assumed that the polynomials \( A, C, D, F \) are monic

\[
A(z) = 1 + a_1 z^{-1} + \cdots + a_n z^{-n_A}
\]  

(1.16)

and similarly for all others with the exception of \( B \)

\[
B(z) = b_0 + b_1 z^{-1} + \cdots + b_n z^{-n_B}
\]  

(1.17)

The degree of a polynomial \( \mathcal{X} \) is denoted as \( n_{\mathcal{X}} \). The model structure described above is called the generalized Box-Jenkins model structure.

We are going to focus on estimation problems where data samples are available for the signals \( u \) and \( y \), but not for \( e \). The parameter vector \( \theta \) is formed from the unknown coefficients of the polynomials appearing in the discussed model and the possibly unknown initial conditions. The number of unknown parameters is going to be denoted as \( n_\theta \). We will assume that the degree of the system polynomials (i.e. the model order) is chosen such that the “original” system can be represented in the chosen model class with parameter vector \( \theta_0 \).

### 1.2.2 Obtaining Estimates

We will think about the unmeasured signal \( e \) as disturbance and we are looking for a model estimate \( \hat{\theta} \) that can explain the data record assuming the smallest possible disturbance. If a set of measurements of \( u \) and \( y \) is available then a candidate \( \varepsilon(\theta) \) for the disturbance can be calculated for every model candidate \( \theta \) that ensures the consistency of the dataset and the model structure. \( \varepsilon \) depends on the registered
data \( u \) and \( y \) as well, but we omit this to ease notation. The estimate based on \( N \) samples can be obtained as

\[
\hat{\theta}_N = \arg\min_{\theta} V_N(\theta) \quad V_N(\theta) = \frac{1}{2N} \sum_{k=1}^{N} \varepsilon^2(\theta)[k]
\] (1.18)

where \( V_N(\theta) \) is defined in a way that it penalizes large values for the disturbance \( \varepsilon(\theta) \) needed to make a given parameter vector \( \theta \) consistent with the dataset. This is the prediction error minimization (PEM) approach with the least squares (LS) cost function. If \( e \) is assumed to be a stochastic signal, then by appropriately changing the cost function \( V_N(\cdot) \) the PEM estimate can also be a maximum likelihood (ML) estimate.

A different approach is to rely on the state space representation of the system and use projection techniques to obtain an estimate. These are the so called subspace methods. The only detail of such methods, that is going to be important in the sequel, is that subspace methods deliver estimates of the state space description. Since this description is not unique, any subspace description up to full rank linear transformation of its state space should be an equally good estimate.

To present all parts of the identification spectrum, it needs to be mentioned that beside the above mentioned parametric methods (that seek directly the parameter vector) there are other methods that calculate an intermediate representation of the system (i.e. its frequency response function or its impulse response); these are called non-parametric methods. Arriving to parametric estimates from the non-parametric ones is possible in a second round of approximation.

### 1.2.3 Evaluating the Uncertainty of the Estimates

Beside delivering the estimate \( \hat{\theta}_N \), it is an integral part of the identification process to somehow quantify the reliability of this estimate. Quality tags can be assigned to an estimate in many different ways, but arguably the most prominent one of these is when the disturbances are considered stochastic and the quality estimate is derived from the distribution of the estimate in this stochastic environment. The standard way of doing this is to consider an infinite amount of measurement data and calculate the asymptotic distribution of the estimate. After proper scaling, as a
consequence of the central limit theorem, this distribution converges to a Gaussian
distribution with some covariance. This covariance itself, or ellipsoidal confidence
regions constructed based on it can serve as reliability tags for the estimate.

For PEM estimates, the calculation of the asymptotic covariance goes as fol-
lows. Assuming that as the amount of data increases towards infinity the delivered
estimate converges to $\theta_0$. This makes it possible to show that the equality

$$\hat{\theta}_N - \theta_0 = -V''_{\infty}^{-1}(\theta_0)V'_{N}(\theta_0)$$  

will hold as $N \to \infty$. By approximating the derivative of the cost function $V'_{N}(\theta_0)$
with a sum of independent and identically distributed random variables it can be
shown that

$$\sqrt{N} (\hat{\theta}_N - \theta_0) \overset{d}{\to} \mathcal{N}(0, V''_{\infty}^{-1}QV''_{\infty}^{-1}) \quad Q = \lim_{N \to \infty} \mathbb{E} \left( NV'_{N}(\theta_0)V'_{N}(\theta_0)^T \right)$$

where $\overset{d}{\to}$ means convergence in distribution and $\mathcal{N}(m, \Sigma)$ is the Gaussian distri-
bution with mean $m$ and covariance $\Sigma$. This limiting distribution can be used to
evaluate the reliability of the estimate.

The core motivation of this dissertation is drawn from the existence of situations
where conclusions based on the asymptotic results are not reliable. Nonetheless,
it needs to be acknowledged that the asymptotic theory has a lot to offer and by
no means is it considered entirely discovered. In order to give a glimpse into this
world, we select some of the interesting results. An intuitive principle in estimation
is that if identification is carried out within the same model class (i.e. Box-Jenkins
models) with increasing model complexity (i.e. higher order model polynomials),
then the variance of the estimates will increase. This has to do with the fact
that information about more parameters needs to be extracted from the same
amount of data, thus their reliability is expected to decrease. This was formally
proven already in the 70’s ([Box and Jenkins, 1990], [Gustavsson et al., 1977]). In
light of this an interesting result is reported in [Ljung, 1985] where bounds are
derived for the variance that are asymptotic both in the number of samples $N$
and the model complexity $n_\theta$, and these are independent of the model class (they
depend on input and noise spectra and the ratio between model complexity and
sample count). Many different results appeared in the same spirit since. Similar bounds were derived in [Ninness and Hjalmarsson, 2003] for any fixed model order $n_\theta$, which are asymptotic only in the sample count $N$. These bounds are used mostly not for uncertainty estimation, but also for experiment design purposes. Among many such results one that had an impressive impact also in practice is [Zhu, 2009], [Zhu, 2001]. Recent work also exists on asymptotic theory, such as [Hjalmarsson and Mårtensson, 2011], where the asymptotic variances are derived in a fundamentally different way compared to [Ljung, 1985] and the derived bounds are also sharper.

1.3 Motivating Example and Problems

As it was mentioned in Section 1.2.3, the most common way of evaluating the reliability of estimates is to rely on their asymptotic distribution. The goal of this section is to illustrate the possible shortcomings of the asymptotic theory on a tailored example.

To this end we simulated a first order output error model with nominal parameters $\theta_0 = [b_1 = 0.7769, f_1 = -0.2231]$

$$y[k] = \frac{0.7769 z^{-1}}{1 - 0.2231 z^{-1}} u[k] + e[k]$$

and the driving noise values $e[k]$ were independent and identically distributed (i.i.d.) exponentially distributed random variables with mean $1/30$ (resulting in a signal to noise ratio $\sim 2 \cdot 10^{-3}$). A sample dataset is shown in Figure 1.4a, where the measurement points are connected just for the sake of illustration. Figure 1.4b shows the estimates obtained from 10000 independent realization of the driving noise. Points that are colored blue fall within the asymptotic confidence ellipsoid corresponding to confidence level 0.8. This means that approximately 80% of the points should be included in this ellipsoid, however actually only 72% of the estimate falls there. In order to enlarge the ellipsoid to contain exactly 80% of the estimates, we need to include also estimates that are marked with orange color. The remaining 20% of the estimates is given with a lighter orange and the nominal value $\theta_0$ is shown in yellow.
Although the asymptotic theory has its merits, in certain situations the results delivered by it are not reliable enough. The underlying causes of this are:

i. the method involves approximating a sum of dependent random variables with independent ones.

ii. the distribution of these variables is assumed to be Gaussian when it is not.

iii. the mapping from the measurements to the estimates is not linear, which again distorts the distribution of the estimates, so it is not Gaussian.

The strength of the asymptotic results lies in the fact that all these effects become invisible as $N \to \infty$. However, they can have a major impact while $N$ is small. In the example above the relative error in the estimated variance attributed to these effects was 10%.

The remainder of this dissertation tries to deal with this problem in two ways:

i. by providing means to create confidence regions for estimates that have exact confidence level for any finite value of $N$. This is done in Chapter 2 and the contents of Chapter 3 are also related to this.

ii. by quantifying the reliability of the asymptotic results and possibly adjusting them to improve their confidence. This is the content of Chapter 4.
1.4 Highly Non-Informative Environment

After introducing what system identification is and showing a simple example we are ready to explain the choice for the title.

There is a certain interplay between different circumstances of the identification scenario that influence the reliability of the results of the identification process. We are going to focus on three such circumstances, which are: the assumptions that we make about the distribution of the noise, the sample count $N$ and the chosen estimator that provides the estimate $\hat{\theta}$.

Many other influencing factors could be listed, such as the choice of the excitation signal or the choice of the model structure, but we are going to consider these as a given and we are going to focus only on the effects of the circumstances mentioned above.

In order to generate classical confidence regions for the estimate $\hat{\theta}$, we need to know the distribution of the estimate. If our knowledge about this distribution is not perfect, but it is still good enough, then we can still construct approximate confidence regions which will inherit this property of being good enough. Our knowledge about the distribution of the estimates depends on the noise distribution assumption, the sample count and the estimator in the following way.

The simplest case is when our assumptions about the noise describe its distribution completely and the estimator transforms the noise in such a way that the distribution of the estimate can be calculated. We consider such an environment to be informative in the sense that confidence regions or other uncertainty estimates are perfectly reliable in such a case. Linear regression problems with the least squares estimator fall into this category if the noise distribution is assumed.

When full noise distribution assumptions are available and the estimator transforms the noise in a way that its distribution cannot be calculated anymore, then we speak about non-informative circumstances. In such situations the distribution of the estimate is approximated and uncertainty quantification is supplied based on this approximate distribution. If we think about linear dynamical systems parameter estimates and the example form Section 1.3 we see that the distribution of the estimate is approximated with a Gaussian distribution. The quality of this approximation improves as $N \to \infty$. Moreover, the full distribution assumption
about the noise can be relaxed, for example this approximation remains valid asymptotically even if we only assume some finite moment conditions on the noise such as Lindeberg’s condition [Billingsley, 1995].

We consider those situations highly non-informative that don’t fit the characterization given in the previous paragraph.

A possibility to get problems of this kind is when the noise distribution does not have finite second moments. It such cases the Gaussian approximation is not going to become valid, not even asymptotically as $N \to \infty$. Chapter 2 describes a set of methods for creating confidence regions around the estimates that do not require any moment conditions, thus these methods are applicable to highly non-informative circumstances as well.

Another example that we consider as highly non-informative is when the $N \to \infty$ asymptotic would make the Gaussian approximation valid but the number of samples at hand is not high enough to ensure that using the asymptotic results is reliable enough. This was illustrated on the example in Section 1.3 and Chapter 4 is going to be focused on problems of this sort.
Data Perturbation Methods

This chapter presents a generic family of hypothesis testing procedures that can be used to characterize exact confidence sets of parameter estimates $\hat{\theta}_N$ in a wide class of estimation problems. The term *exact* refers to the fact that the confidence level of the tests is exactly the required one, irrespective of the sample count $N$, or the complexity of the estimation problem (linear or nonlinear).

The described method draws its inspiration from the SPS method introduced in a series of papers by the group of authors [Campi et al., 2012]. The main contributions presented here with respect to previous publications on SPS are that

- a more modular description is given here, that allows a deeper understanding of the method by separating assumptions about the noise distribution from other parts of the algorithms.

- all previous publications concentrate on the case when the joint distribution of the noise is symmetric. This condition is largely relaxed.

- the notion of performance measures is introduced and elaborated.

- the proofs of some results related to linear regression problems are included that were not published so far, offering new insights.

- a more natural solution, compared to previous publications, is given to the problem of estimating dynamical systems parameters in this framework.
Before we start the detailed description of data perturbation methods, we give a short overview on distribution free statistics in Section 2.1. To convince the reader that it is worth the effort to read the rest of the chapter we included an example in Section 2.2 that is worked out in great detail and shows many important features of the proposed methodology. This is followed by the description of the data perturbation framework for generic estimation problems, irrespective of model structures, in Section 2.3. Results related to linear regression problems are given in Section 2.4, while the case of linear dynamical systems is treated in Section 2.5. The problems solved in this chapter cover only a fraction of the identification workflow. To give a more complete picture, we provide notes related to the not covered parts of the workflow in Chapter 2.6. Concluding remarks are given in Section 2.7.

2.1 Introduction to Distribution Free Statistics

The most elementary example of estimation is when the mean of a Gaussian population is estimated. Let \( x_i, i \in [N] = 1, \ldots, N \) be a set of numbers drawn independently from a Gaussian distribution \( \mathcal{N}(\mu_0, \sigma^2) \), with known variance \( \sigma^2 \). The estimate is calculated as the sample mean \( \hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i \), and it is known that the distribution of the estimate is \( \mathcal{N}(\mu, \sigma^2/N) \). Based on the distribution of \( \hat{\mu} \), the tightest confidence region can be calculated that has a prescribed confidence level. By making very detailed assumptions, it is possible to exactly calculate the distribution of the estimate. This is what makes it possible to make decisions between the hypothesis \( H_0 : \mu_0 = \hat{\mu} \) and \( H_1 : \mu_0 \neq \hat{\mu} \) with exact confidence levels.

It was recognized early in the statistics community that having such strong assumptions on the samples (exactly knowing their distribution) is not acceptable in many cases. Although the confidence regions obtained at the end are rigorous if the assumptions are fulfilled, we do not know how much the actual data violates these assumptions (and it most certainly violates them to some extent). This makes the reliability of the results questionable. This served as motivation for a large group of statisticians to deliver hypothesis tests that do not rely on distribution assumptions. This branch of statistics, called nonparametric or distribution free statistics, had its golden era around in the first half of the twentieth century.
Some of the early works on the topic are [Pitman, 1937a] and [Pitman, 1937b], from which permutation tests can be originated. A large family of rank tests can be traced back to the works of [Mann and Whitney, 1947] and [Wilcoxon, 1945], which are reviewed in [Kendall and Gibbons, 1990]. Other methods of the same kind are jackknifing originating from [Quenouille, 1949] and [Quenouille, 1956], bootstrapping [Efron, 1979] and cross validation methods overviewed in [Hastie et al., 2009]. A general and thorough overview of nonparametric statistical methods can be found in [Hollander and Wolfe, 1999].

We are going to focus our attention on statistical hypothesis tests under relaxed distribution assumptions, which implies that we are considering probabilistic problem formulations. Before continuing in that direction we would like to mention that this is not the only possible approach. The framework of unknown but bounded errors is such an alternative in which case some hard constraints are assumed about the error variables (i.e. \( \ell_\infty \) norm, power constraint, average power in a specific time window, etc.) [Cerone, 1993]. This is also an active line of research [Meslem et al., 2014], but it operates under fundamentally different assumptions. Beside the many dissimilarities, however, there are also similarities with the approach presented in this chapter. These similarities concern the computational complexity of the methods and they are going to be commented in Section 2.5.5.

In order to make hypothesis testing possible on a given confidence level, a random variable (a statistic) is needed with known distribution. Almost all methods, that can be used to create such a statistic without fully detailed assumptions about the distribution of the underlying dataset, rely on some symmetry (either in the dataset or the algorithm) that can still be exploited to end up with something with known distribution. These symmetries are best captured by group theory. The role of group representations in statistics is nicely presented in [Diaconis, 1988].

The reason why most of these methods had a difficulty penetrating the system identification community lies in the fact that these are mostly designed to handle datasets where the order of samples is arbitrary. The “tricks” needed to obtain the statistics with known distributions often involve rearranging samples or leaving out samples. These are problematic in case of dependent datasets, where the samples are linked through some specific dynamics and these “tricks” would make the dataset inconsistent with that. Nonetheless, there are results in the dependent
setup as well. Some of the earlier examples for moving average and autoregressive models are [Bose, 1988] and [Bose, 1990].

Interest in exact identification methods was renewed with the publication of [Campi and Weyer, 2005], introducing the Leave-out Sign-dominant Correlation Regions (LSCR) method. This method was successfully applied to many problems in system identification covering nonlinear systems [Dalai et al., 2007], the problem of unmodelled dynamics [Campi et al., 2009], non parametric frequency response function estimates [Ko et al., 2007], also prediction and filtering [Weyer and Campi, 2011]. This method overcomes the issue caused by the dependency between samples by carefully selecting samples that are used in calculating correlations. If this is done correctly (the selected samples are determined by an underlying group) and evaluated for all possible combinations (in the sense of the underlying group) then exact confidence regions can be constructed. The single biggest drawback of the LSCR method is its computational complexity. In order to deliver its confidence regions, calculations need to be carried out over all elements of a group whose size is growing linearly with the sample count. What we get in return for the computational burden (beside the exact confidence level) is that the delivered result is determined by the dataset. Calculating a confidence region twice on the same dataset will result in the same confidence region.

A different method from the same group of authors is the method of Sign-perturbed Sums (SPS), which was developed in a series of papers [Csáji et al., 2012a, Csáji et al., 2012b, Weyer et al., 2013, Csáji et al., 2014, Csáji et al., 2015]. This method is also a hypothesis test with exact confidence level but with a much smaller computational complexity than LSCR. The price we pay for the reduced computational complexity is that the delivered results are no longer determined by the data alone. The algorithm itself contains random elements, which will make the outcome random.

The work presented in the remainder of this chapter relies heavily on the SPS method. It is described in a more detailed manner and is significantly extended in terms of it assumptions. It needs to be mentioned, that although the methodology extends very well to a more general setup compared to how the SPS was introduced, there are results that require some specific variant of the SPS, hence these results
might not carry over to the more general setting. Some of these will be explicitly mentioned in the sequel. Part of the presented material was published in [1].

2.2 Example

The goal of this section is to serve as appetizer for the rest of the chapter. We are going to work out an example that is still simple enough to be presented in detail but illustrates the flexibility of data perturbation methods. Some details that are not relevant to discussing this example are swept under the rug, but hopefully the example is convincing enough to motivate reading the following sections for the rigorous description and proofs of the claimed properties.

2.2.1 Problem Data

This section contains every detail of the example problem setup. This includes the description of the model structure and the noise model along with the actual sample.

Model structure: We chose a simple nonlinear model with a single scalar parameter to illustrate that the presented method can be applied to a wide range of models. The nominal model is

\[ y_0[k] = \arctan(\theta_0 x[k]) \quad k \in [N] \]  

with \( \theta_0 = 1 \).

Noise model: The sample data is generated using the additive noise model, so the actual measurements \( y[k] \) are generated as

\[ y[k] = \arctan(\theta_0 x[k]) + e[k] \quad k \in [N] \]  

where \( e[k] \) is assumed to be a sequence of independent and identically distributed random variables with zero mean. No additional knowledge about the noise is required to make the statements in this example rigorous, so we don’t even discuss how the actual realization is generated.
Table 2.1: Sample data

<table>
<thead>
<tr>
<th>$k$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x[k]$</td>
<td>1.2557</td>
<td>-1.5826</td>
<td>-2.2836</td>
<td>0.6438</td>
<td>-0.2992</td>
</tr>
<tr>
<td>$y_0[k] = \tan(x[k])$</td>
<td>0.8983</td>
<td>-1.0073</td>
<td>-1.1580</td>
<td>0.5720</td>
<td>-0.2907</td>
</tr>
<tr>
<td>$e[k]$</td>
<td>-0.0308</td>
<td>0.0023</td>
<td>-0.0365</td>
<td>-0.3915</td>
<td>0.1333</td>
</tr>
<tr>
<td>$y[k] = y_0[k] + e[k]$</td>
<td>0.8675</td>
<td>-1.0050</td>
<td>-1.1946</td>
<td>0.1805</td>
<td>-0.1574</td>
</tr>
</tbody>
</table>

Black box estimator: Throughout this example we are going to assume that an estimator algorithm is available that can provide an estimate $\hat{\theta}$ to every set of input-output measurements $x[k], y[k]$. We are going to use the least squares estimator that delivers the estimate

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}} \frac{1}{2N} \sum_{k=1}^{N} (y[k] - \tan(\theta x[k]))^2$$

We specify the estimator to enable the reader to reproduce the example numerically, but any other function that maps measurements to parameters could be used.

Sample data: The sample data is given in Table 2.1. The least squares estimate belonging to this data set is $\theta_{LS} = 0.8040$.

The visual representation of the nominal model, the sample points and the least squares estimated model is given in Figure 2.1.
2.2.2 Membership Tests

As it is going to be described later, data perturbation methods offer a hypothesis test for the null hypothesis $H_0 : \theta_0 = \theta$ against the alternative hypothesis $H_1 : \theta_0 \neq \theta$. This section executes a proposed algorithm step-by-step that corresponds to a $2/3$ confidence test.

**Data perturbation setup:** Beside the random noise in the measurements, the outcome of a DP hypothesis test will also depend on some extra random variables that are generated when the test is performed. The data perturbation setup $\Pi$ is composed of these random objects. The noise model assumptions and the chosen confidence level determine that for this example $\Pi$ consists of two uniformly chosen random permutations $\tilde{G}_{\Pi}^1, \tilde{G}_{\Pi}^2$ and the identity permutation $\tilde{G}_{\Pi}^3$

$$
\tilde{G}_{\Pi}^1 = \langle 1, 2, 3, 4, 5 \rangle \quad \tilde{G}_{\Pi}^2 = \langle 4, 1, 2, 5, 3 \rangle \quad \tilde{G}_{\Pi}^3 = \langle 4, 1, 5, 3, 2 \rangle
$$

(2.4)

Upper case letters will refer to the column vector form of the given signal, such as $E = (e[k])_{k=1}^n$. With a slight abuse of notation we are going to write that $\tilde{G}_{\Pi} E$ is the rearranged version of the vector $E$ as

$$
\left( \tilde{G}_{\Pi} E \right)(k) = e \left[ \tilde{G}_{\Pi}(k) \right] \quad k \in [N]
$$

(2.5)

2.2.2.1 Membership Test of $\theta_0$

This section describes the steps to perform the hypothesis test of $\theta = \theta_0$ on the given data $X$ and $Y$ with respect to the chosen data perturbation setup $\Pi$.

**Perturbed data sets:** The first step is to generate three perturbed data sets based on the given data and the data perturbation setup. In order to do this, we calculate an estimated noise sequence as

$$
E(\theta) = Y - \text{atan}(\theta X)
$$

(2.6)

Since we are testing $\theta = \theta_0$, $E(\theta) = E$. For every permutation in $\Pi$ we generate perturbed noise sequences using this estimated noise sequence, these are $\tilde{G}_{\Pi} E$, $\tilde{G}_{\Pi}^2 E$ and $\tilde{G}_{\Pi}^3 E$. With these perturbed noises, we can compute the perturbed data
Table 2.2: Perturbed data sets corresponding to $\theta = \theta_0$

<table>
<thead>
<tr>
<th>$k$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_0[k] = \atan(x[k])$</td>
<td>0.8983</td>
<td>-1.0073</td>
<td>-1.1580</td>
<td>0.5720</td>
<td>-0.2907</td>
</tr>
<tr>
<td>$e[k]$</td>
<td>-0.0308</td>
<td>0.0023</td>
<td>-0.0365</td>
<td>-0.3915</td>
<td>0.1333</td>
</tr>
<tr>
<td>$\tilde{G}_2^H E$</td>
<td>-0.3915</td>
<td>-0.0308</td>
<td>0.0023</td>
<td>0.1333</td>
<td>-0.0365</td>
</tr>
<tr>
<td>$y^{(2)}[k] = y_0[k] + \tilde{G}_2^H E[k]$</td>
<td>0.5068</td>
<td>-1.0380</td>
<td>-1.1558</td>
<td>0.7053</td>
<td>-0.3272</td>
</tr>
<tr>
<td>$\tilde{G}_3^H E$</td>
<td>-0.3915</td>
<td>-0.0308</td>
<td>0.1333</td>
<td>-0.0365</td>
<td>0.0023</td>
</tr>
<tr>
<td>$y^{(3)}[k] = y_0[k] + \tilde{G}_3^H E[k]$</td>
<td>0.5068</td>
<td>-1.0380</td>
<td>-1.0248</td>
<td>0.5355</td>
<td>-0.2884</td>
</tr>
</tbody>
</table>

set as $Y^{(i)} = \atan(\theta X) + \tilde{G}_i^H E(\theta)$. Since $\theta = \theta_0$, we have that $Y^{(1)} = Y$. The perturbed noise sequences and data sets for $i = 2$ and $i = 3$ are given in Table 2.2.

**Performance of $\theta_0$:** In order to be able to decide whether to accept $\theta$ or not, we need to measure the performance of this model on the three available perturbed data sets. To do this, we are going to calculate the least squares estimates $\hat{\theta}^{(i)}$ for each of these data sets. Because $\theta = \theta_0$, we know that $\hat{\theta}^{(1)} = \hat{\theta}_{LS} = 0.8040$. The other two estimates are

$$\hat{\theta}^{(2)} = 0.8760 \quad \hat{\theta}^{(3)} = 0.7490$$  \hspace{1cm} (2.7)

The performance of the tested model $\theta$ on the $i$-th data set is defined as

$$Z_i = \frac{1}{N} \sum_{k=1}^{N} \left( \atan(\theta x[k]) - \atan(\hat{\theta}^{(i)} x[k]) \right)^2$$  \hspace{1cm} (2.8)

and it can be interpreted as the mean square difference between the predictions of the tested model and the perturbed estimate. The performances of $\theta_0$ on the above given perturbed data sets are

$$Z_1 = 0.0083 \quad Z_2 = 0.0030 \quad Z_3 = 0.0146$$  \hspace{1cm} (2.9)

**Concluding the membership:** In order to have well structured (in case of this example this means connected and bounded) confidence regions we accept $\theta$ if $Z_1$ is not the largest value among $Z_i$, thus $\theta = \theta_0$ is accepted in this case.
Table 2.3: Perturbed data sets corresponding to $\theta = \theta_{LS}$

<table>
<thead>
<tr>
<th>$k$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y^0_{\theta_{LS}}[k] = \text{atan}(\theta_{LS}x[k])$</td>
<td>0.7902</td>
<td>-0.9047</td>
<td>-1.0721</td>
<td>0.4777</td>
<td>-0.2361</td>
</tr>
<tr>
<td>$e(\theta_{LS})[k]$</td>
<td>0.0774</td>
<td>-0.1003</td>
<td>-0.1225</td>
<td>-0.2971</td>
<td>0.0786</td>
</tr>
<tr>
<td>$\bar{G}^1 E(\theta_{LS})$</td>
<td>0.0774</td>
<td>-0.1003</td>
<td>-0.1225</td>
<td>-0.2971</td>
<td>0.0786</td>
</tr>
<tr>
<td>$y^{(1)}[k] = y^0_{\theta_{LS}}[k] + \bar{G}^1 E(\theta_{LS})[k]$</td>
<td>0.8675</td>
<td>-1.0050</td>
<td>-1.1946</td>
<td>0.1805</td>
<td>-0.1574</td>
</tr>
<tr>
<td>$\bar{G}^2 E(\theta_{LS})$</td>
<td>-0.2971</td>
<td>0.0774</td>
<td>-0.1003</td>
<td>0.0786</td>
<td>-0.1225</td>
</tr>
<tr>
<td>$y^{(2)}[k] = y^0_{\theta_{LS}}[k] + \bar{G}^2 E(\theta_{LS})[k]$</td>
<td>0.4930</td>
<td>-0.8274</td>
<td>-1.1724</td>
<td>0.5563</td>
<td>-0.3585</td>
</tr>
<tr>
<td>$\bar{G}^3 E(\theta_{LS})$</td>
<td>-0.2971</td>
<td>0.0774</td>
<td>0.0786</td>
<td>-0.1225</td>
<td>-0.1003</td>
</tr>
<tr>
<td>$y^{(3)}[k] = y^0_{\theta_{LS}}[k] + \bar{G}^3 E(\theta_{LS})[k]$</td>
<td>0.4930</td>
<td>-0.8274</td>
<td>-0.9934</td>
<td>0.3552</td>
<td>-0.3363</td>
</tr>
</tbody>
</table>

2.2.2.2 Membership Test of $\theta_{LS}$

This section describes the steps to perform the hypothesis test of $\theta = \theta_{LS} = 0.8040$ on the given data $X$ and $Y$ with respect to the chosen data perturbation setup $\Pi$. Perturbed data sets: Since $\theta = \theta_{LS} \neq \theta_0$, we now have that $E(\theta) \neq E$. Nevertheless, the same calculations can be performed as previously. The perturbed data sets are given in Table 2.3.

Performance of $\theta_{LS}$: The least squares estimates belonging to these perturbed data sets are

$$\hat{\theta}^{(1)} = 0.8040 \quad \hat{\theta}^{(2)} = 0.7290 \quad \hat{\theta}^{(3)} = 0.6060$$  \hspace{1cm} (2.10)

The performances of $\theta = \theta_{LS}$ can be calculated again using (2.8) which gives

$$Z_1 = 0 \quad Z_2 = 0.0017 \quad Z_3 = 0.0139$$  \hspace{1cm} (2.11)

Concluding the membership: Notice, that $Z_1 = 0$ and this result is independent of the actual data set. Since the performance was defined as a strictly positive function, this shows that $\theta_{LS}$ will always be accepted, as 0 can never be the greatest of 3 different non-negative values.
Table 2.4: Perturbed data sets corresponding to $\theta \to \infty$

<table>
<thead>
<tr>
<th>$k$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_0^\infty[k]$ = $\text{atan}(\infty x[k])$</td>
<td>$\pi/2$</td>
<td>$-\pi/2$</td>
<td>$-\pi/2$</td>
<td>$\pi/2$</td>
<td>$-\pi/2$</td>
</tr>
<tr>
<td>$e(\infty)[k]$</td>
<td>-0.7033</td>
<td>0.5658</td>
<td>0.3762</td>
<td>-1.3902</td>
<td>1.4133</td>
</tr>
<tr>
<td>$\tilde{G}_1^\Pi E(\infty)$</td>
<td>-0.7033</td>
<td>0.5658</td>
<td>0.3762</td>
<td>-1.3902</td>
<td>1.4133</td>
</tr>
<tr>
<td>$y^{(1)}[k] = y_0^\infty[k] + \tilde{G}_1^\Pi E(\infty)[k]$</td>
<td>0.8675</td>
<td>-1.0050</td>
<td>-1.1946</td>
<td>0.1805</td>
<td>-0.1574</td>
</tr>
<tr>
<td>$\tilde{G}_2^\Pi E(\infty)$</td>
<td>-1.3902</td>
<td>-0.7033</td>
<td>0.5658</td>
<td>1.4133</td>
<td>0.3762</td>
</tr>
<tr>
<td>$y^{(2)}[k] = y_0^\infty[k] + \tilde{G}_2^\Pi E(\infty)[k]$</td>
<td>0.1805</td>
<td>-2.2741</td>
<td>-1.0050</td>
<td>2.9841</td>
<td>-1.1945</td>
</tr>
<tr>
<td>$\tilde{G}_3^\Pi E(\infty)$</td>
<td>-1.3902</td>
<td>-0.7033</td>
<td>1.4133</td>
<td>0.3762</td>
<td>0.5658</td>
</tr>
<tr>
<td>$y^{(3)}[k] = y_0^\infty[k] + \tilde{G}_3^\Pi E(\infty)[k]$</td>
<td>0.1805</td>
<td>-2.2741</td>
<td>-0.1574</td>
<td>1.9470</td>
<td>-1.0050</td>
</tr>
</tbody>
</table>

2.2.2.3 Membership Test of Faraway Models

Since the test is defined for individual models, we are restricted to be able to test only a bounded number of models. It is not hard to see that the $Z_i$ values are continuous functions of the tested model, so evaluating a fine enough grid within a bounded set will give us a reliable picture about which models are accepted in that set. But it is also important to see what happens outside. Since

$$\lim_{\theta \to \infty} \text{atan}(\theta x[k]) = \text{sign}(x[k]) \frac{\pi}{2}$$

we can test the membership of models as $\theta \to \infty$.

**Perturbed data sets:** The perturbed data sets are given in Table 2.4.

**Performance of $\theta \to \infty$:** The least squares estimates belonging to these perturbed data sets are

$$\hat{\theta}^{(1)} = 0.8040 \quad \hat{\theta}^{(2)} = \infty \quad \hat{\theta}^{(3)} = 4.9870$$

and the corresponding performances are

$$Z_1 = 0.8548 \quad Z_2 = 0.0280 \quad Z_3 = 0.0978$$
Concluding the membership: Since \( Z_1 \) is the greatest, we have that there is an upper limit \( \theta_U \), such that every \( \theta > \theta_U \) is going to be rejected by the membership test. Similar calculations can show the existence of a lower limit \( \theta_L \). When combined, these results show that the confidence region is bounded, and it is a subset of \( [\theta_L, \theta_U] \).

2.2.3 The Confidence Region

As we have tested \( \theta = \theta_0 \) and \( \theta = \theta_{LS} \) we could test any arbitrary parameter \( \theta \in \mathbb{R} \) for membership with respect to the given data \( (X,Y) \) and the selected data perturbation setup \( \Pi \). Let \( C_{2/3}^\Pi \subset \mathbb{R} \) be the set of points that pass the membership test. That is, \( C_{2/3}^\Pi \) is the confidence set belonging to the given data \( (X,Y) \) and the data perturbation setup \( \Pi \). This confidence region has the prescribed \( 2/3 \) confidence level and it is shown in Figure 2.2.

We would like to note that it can be proven analytically that no values outside \([-0.5, 1.3]\) will pass the membership test, thus Figure 2.2 shows the entire confidence region. This is due the monotonicity of the model in \( x \) and the chosen performance measure.

2.2.4 Conclusions

The goal of this section was to give a glance at the complexity of the problems that can be handled with data perturbation methods. The most notable features are that

1. mild assumptions are required about the noise model.

2. the model structure can be arbitrarily complex.

![Figure 2.2: Example confidence set \( C_{2/3}^\Pi \)](image)
3. the resulting confidence region $C^\Pi$ is still a random subset of the parameter space, even if the measurement data is fixed. This is due to the random selection of the data perturbation setup $\Pi$.

The remainder of this chapter gives the precise framework for constructing similar hypothesis testing algorithms for a large variety of model structures and noise structure assumptions. The focus will be put on linear regression and linear dynamical systems model structures and some practically relevant noise structure assumptions will be discussed explicitly.

### 2.3 Definition of Data Perturbation Methods

The goal of this section is to describe a generic method to create exact confidence level hypothesis tests for parameter estimates. Before doing that, the problem setting needs to be formally defined. We are going to specify a model class through a probabilistic model description and we will assume that the measurements can be described by a model from this model class.

**Definition 2.1 (Model structure).** A model structure is a set of parameterized mappings $\mathcal{M} = \{\mathcal{M}(\theta) | \mathcal{M}(\theta) : \mathcal{X} \times \mathcal{E} \to \mathcal{Y}\}$ where $\mathcal{X}$ is the set of possible measured inputs, $\mathcal{E}$ is the set of possible noise disturbances, $\mathcal{Y}$ is the set of possible outputs and the parameter vector $\theta$ is an element of the set of possible parameters $\Theta \subset \mathbb{R}^{n_\theta}$.

**Definition 2.2 (Model structure defining function).** The behavior of a specific model $\mathcal{M}(\theta) \in \mathcal{M}$, for a concrete set of signal values $X \in \mathcal{X}, E \in \mathcal{E}$ and $Y \in \mathcal{Y}$, is given by the relation $Y = f(\theta, X, E)$, where $f$ is assumed to be a known function $f : \Theta \times \mathcal{X} \times \mathcal{E} \to \mathcal{Y}$. We say that the model structure $\mathcal{M}$ is defined by the function $f$.

The distinction between the input values $X$ and noise values $E$ is made explicit because we are going to consider probabilistic models where $E$ is considered as a random variable with some unknown distribution having specific symmetry properties. The definition of the model class becomes complete only with the description of these properties. All results, reported in this chapter, will contain this
additional piece of information, thus completing the definition of the model class in an appropriate way.

**Assumption 2.3** (Model structure assumption). *We assume that the measurements come from a model $M(\theta_0)$ parameterized by the nominal parameter vector $\theta_0$ and they are described in the following form*

$$Y = f(\theta_0, X, E)$$  \hspace{1cm} (2.15)

*where $X$ contains known data not contaminated by noise, $E$ contains the randomness contaminating the measurements, $Y$ contains the observables, $f : \Theta \times \mathcal{X} \times \mathcal{E} \rightarrow \mathcal{Y}$ is a known mapping from the model parameters, inputs and noises to the observables.*

Assumption 2.3 means that we know perfectly how noise contaminates our measurements. We are going to describe a family of methods with the goal to create a hypothesis test for the null hypothesis $H_0 : \theta_0 = \theta$ against the alternative hypothesis $H_1 : \theta_0 \neq \theta$.

When these hypothesis tests are defined, it will be implicitly assumed that there is a known estimation algorithm that can provide estimates $\hat{\theta}$ for the nominal parameter $\theta_0$ based on the pair $X$ and $Y$. Most estimation algorithms require that the function $f$, which defines the model structure, is differentiable in its first argument (see Definition 4.3 in [Ljung, 1999]). This assumption is fulfilled for all model structures parameterizations considered in this chapter.

**Definition 2.4** (Invertibility with respect to noise). A model structure $\mathcal{M}$ defined by the function $f : \Theta \times \mathcal{X} \times \mathcal{E} \rightarrow \mathcal{Y}$ is invertible with respect to noise if there exists a mapping $f^* : \Theta \times \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{E}$ such that

$$\forall \theta \in \Theta, \forall X \in \mathcal{X}, \forall Y \in \mathcal{Y} : (\exists! E \in \mathcal{E} : Y = f(\theta, X, E)) \Rightarrow E = f^*(\theta, X, Y)$$

where $\exists!$ denotes unique existence.

**Assumption 2.5** (Invertibility with respect to noise). *We assume that the nominal model $\theta_0$ belongs to a model structure which is invertible with respect to noise in the sense of Definition 2.4.*
This invertibility assumption together with the model structure assumption means that for every set of data $D = (Y, X)$ we can explicitly and uniquely calculate the contaminating noise $E$ for every possible model parameter $\theta$, if the data can be generated with that model.

**Definition 2.6** (Transformation invariance). Let $\mathbb{P} : E \rightarrow [0, 1]$ be the joint distribution of a random vector. Let $\mathcal{G}$ be a set of transformations such that, together with their natural composition operation, these transformations form a compact group $G = (\mathcal{G}, \cdot)$. The distribution is invariant with respect to $G$ if $\forall E \subset \mathcal{E}$, $\forall g \in \mathcal{G}$ $\mathbb{P}(E) = \mathbb{P}(gE)$.

**Definition 2.7** (Composite set). A set $\mathcal{E}$ is called a composite set if there exist sets $\mathcal{X}_1$ and $\mathcal{X}_2$ and bijective function $r : \mathcal{X}_1 \times \mathcal{X}_2 \rightarrow \mathcal{E}$ with coordinatewise inverse functions $r_1^{-1} : \mathcal{E} \rightarrow \mathcal{X}_1$ and $r_2^{-1} : \mathcal{E} \rightarrow \mathcal{X}_2$ such that

$$\forall e \in \mathcal{E} : e = r(r_1^{-1}(e), r_2^{-1}(e))$$

The composite assumption is the formal way of saying that $\mathcal{E}$ can also be thought of as a two dimensional set $\mathcal{X}_1 \times \mathcal{X}_2$, there exists a one-to-one mapping between points of these sets.

As we will see later, data perturbation methods can be constructed for hypothesis testing of the parameter vector $\theta$ whenever Assumptions 2.3 and 2.5 are fulfilled by the model structure that we estimate and the joint distribution of the contaminating noise adheres to Definitions 2.6 and 2.7.

The remainder of this section is organized as follows. Section 2.3.1 contains lemmas that are going to be needed to prove the main theorem in Section 2.3.2. As an example, we show in Section 2.3.3 that the SPS method is a DP method belonging to the symmetric noise distributions. We show other examples of transformation invariant noises in Section 2.3.4. The notion of performance measures is going to be introduced later in this section; we give general comments about them in Section 2.3.5.
Chapter 2. Data Perturbation Methods

2.3.1 Probability Prerequisites

Exchangeable random variables play a central role in the theory of data perturbation methods. An excellent account of their properties is given in [Aldous, 1985]. We are going to prove some properties of such random variables in this section, together with consequences of transformation invariance.

First we are going to show that if a random variable $E$ takes its values in a set $E$, that is composite with respect to $G$ and an arbitrary set $X$, and its joint distribution is invariant under the transformations with respect to a compact group $(G, \cdot)$, then $E$ can be realized as a function of two independent random variables over $X$ and $G$ respectively.

**Lemma 2.8** (Independent realization of composite random variables). Let $E$ be an $(E, \sigma_E)$ valued random variable defined over the probability space $(\Omega, \sigma_\Omega, P)$. Assuming that

i. the set $E$ is a composite with respect to $G$ and $X$

ii. the probability measure $P$ is invariant under $(G, \cdot)$

iii. $(G, \cdot)$ is a compact group

$E$ can also be defined over the probability space $(\Omega_G \times \Omega_X, \sigma_{\Omega_G} \times \sigma_{\Omega_X}, U_G \times P_X)$ where $U_G$ denotes the uniform distribution over $G$ and $P_X$ is some unknown measure.

**Proof.** Let $v = (X_1, X_2)$ be a two dimensional random variable, defined as $(r_1^{-1}(E), r_2^{-1}(E))$. Because of condition $i$, this is a bijective transformation with the inverse $E = r(X_1, X_2)$. This makes the calculation of marginal and conditional distributions of $X_1$ and $X_2$ possible. We will show that $X_1$ and $X_2$ are independent; this will prove that $E$ can be generated using the transformation $r$ and independent realizations of $X_1$ and $X_2$. Let $P_1$ and $P_2$ denote the marginal distributions of $X_1$ and $X_2$ respectively.

The marginal distribution of $X_1$ can be calculated as

\[
P_1(X_1 \in A) = P(r_1^{-1}(E) \in A) = P(r_1^{-1}(E) \in gA) = P(r_1^{-1}(E) \in Ag)
\]

where $g \in G$ and the group invariance of $P$ was used. The unique (up to a constant) translation invariant finite signed measure over a compact group $(G, \cdot)$ exists and it
is the Haar measure (see Theorem 6.8 in [Knapp, 2005]). This proves that uniform
distribution over \( G, U_G \), exists and it is also the marginal distribution of \( X_1 \). The
same argument shows that the conditional distributions of \( X_1 \) are also uniform.

\[
P(r_1^{-1}(E) \in A| r_2^{-1}(E) \in B) = \\
P(r_1^{-1}(E) \in gA| r_2^{-1}(E) \in B) = \\
P(r_1^{-1}(E) \in Ag| r_2^{-1}(E) \in B) = U_G(A)
\]

Using these, the independence of \( X_1 \) and \( X_2 \) follows by writing their joint distri-
bution in product form as

\[
P(X_1 \in A, X_2 \in B) = P(r_1^{-1}(E) \in A, r_2^{-1}(E) \in B) = \\
P(r_1^{-1}(E) \in A| r_2^{-1}(E) \in B)P(r_2^{-1}(E) \in B) = \\
U_G(A)P(r_2^{-1}(E) \in B) = U_G(A)P_2(B) = P_1(A)P_2(B)
\]

This proves that if \( P_X = P_2 \) is chosen, then \( E \) is realized as transformation of two
independent random variables with \( U_G \) and \( P_X \).

The next ingredient that we will need is the randomization property of groups.
The meaning of this property is that by multiplying a uniformly chosen random
element \( g_1 \) of the group with another one \( g_2 \), the product \( g_1g_2 \) is also going to be
uniformly distributed, moreover \( g_1 \) and \( g_1g_2 \) are independent of each other. We
are going to use this property for collections of variables, so the following lemma
proves this property in the multidimensional setting.

**Lemma 2.9** (Randomization property). Let \((G, \cdot)\) be a compact group and let \( U_G \)
denote the uniform distribution over \( G \). Let \( X_1, \ldots, \tilde{X}_m \) be \( U_G \) distributed
i.i.d. random variables, and let \( X_k = X_1 \tilde{X}_k \) for \( k = 2, \ldots, m \). The sequence of
variables \((X_1, \ldots, X_m)\) is also an i.i.d. sequence of \( U_G \) distributed random vari-
ables.

**Proof.** As \( U_G \) is invariant under the group, so is the joint distribution of \( m - 1 \)
independent variables from this distribution, \( U_G^{m-1} \). What needs to be shown is
that the product \( X_1 \tilde{X}_k \) does not compromise the independence between \( X_1 \) and
The joint distribution of these variables can be written as
\[ P(X_1 \in A, (X_2, \ldots, X_m) \in B) = P((X_2, \ldots, X_m) \in B|X_1 \in A)P(X_1 \in A) \]

Using the property of groups that \( X_1^{-1} \) exists and the definition \( X_k = X_1 \tilde{X}_k \) we get
\[ P(X_1^{-1}(X_2, \ldots, X_m) \in X_1^{-1}B|X_1 \in A)P(X_1 \in A) = P((\tilde{X}_2, \ldots, \tilde{X}_m) \in X_1^{-1}B|X_1 \in A)P(X_1 \in A) \]

By invariance of \( U_{m-1}^G \) and the independence between \( X_1 \) and \( \tilde{X}_k \)
\[ P((\tilde{X}_2, \ldots, \tilde{X}_m) \in B|X_1 \in A)P(X_1 \in A) = P((\tilde{X}_2, \ldots, \tilde{X}_m) \in B)P(X_1 \in A) = U_{m-1}^G(B)U^G(A) \]

This shows that the joint distribution of \( (X_1, X_2, \ldots, X_m) \) is the product of \( U^G \) and \( U_{m-1}^G \). The marginal distribution of \( X_1 \) is obviously \( U^G \), so what is left to complete the proof is to show that the marginal distribution of \( (X_2, \ldots, X_m) \) is \( U_{m-1}^G \).

\[ P((X_2, \ldots, X_m) \in B) = \int_{g \in G} P((\tilde{X}_2, \ldots, \tilde{X}_m) \in g^{-1}B)dU(g) = \]
\[ \int_{g \in G} P((\tilde{X}_2, \ldots, \tilde{X}_m) \in B)dU(g) = \]
\[ P((\tilde{X}_2, \ldots, \tilde{X}_m) \in B)\int_{g \in G} 1dU(g) = U_{m-1}^G(B) \]

where, again, the invariance of the distribution is used.

**Definition 2.10** (Exchangeable random variables). A finite sequence \( (X_1, \ldots, X_N) \) of random variables is called exchangeable if
\[ (X_1, \ldots, X_N) \overset{d}{=} (X_{\pi(1)}, \ldots, X_{\pi(N)}) \tag{2.17} \]
for all permutations \( \pi \) over \([N]\), where \( \overset{d}{=} \) denotes equality in distribution.

By definition all i.i.d. sequences are exchangeable. The goal of exchangeability is to generalize the notion of i.i.d. random variables to conditionally independent
random variables, or putting it differently, mixtures of i.i.d. sequences. One way of creating such mixtures of i.i.d. sequences is to consider a family of distributions \( P_\theta \), parameterized with a set of parameters \( \theta \), draw this parameter from any distribution, then draw \( X_i \) independently from distribution \( P_\theta \) (see Chapter 1 of [Aldous, 1985]).

**Lemma 2.11** (Mixture of i.i.d. sequences is exchangeable). Let \( G_1, \ldots, G_m \) be an i.i.d. sequence of random variables with distribution \( P_G \) and let \( X \) be a random variable with distribution \( P_X \). If \( r : G \times X \rightarrow E \) is a deterministic function and \( E_i = r(G_i, X), i \in [m] \) is defined, then the sequence \( (E_1, \ldots, E_m) \) is an exchangeable sequence.

**Proof.** We can think about \( X \) as the parameter of the distribution, and the statement follows from Definition 2.10. \[ \square \]

**Remark 2.12.** If a function \( Z : E \rightarrow \mathbb{R} \) is used to map the random variables \( E_i \) from Lemma 2.11 to real numbers as \( Z_i = Z(E_i) \), then \( (Z_1, \ldots, Z_m) \) is going to be an exchangeable sequence of scalars. This follows by using the function \( r' = Z \circ r \).

**Definition 2.13** (Well defined increasing ordering with respect to a permutation). The well defined increasing ordering of values \( Z_1, \ldots, Z_m \) with respect to a permutation \( \pi \) over \([m]\) is another permutation \( O \) over \([m]\) defined such that \( i \) precedes \( j \) in \( O \) if 1. \( Z_i < Z_j \) or if 2. \( Z_i = Z_j \) and \( i \) precedes \( j \) in \( \pi \).

The well defined increasing ordering of real values is the usual ordering if there are no equal values to be ordered. If there are ties, then these ties are uniquely solved by the given permutation \( \pi \).

One property of i.i.d. sequences of scalar random variables, that is generalized by exchangeable sequences, is that their well defined ordering with respect to a uniformly chosen permutation is uniform over all possible orderings. This is formalized in Lemma 2.14.

**Lemma 2.14** (Uniform ordering of exchangeable random variables). Let \( (Z_1, \ldots, Z_m) \) be a sequence of exchangeable scalar random variables on \( \mathbb{R} \) and let \( \pi \) be a uniformly chosen permutation over \([m]\). If \( O_\pi(Z_1, \ldots, Z_m) \) is the well defined
ordering of the values $Z_i$ with respect to $\pi$, as defined in Definition 2.13, then for every permutation $\tilde{\pi}$
\[ \mathbb{P}(O_{\tilde{\pi}}(Z_1, \ldots, Z_m) = \tilde{\pi}) = \frac{1}{m!} \]

Proof. What characterizes a uniform random permutation $\pi$ is that for any pair of values $(i, j) \in [m]^2$
\[ \mathbb{P}(\pi_i < \pi_j) = \frac{1}{2} \quad (2.18) \]
This is the property of $O_{\pi}(Z_1, \ldots, Z_m)$ that we are going to verify. To help notation,
let $I_i$ be defined as the position of $Z_i$ in $O_{\pi}(Z_1, \ldots, Z_m)$, so what we need to show is that $\mathbb{P}(I_i < I_j) = \frac{1}{2}$ for all $(i, j)$ pairs.
\[ \mathbb{P}(I_i < I_j) = \mathbb{P}(I_i < I_j|Z_i \neq Z_j)\mathbb{P}(Z_i \neq Z_j) + \mathbb{P}(I_i < I_j|Z_i = Z_j)\mathbb{P}(Z_i = Z_j) \quad (2.19) \]
Because of the exchangeability of $Z_i$-s, $\mathbb{P}(I_i < I_j|Z_i \neq Z_j) = \mathbb{P}(I_j < I_i|Z_j \neq Z_i)$.
Using that $\mathbb{P}(I_i < I_j|Z_i \neq Z_j) + \mathbb{P}(I_j < I_i|Z_j \neq Z_i) = 1$ we get that $\mathbb{P}(I_i < I_j|Z_i \neq Z_j) = \frac{1}{2}$. In the case $Z_i = Z_j$ the order $I_i < I_j$ is defined by $\pi$, which is uniformly distributed, so $\mathbb{P}(I_i < I_j|Z_i = Z_j) = \frac{1}{2}$. Substituting these into (2.19), we get
\[ \mathbb{P}(I_i < I_j) = \frac{1}{2} (\mathbb{P}(Z_i \neq Z_j) + \mathbb{P}(Z_i = Z_j)) = \frac{1}{2} \]

2.3.2 The General Framework

As promised at the beginning of the section, we are going to present an algorithm that can be used for hypothesis testing for a parameter vector $\theta$, based on a set of measurements coming from a system fulfilling Assumptions 2.3 and 2.5.
\[ Y = f(\theta_0, X, E) \quad (2.20) \]
where $\theta_0$ is the nominal parameter, $X$ is considered as the measured input of the system, $E$ plays the role of the contaminating noise, and $Y$ is the observation. The method is going to be applicable when the noise distribution is invariant under a compact group $(G, \cdot)$ (see Definition 2.6) and it takes values from a set that is
a composite of $\mathcal{G}$ and some other set $\mathcal{X}$ (see Definition 2.7). We need two more definitions before giving the algorithm.

**Definition 2.15** (Performance measure). A performance measure is a function $Z : \theta \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, that maps a parameter vector $\theta$ and a set of measurements $(X,Y)$ to a scalar value.

**Definition 2.16** (Data perturbation setup). A data perturbation setup $\Pi$ of order $m$ over a compact group $(\mathcal{G}, \cdot)$ consists of $m-1$ uniformly and independently chosen elements $G_{\Pi}^1, \ldots, G_{\Pi}^m$ of $\mathcal{G}$, and a uniformly chosen permutation $\pi^\Pi$ of the numbers $[m]$. To help notation $G_{\Pi}^1$ is defined as the identity element of the group $(\mathcal{G}, \cdot)$.

In order to create a hypothesis test with confidence level $p$, this value should be written in the form $p = \frac{k}{m!}$, with $k$ and $m$ being integers. A subset $\mathcal{A}$ of all possible permutations of $[m]$ should be selected such that $|\mathcal{A}| = k$ and a data perturbation setup $\Pi$ should be generated at random. If the model structure and the noise distribution adhere to the assumption mentioned earlier, then Algorithm 2.1 will accept $\theta_0$ with probability $p$. This is formalized in Theorem 2.17.

**Algorithm 2.1** The data perturbation method

```
1: procedure TestModel($\theta, X, Y, \mathcal{A}, \Pi$)
2: if $f^*(\theta, X, Y)$ is not defined then
3:   Reject $\theta$.
4: else
5:   Calculate $E(\theta) = f^*(\theta, X, Y)$.
6:   Create $m$ perturbed datasets as $Y^{(i)} = f(\theta, X, \tilde{G}_{i}^{\Pi}E(\theta)) \quad \forall i \in [m]$.
7:   Define $Z_i$ as the performance of the model $\theta$ on the different datasets $Z_i = Z(\theta, X, Y^{(i)}) \forall i \in [m]$.
8:   Calculate the well defined increasing ordering $O$ of the values $Z_1, \ldots, Z_m$ with respect to $\pi^\Pi$.
9:   if $O \in \mathcal{A}$ then
10:      Accept $\theta$.
11:   else
12:      Reject $\theta$.
13: end if
14: end if
15: end procedure
```
Theorem 2.17. Let the dataset \( D = (Y, X) \) be generated by a model that satisfies Assumptions 2.3 and 2.5 with underlying model parameter \( \theta_0 \), so \( Y = f(\theta_0, X, E) \) and \( E_\theta = f^*(\theta, X, Y) \) is defined such that \( E = E_{\theta_0} = f^*(\theta_0, X, Y) \). Let the distribution of the noise \( E \) be invariant under a compact group \((G, \cdot)\), in the sense of Definition 2.6 and its domain \( \mathcal{E} \) be a composite of \( G \) and some other set \( \mathcal{H} \), in the sense of Definition 2.7. Let \( Z \) be any performance measure, \( A \) be an arbitrary subset of all permutations over \([m]\) and let \( \Pi \) be a data perturbation setup of order \( m \) over \( G \) selected uniformly at random. Under these conditions

\[
\mathbb{P}(\text{TestModel}(\theta_0, X, Y, A, \Pi) = \text{Accept}) = \alpha(A) = \frac{|A|}{m!}
\]

where \( \text{TestModel} \) is defined in Algorithm 2.1.

As this theorem is a generalized version of Theorem 1. in [Csáji et al., 2012a] we feel the need to point out three differences compared to that before proving it. First, [Csáji et al., 2012a] proves the particular case of this theorem for symmetrically distributed noise distributions, which is a condition that is significantly relaxed here. This is going to be elaborated in Section 2.3.4. Also, it is defined for linear regression and linear dynamical systems parameter estimates, but here it is extended to a large class of nonlinear estimation problems. Second, two specific performance measures are used in [Csáji et al., 2012a]. It is important to emphasize that the specific choice of the performance measure has no impact on the confidence of the test, it is only important to shape the confidence regions belonging to these tests. This will be discussed in more detail in Section 2.4 and 2.5, where specific model structures will be examined and the selected performance measures will reflect these model structures. Third, the tie resolution based on the random permutation \( \pi^\Pi \) makes implementing the algorithm easier compared to the tie resolution given in [Csáji et al., 2012a]. Beside these differences the proof of Theorem 2.17 follows the same train of thoughts as that in [Csáji et al., 2012a].

Proof. Based on the invertibility assumption the exact noise realization can be extracted from the data if \( \theta_0 \) is tested, indeed

\[
E_{\theta_0} = f^*(\theta_0, X, Y) = E
\] (2.21)
The noise random variable $E$ takes its values from a composite set $\mathcal{E}$, that is a composite of $\mathcal{G}$ and $\mathcal{H}$, moreover $\mathcal{G}$ is a set over which a compact group $(\mathcal{G}, \cdot)$ can be defined. Based on Lemma 2.8 we know that $E$ can be generated using a bijective mapping $r : \mathcal{G} \times \mathcal{H} \to \mathcal{E}$ of two independent random variables. Let $G_1 \sim \mathcal{U}_\mathcal{G}$ be a uniformly distributed random variable over $\mathcal{G}$ and $H$ be the other $\mathcal{P}_\mathcal{H}$ distributed random variable, such that $E = r(G_1, H)$.

Let us define $G_i = G_1 G_i^{\Pi}$ for all $i \in [m]$ using the data perturbation setup $\Pi$. Mind that $G_1^{\Pi} = 1$, so the definition is not conflicting on $G_1$. As a consequence of the randomization property in Lemma 2.9, we know that $(G_1, \ldots, G_m)$ is a sequence of i.i.d. random variables with law $\mathcal{U}_\mathcal{G}$.

By defining $E_i = r(G_i, H)$ and using Lemma 2.11 we know that the variables $E_i$ are exchangeable and so are $Y^{(i)}$. In accordance with Remark 2.12 also $Z_i$ are exchangeable.

As a consequence of Lemma 2.14, for every permutation $\pi$ of the numbers $[m]$ we have that
\[
\mathbb{P}(O_{\pi n}(Z_1, \ldots, Z_m) = \pi) = \frac{1}{m!}
\]

Since the cardinality of $\mathcal{A}$ is $k$, the total probability acceptance is $\frac{k}{m!}$, no matter how $\mathcal{A}$ is assembled.

\begin{remark}
As it can be seen from the proof, the intermediate step of defining the perturbed outputs $Y^{(i)}$ is not really necessary in Algorithm 2.1. If the performance measures would defined as $\theta \times \mathcal{X} \times \mathcal{E} \to \mathbb{R}$ instead of $\theta \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, it would not make a difference (as $f^*$ is a bijection between $\theta \times \mathcal{X} \times \mathcal{E}$ and $\theta \times \mathcal{X} \times \mathcal{Y}$). The reason for defining the performance measure and the algorithm this way will be justified in later sections, where particular performance measures will be interpreted more naturally in this setting.
\end{remark}

\begin{remark}
A nice feature of Lemma 2.8, and the proof of Theorem 2.17, is that it is exactly the distribution $\mathbb{P}_{\mathcal{H}}$ that can be arbitrary in the problem setup. Every other element is assumed in one way or another.

Theorem 2.17 relies on the fact that, under the null hypothesis, we know the distribution of the ordering of values $Z_1, \ldots, Z_m$ exactly. It is important to note that without $Z_1$, the distribution of the ordering of $Z_2, \ldots, Z_m$ is known also under the alternative hypothesis. This is formalized in the following lemma.
Lemma 2.20 (Uniform ordering of $Z_2, \ldots, Z_m$). The random variables $Z_2, \ldots, Z_m$ in Algorithm 2.1 are exchangeable, thus they are uniformly ordered.

Proof. The simplest way to see this is to consider the values $Z_i$ as a deterministic function of $\theta_0, X, E$ and $G^\Pi_i$. If we think about $(\theta_0, X, E)$ as parameters of a distribution then the statement follows from Definition 2.10 (similarly to the proof of Lemma 2.11). \hfill \Box

2.3.3 SPS as a DP Method

As the SPS method was the first DP method, and a main source of inspiration, it is natural that we showcase the framework with it. The SPS method was defined for linear regression and linear dynamical systems parameter estimation problems with symmetric noise distribution.

It is clear that both linear regression and linear dynamical systems models fulfill Assumptions 2.3 and 2.5. Assumption 1 from [Csáji et al., 2012b] ("The true system that generates the data is in the model class") is the exact correspondent of Assumptions 2.3. While Assumption 2 from [Csáji et al., 2012b] ("The transfer function $H(z^{-1}; \theta^*)$ has a stable inverse. Moreover, $G(0; \theta^*) = 0$ and $H(0; \theta^*) = 1"), together with assumed initial conditions is the version of Assumption 2.5 corresponding to linear dynamical systems.

We will discuss transformation invariant noise distributions in more detail in Section 2.3.4, but we illustrate here how a symmetric real valued random variable fits the conditions of Theorem 2.17. If the variable is symmetric, then its values come from a set $[-K, K]$ ($K$ possibly infinite). This set is a composite of $G = \{-1, +1\}$ and $H = [0, K]$, with the mapping $r$ being the simple product of the values from these sets. The group $(G, \cdot)$ with the regular product is a finite group, which is compact by definition.

The performance measures of SPS are going to be discussed later when particular model structures are examined in Sections 2.4 and 2.5.

The acceptance condition in the SPS method is based on the position of $Z_1$ in the ordered sequence. This can be simply replaced with the set of permutations where the position of 1 fulfills the same condition. The particular choice of the acceptance set $A$ will also be discussed later.
2.3.4 Transformation Invariant Noise Sources

The conditions about noise distributions given in Theorem 2.17 are rather abstract, so the goal of this section is to provide three comprehensible examples of such distributions. All these examples are going to be distributions that are invariant under a subgroup of the orthogonal group, which is the group of orthogonal matrices.

**Definition 2.21 (Orthogonal group).** The orthogonal group of order $N$ is the group $O(N)$ defined over the set of matrices $G \in \mathbb{R}^{N \times N}$ such that $G^T G = GG^T = I_N$, with the group operation being the regular matrix multiplication.

It is easy to show that the orthogonal group, and its subgroups are compact, but we include the proof to justify their use in the rest of this section.

**Lemma 2.22 (Compactness of $O(N)$ and its subgroups).** The orthogonal group and its subgroups are compact.

**Proof.** Orthogonal matrices form a closed subspace in $\mathbb{R}^{N^2}$ as they can be considered as vectors in this space and the set is defined by the equality constraint $GG^T = I_N$. When subgroups are defined, the extra restrictions also come in form of equality constraints to specify the structure or the determinant of the matrix. As $GG^T = I_N$, we know that $\sum_{j=1}^{N} g_{i,j}^2 = 1$ for all $i \in [N]$. Thus, the entries $g_{i,j}$ of the matrix have absolute value less than one. So the matrices form a bounded subspace of $\mathbb{R}^{N^2}$. The Heine-Borel theorem (Theorem 3 in [Knopp, 1996]) states that a closed and bounded subset of the Euclidean space is compact. As restrictions with additional equality constraints preserve closedness, and the same boundedness argument applies, subgroups are also compact. \hfill \Box

In order to show that a noise distribution fulfills the requirements of Theorem 2.17, it needs to be shown that it is invariant under a group, and its value set is a composite involving the elements of that group.

2.3.4.1 Symmetric Noise Sequences

The first example concerns symmetric distributions over $\mathbb{R}^N$. For a vector $v \in \mathbb{R}^N$ let $\bar{v}^i$ denote the vector $\bar{v}^i = [v_1, \ldots, v_{i-1}, -v_i, v_{i+1}, \ldots, v_N]^T$, where the $i$-th
coordinate is flipped. Let $E$ be a random variable over $\mathbb{R}^N$ such that its distribution is symmetric in every coordinate as

$$\mathbb{P}(E \in A) = \mathbb{P}(E^\pi \in A) \quad \forall i \in [N], A \subset \mathbb{R}^N \quad (2.22)$$

Let $\mathcal{G}$ be the set of square matrices of size $N$ with nonzero entries only on the diagonal, with diagonal entries having only values $\pm 1$. With the regular matrix multiplication this is a compact group, as stated in Lemma 2.22.

Let $\mathbb{R}_+^N$ denote the subset of $\mathbb{R}^N$ with nonnegative entries, $|v|_c$ denote the coordinatewise absolute value vector and $\text{sign}_c(v)$ be the coordinatewise sign function. Clearly $\mathbb{R}^N$ is a composite of $(\mathcal{G}, \cdot)$ and $\mathcal{H} = \mathbb{R}_+^N$, as

$$\forall E \in \mathbb{R}^N \quad E = \text{diag}(\text{sign}_c(E)) |E|_c \quad (2.23)$$

We mention that the hidden distribution $\mathbb{P}_\mathcal{H}$ is the distribution of the variable $|E|_c$ is not needed. Nonetheless, in case of distributions that have a density function, calculation of the density of $|E|_c$ is not particularly difficult.

### 2.3.4.2 Exchangeable Noise Sequences

Let $\mathcal{G}$ denote the set of permutation matrices of size $N$. These are matrices with entries $\{0, 1\}$, with only one nonzero entry in every row and column. These matrices form a finite subgroup of $O(N)$, so it is a compact group. If $E \in \mathbb{R}^N$ is a random vector such that its entries are exchangeable then, by Definition 2.10

$$\mathbb{P}(E \in A) = \mathbb{P}(GE \in A) \quad \forall G \in \mathcal{G}, A \subset \mathbb{R}^N \quad (2.24)$$

Let $\mathbb{R}_+^N$ denote the subset of $\mathbb{R}^N$ with (not strictly) increasing entries, $o(v) \in \mathbb{R}^N$ denote the ordered values of vector $v$ and $o_\pi(v) \in \mathcal{G}$ be the matrix representation of the permutation that correspond to the actual order of entries in $v$ with respect to any fixed permutation $\pi$. Clearly $\mathbb{R}^N$ is a composite of $G$ and $\mathcal{H} = \mathbb{R}_+^N$, as

$$\forall E \in \mathbb{R}^N \quad E = o_\pi(E) o(E) \quad (2.25)$$
We mention that the hidden distribution \( P_H \) is the joint distribution of the order statistics of \( E \). Which is known to be notoriously difficult to calculate even for simple distributions. Since this is the very object about which no assumptions are made, it can offer a major simplification compared to classical hypothesis tests as the effect it has on the statistic itself does not need to be calculated.

### 2.3.4.3 Power Defined Noise Sequences

The last example of “good” distributions that we present is the distribution of noise vectors \( E \) that are uniformly distributed conditionally on their power. To keep the notation simple we illustrate the concept assuming that the distribution has a density function \( f \), but this is not a necessary restriction. A distribution is power defined if its density \( f \) depends on its argument only through the sum of squares of its entries, so there exists a function \( g \) such that \( f \) can be written in the form

\[
f(E) = g \left( \sum_{i=1}^{N} E_i^2 \right)
\]

Such distributions are usually called circularly symmetric, but we use the term power defined so we can reserve symmetry to always mean coordinatewise symmetry.

In case of such random variables \( \mathbb{R}^N \) can be written as a composite of \( \mathbb{R} \) and the set defining the special orthogonal group \( \text{SO}(N) \). \( \text{SO}(N) \) is the subgroup of \( \text{O}(N) \) restricted to real rotations in \( \mathbb{R}^N \), namely the group over those orthogonal matrices \( G \) that have \( \det(G) = 1 \) (so \(-1\) is not allowed). There is no closed form representation of these matrices for large values of \( N > 3 \), but it is known that they are parameterized by \( \frac{N(N-1)}{2} \) dimensional real vectors (see Proposition 3 in Chapter VI of [Curtis, 1979]). Be defining \( t(v) = \sum_{i=1}^{N} v_i^2 \) and \( G(v) \) as the matrix representation of the rotation that moves the point \( [1, 0, \ldots]^T \) to \( v/t(v) \) we have that

\[
\forall E \in \mathbb{R}^N \quad E = G(E) \begin{bmatrix} t(E) \\ \vdots \end{bmatrix}
\]
There are some interesting comments that should be made about this particular noise distribution. Firstly, when the elements $G_i$ in the data perturbation setup need to be selected, we should be able to draw elements uniformly from $\text{SO}(N)$. In the previous examples the uniform distribution could be implemented easily over the sets in question. In case of $\text{SO}(N)$ this is no longer trivial. The simplest known algorithm for generating such samples is to use QR decomposition of a matrix that has independent Gaussian entries, under the condition that the resulting $R$ matrix has only positive diagonal entries. This algorithm is very inefficient in high dimensions as many retries are needed before an $R$ matrix with proper diagonal is obtained. A more efficient algorithm is given in [Diaconis and Shahshahani, 1987].

The other significant difference compared to the previous examples is that $G$ is not a finite set. This has the consequence that the values $Z_i$ will be equal with probability zero (assuming that the performance measure is not degenerate, see later in Theorem 2.24), rendering the tie resolution using the uniform permutation $\pi$ unnecessary.

### 2.3.5 About Performance Measures in General

As we have discussed in detail noise distributions that enable the usage of data perturbation methods, we turn our attention to the other significant ingredient: performance measures. Theorem 2.17 describes the behavior of DP methods at $\theta_0$ for arbitrary performance measure $Z$, but it is also important how a hypothesis test behaves for models $\theta \neq \theta_0$, which can be quantified using the power of the test.

We will see in the rest of the chapter that the choice of performance measure will effect the shape of the generated confidence regions. The probability that the nominal parameter $\theta_0$ will be part of the confidence region is independent of the chosen performance measure. By choosing good performance measures we can ensure that relevant parts of the parameter space will be selected as the confidence region.

We will also see that the performance measure suggested to be optimal in terms of tightness will be computationally expensive for complex model structures. It can be stated in general that there is a trade-off between tightness of the generated
confidence regions and the computational complexity. On one hand there is the prediction based performance measure that is defined in Section 2.5.1, this measure is computationally expensive but it is thought to produce tight confidence regions. On the other hand choosing a constant performance measure \( Z(\theta, X, Y) = c \) results in a random coin flip which either marks the empty set or the entire parameter space as the confidence region. This approach is obviously computationally as cheap as it gets but the resulting confidence region is useless. This trade-off between computational complexity and tightness has to be always evaluated for the problem at hand.

Definition 2.23 (Power of a hypothesis test). When testing a null hypothesis \( H_0 : \theta = \theta_0 \) against the alternative \( H_1 : \theta \neq \theta_0 \), the power of the test \( \Phi \) is defined as

\[
\Phi(\theta) = P(H_0 \text{ rejected} | H_1 \text{ is true})
\]

Figure 2.3 illustrates different shapes of power functions with respect to \( \Delta \theta = \theta_0 - \theta \). On the left, Figure 2.3a shows the shape of a power function that most hypothesis tests have. Namely, as we are testing parameters further away from \( \theta_0 \) the probability of rejection increases, preferably converges to 1.

DP methods use lots of random variables in the data perturbation setup \( \Pi \) to come up with a yes or no answer, saying yes with a given probability \( p \). An obvious way of doing this would be to toss a coin that is heads with probability \( p \) and make a decision based on that. This method would have the desired confidence also, and its power function is shown in Figure 2.3b. To no surprise, making decisions solely based on coin tosses is not of too much use.
The choice of the performance measure $Z$ and the set of accepted permutations $A$ will define the shape of the confidence regions, and that of the power functions belonging to particular DP methods. This is important because the method, as it is now, is nothing more than a “fancy coin toss”, so special care should be devoted to finding the proper performance measures and accepted permutations to make the corresponding power function look more like Figure 2.3a instead of Figure 2.3b.

The next theorem identifies performance measures that should be avoided at all cost, as they are useless.

**Theorem 2.24** (Invariant performance measures degenerate to a coin toss). Let a DP method be defined for a problem with underlying noise distribution invariant under the compact group $(G, \cdot)$, using a performance measure $Z : \theta \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, with accepted permutation over $[m]$ in $A$. If $Z$ is also invariant under $G$ in the sense

$$Z(\theta, X, Y) = Z(\theta, X, f(\theta, X, G f^*(\theta, X, Y))) \quad \forall G \in G, \theta \in \theta, X \in \mathcal{X}$$

(2.29)

then the method is equivalent to a random coin toss with the probability of being heads $\alpha(A) = \frac{|A|}{m!}$.

**Proof.** The different values $Z_i$ are generated in Algorithm 2.1 as

$$Z_i = Z(\theta, X, f(\theta, X, G^\Pi f^*(\theta, X, Y)))$$

so if the performance measure is invariant in the sense (2.29), all these values will be equal. This will result in the ordering $O$ to be exactly equal to the tie resolution order $\pi^\Pi$ from the DP setup $\Pi$. Which means that this is a uniformly chosen permutation independent from everything, including the measurement data. Thus acceptance of a parameter vector $\theta$ does not depend on the measurement data at all, it just depends on a uniformly distributed permutation $\pi^\Pi$ falling into the set of accepted permutations $\mathcal{A}$. This is equivalent to a coin toss with heads probability $\frac{|A|}{m!}$. \qed
One such example, that matches all three noise distributions discussed earlier, is the usual least squares performance measure defined as

$$V_{X,Y,\theta}(\alpha) = \sum_{i=1}^{N} e_i^2 \quad (e_i)_{i=1}^{N} = E(\alpha) = f^*(\alpha, X, Y) \quad (2.30)$$

Perturbing the dataset in a way that the sum of squares of the prediction errors is preserved will change the cost function $V_{X,Y,\theta}(\cdot)$ to $V_{X,Y,\theta}'(\cdot)$, where $Y'$ is the perturbed output, but these two different LS cost functions will have the property that $V_{X,Y,\theta}(\theta) = V_{X,Y,\theta}'(\theta)$, which is exactly the condition that needs to be avoided. This shows that the LS cost function should not be used as a performance measure for symmetric, exchangeable or power defined noises. The following sections will discuss performance measures that are adequate for such problems in case of linear regression and Box-Jenkins model structures.

### 2.3.6 Evaluating the Risk of a Model

When a parameter estimate is defined as the minimizer of a cost function, assigning a risk value to any given parameter vector is quite straightforward, the value of the cost function is used. In case of the least squares cost function this associated risk is the estimate of the variance. Although this example of risk was given under a distribution assumption, the concept of risk can also be defined in the context of data perturbation methods. We would like to emphasize that DP methods can be constructed in cases where estimating the variance is not even meaningful. For example in case of coordinatewise symmetric noise the individual noise components need not be identically distributed, which renders variance estimation meaningless.

The risk that a hypothesis test can associate to a parameter can be defined as the minimum confidence level under which the hypothesis test would accept the parameter. This notion of risk is not the same, but it is closely related to the classical definition of risk which is usually given as the expected loss of a given loss function.

**Definition 2.25** (Risk). Let $C_\alpha$ be the confidence region generated by a given hypothesis test on a given dataset corresponding to confidence level $\alpha$. The risk of
a model $\theta$ is defined as

$$R(\theta) = \arg\min_{\alpha} \{ \theta \in C_{\alpha} \}$$

So far we have selected the parameters of the DP setups in a way to match a prescribed confidence level we wanted the associated test to have. We change this attitude a little bit in order to be able to find upper and lower bounds for $R(\theta)$. First we have to select the resolution we want to have for these bounds. If $\overline{R}(\theta)$ and $\underline{R}(\theta)$ denote the upper and lower bound of $R(\theta)$ then we select $m$ such that $\frac{1}{m} < \overline{R}(\theta) - \underline{R}(\theta)$ is an acceptable lower bound for the desired approximation.

**Definition 2.26 (Monotone DP method).** A DP method is called monotone if $A_{\alpha_1} \subset A_{\alpha_2}$ if $\alpha_1 \leq \alpha_2$, where $A_{\alpha_1}$ and $A_{\alpha_2}$ are the sets of accepted orderings for confidence levels $\alpha_1$ and $\alpha_2$ respectively.

Lets assume that a DP method is monotone in the sense of Definition 2.26. We will see in later sections that all DP methods proposed in this dissertation satisfy the requirements of this definition.

The risk $R(\theta)$ can be approximated by randomly selecting a data perturbation setup $\Pi$ and defining the bounds as

$$\overline{R}(\theta) = \arg\max_k \{ \theta \text{ is accepted with respect to the set } A_{k/m} \}$$

$$\underline{R}(\theta) = \arg\min_k \{ \theta \text{ is rejected with respect to the set } A_{k/m} \}$$

When the aforementioned monotonicity of the accepted orderings is provided, it allows drawing level set contours, and the set for a given confidence level can be chosen as one of these nested sets. Moreover this notion of upper and lower bounds for the risk can be extended from parameter points to sets of parameters, defining an overall upper and lower bound for parameters within the given set. This opens up the possibility to discover confidence regions using interval analytic methods. This is going to be elaborated later in Section 2.5.5.

### 2.4 DP Methods for Linear Regression Problems

The goal of this section is to find performance measures for parameter estimation of linear regression problems with well behaved corresponding power functions.
When the SPS method was introduced in [Csáji et al., 2012a], two possible performance measures were suggested. We will examine the implications of using these performance measures. It will be shown in Section 2.4.1, that one of the performance measures used in [Csáji et al., 2012a] is, in some sense, the natural performance measure to be used. The examination of the other suggested performance measure is carried out in Section 2.4.2, where we will see that it cannot be suggested in good faith. The price that we pay for different assumptions is going to be illustrated in Section 2.4.3.

We will assume for the rest of the section, that the measurements come from a model structure as

$$ Y = X^T \theta + E $$

where $Y, E \in \mathbb{R}^N$, $\theta \in \mathbb{R}^{n_{\theta}}$ and $X \in \mathbb{R}^{n_{\theta} \times N}$ such that the random noise $E$ is independent from the regressor matrix $X$. It is evident that this model fulfills Assumptions 2.3 and 2.5.

### 2.4.1 Performance Measure for Noise Distributions Invariant Under Orthogonal Transformation

The performance measure that is used in [Csáji et al., 2012a] can be written as

$$ Z(\theta, X, Y) = (Y - X^T \theta)^T X^T [XX^T]^{-1} X (Y - X^T \theta) $$

and the main result in this section is that this performance measure works well for linear regression problems as long as the noise distribution fulfills the requirements of Theorem 2.17 with a group $(\mathcal{G}, \cdot)$ that is a subgroup of the orthogonal group $O(N)$, such as the ones presented in Section 2.3.4. We define a condition for sufficient excitation and elaborate on it before stating this result precisely.

**Definition 2.27** (Sufficiently exciting input). Let $(\mathcal{G}, \cdot)$ denote a subgroup of $O(N)$, and $G$ be the matrix representation of a group element. We say that the problem input $X$ is sufficiently exciting with respect to $G$ if

$$ Q_G(X) = XX^T - XG^T X^T [XX^T]^{-1} XGX^T > 0 $$

(2.34)
holds, where $> 0$ is understood as the matrix being positive definite.

We would like to emphasize that this sufficient excitation condition can be much more restrictive than the usual one, requiring the existence of $[XX^T]^{-1}$. The easiest way to illustrate this is to consider the one dimensional problem with $n_\theta = 1$ and exchangeable noise distribution. The only not exciting input in the usual sense would be the constant zero input. However the condition in (2.34) will fail for every constant input $x_i \equiv c$ as $Q_G(X) = 0$ for all permutations in this case. Similar examples can be created also for the other distributions discussed in Section 2.3.4.

**Theorem 2.28.** If the joint distribution of the noise vector $E$ matches the requirements of Theorem 2.17 with a subgroup $G$ of $O(N)$ then data perturbation methods can be constructed for the parameter vector $\theta_0$ of a linear regression problem that will result in connected confidence regions containing the least squares estimate. These confidence regions will also be bounded if the input is sufficiently exciting with respect to every perturbation matrix $G_{\Pi}^k$ for $k \geq 2$ in the used DP setup $\Pi$ and the required confidence level is less than $1 - \frac{1}{m}$.

The performance measure for these tests is

$$Z(\theta, X, Y) = (Y - X^T \theta)^T X^T [XX^T]^{-1} X (Y - X^T \theta) \quad (2.35)$$

and the accepted set of orderings $\mathcal{A}$ is selected such that permutations are included in $\mathcal{A}$ in increasing order with respect to the position of 1 in them until $\alpha(\mathcal{A})$ reaches the desired confidence level.

**Proof.** The fact that DP methods can be defined is given as all conditions of Theorem 2.17 are required explicitly.

We are going to prove the statements of the theorem for confidence regions belonging to DP methods with confidence level $\frac{1}{2}$. Sets for other confidence levels are created as a union of intersections of such sets, preserving the stated structural properties. For ease of notation $G_{1}^{\Pi}$ is going to be denoted as $G_1$ throughout the proof.

As the confidence is required to be less than $1 - \frac{1}{m}$, we know that the set of acceptable permutations $\mathcal{A}$ accepts no ordering with 1 in its last position. Thus in
order to show boundedness we need to prove that $Z_1$ will outgrow $Z_i$ as $\|\theta_0 - \theta\| \to \infty$ and we do this by showing that the difference $Z_1 - Z_i \to \infty$.

Using the definition of $Z_i = Z(\theta, X, Y^{(i)})$ and that $Y^{(i)} = X^T \theta + G_i(Y - X^T \theta)$ we get that $Y^{(i)} - X^T \theta = G_i(Y - X^T \theta)$, so the values $Z_i$ can be expressed directly using the problem data as

$$Z_i = (Y - X^T \theta)^T G_i^T X^T [XX^T]^{-1} X G_i (Y - X^T \theta) \quad (2.36)$$

Rewriting this expression in terms of $\theta_0 - \theta$ we get that

$$Z_i = (Y - X^T \theta_0)^T G_i^T X^T [XX^T]^{-1} X G_i (Y - X^T \theta_0) + 2 (Y - X^T \theta_0)^T G_i^T X^T [XX^T]^{-1} X G_i X^T (\theta_0 - \theta) + (\theta_0 - \theta)^T X G_i^T X^T [XX^T]^{-1} X G_i X^T (\theta_0 - \theta) \quad (2.37)$$

Since $Z_i$ is a quadratic function of $\theta_0 - \theta$, the limit $Z_1 - Z_i \to \infty$ as $\|\theta_0 - \theta\| \to \infty$ holds if $Q_i > 0$, where $Q_i$ is defined as

$$Q_i = XX^T - X G_i^T X^T [XX^T]^{-1} X G_i X^T \quad i \geq 2 \quad (2.38)$$

Using the fact that $G_i$ is an element from a subgroup of the orthogonal group $O(N)$, by definition $G_i G_i^T = I_N$. This allows $Q_i$ to be rewritten as

$$Q_i = X (I_N - G_i^T X^T [X G_i G_i^T X^T]^{-1} X G_i) X^T \quad (2.39)$$

The term in the middle is the difference of the identity matrix and a projection matrix defined by $G_i^T X^T$ showing that the eigenvalues of the middle term are either 0 or 1. This shows that $Q_i \geq 0$. If the input is sufficiently exciting with respect to all $G_i$ for $i \geq 2$ then $Q_i > 0$ also holds. This ensures that if the distance $\theta_0 - \theta$ is large enough, irrespective of its direction, $\theta$ will be excluded from the confidence set. This proves boundedness of the confidence sets under the given conditions.

As the level $1/2$ confidence set is characterized by a lower level set of a convex quadratic function, it is always connected and convex. Connectivity is preserved by union of intersections of such sets, this shows connectivity.
Chapter 2. Data Perturbation Methods

Showing that the least squares estimate is always in the confidence region is done by evaluating (2.36) with $G_1 = I_N$ and $\theta = \hat{\theta}^{LS} = [XX^T]^{-1}XY$.

\[
(Y - XX^TXX^T^{-1}XY)^T XX^TXX^T^{-1}X(Y - XX^TXX^T^{-1}XY) =
\]

\[
(XY - XX^TXX^T^{-1}XY)^T XX^T^{-1}(XY - XX^TXX^T^{-1}XY) =
\]

\[
(XY - XY)^T XX^T^{-1}(XY - XY) = 0
\]

Since $Z$ is a quadratic function, $Z_i \geq 0$ for all $i \in [m]$, thus $Z_1 = 0$ is always smaller or equal than the other $Z_i$ values, making sure that it is in the beginning of the ordering. This ensures that the least squares estimate $\hat{\theta}^{LS}$ is accepted.

Before moving on, some comments are in order about Theorem 2.28. Mind that the ordering $O$ in Algorithm 2.1 was created as an increasing ordering of the values $Z_i$. The theorem defines $A$ such that it includes permutations that have 1 at their first position. Once all these are added to $A$ and $\alpha(A)$ is still lower then the desired confidence level, then permutations containing 1 in the second position can be added, and so on. [Csáji et al., 2012a] defined the set $A$ such that all permutations that contains 1 at a given position $k$ are either included in $A$ or not, depending only on $k$, the “rank” of these permutations. This makes the definition simpler, but looses a little bit on the resolution of what can be achieved in terms of confidence levels.

The main reason why the theorem is formulated like this is because a connected and bounded confidence region is considered “good” in general, and we want to emphasize that this combination of the chosen performance measure and accepted orderings is the closest that we can get to this, covering a wide set of problems. However, even methods fulfilling the requirements of Theorem 2.28 can in some cases produce unbounded confidence regions. There are two possible ways how a confidence region can become unbounded. One of these is if the required confidence is approximately one. This is not unexpected, as every hypothesis test that relies on noise distributions with unbounded support will have to accept growing confidence regions as the confidence level approaches one. A slight difference with respect to usual hypothesis tests, is that the confidence level of confidence sets belonging to DP methods can only increase in discrete steps of $1/m!$. If the required confidence level is at least $1 - 1/m$ then the resulting confidence regions will be unbounded.
irrespective of any other problem specific property. $A$ will allow $Z_1$ to appear in any position of the ordering and the ordering of the other $Z_i$ values is an exchangeable sequence even if the null hypothesis fails (see Lemma 2.20). This problem can always be handled by choosing a sufficiently high value of $m$.

The other situation that will result in unbounded confidence regions is when $Z_i = Z_1$ occurs for some $i \in [m]$ (the data perturbation setup $\Pi$ contains some group elements $G^\Pi_k = I_N$). This will result in a partial degeneration to a coin toss, similarly to the case shown in Theorem 2.24. The probability of such a DP setup is small, if $G$ is large enough. If the group of perturbations is of continuum size then this is a zero probability event. If the group is finite then the probability of such a DP setup decreases to zero quickly in $N$. For symmetric noises the probability is $\sim (1 - 2^N)^\kappa$, while for exchangeable noises it is $\sim (1 - N!)^\kappa$, where $\kappa$ depends on the required confidence level and the particular choice of $m$.

In order to illustrate how DP methods works compared to the classical theory, we show confidence regions belonging to the same dataset created using different DP setups. Figure 2.4 shows the border of four DP confidence regions and the confidence region assuming Gaussian noise (that was used to create the dataset).
2.4.2 An Instructive Example for a Performance Measure

Theorem 2.28 provides performance measure (2.35) and the set of accepted orderings for a wide range of noise distributions that works well in case of linear regression problems. The question arises if there are other choices for these building blocks that could also be used with similar consequences. The goal of this section is to illustrate that this is by no means an easy task. We will replace the performance measure (2.35) with a different one, also suggested in [Csáji et al., 2012a], and we will show that this performance measure has worse properties than (2.35). Only the symmetric noise case will be considered formally, to help giving examples. We will comment on the general behavior of this alternative performance measure at the end of the section.

As we are dealing with symmetric noise distributions we know from Section 2.3.4.1 that the perturbation matrices $G_i$ in the DP setup are diagonal matrices with ±1 entries on the diagonal, with equal probability. The modified performance measure will require a different condition for sufficient excitation, other than Definition 2.27.

**Definition 2.29** (Symmetrically exciting input). If $G$ is a diagonal $N$ dimensional matrix with diagonal entries ±1 and $X \in \mathbb{R}^{n_b \times N}$ is the input matrix then let $X_+$ denote the submatrix of $X$ formed by the columns at positions in $X$ where $G$ contains +1 on the diagonal. Let $X_-$ be defined similarly as the complement of $X_+$. The input $X$ is called symmetrically exciting with respect to $G$ if

$$Q_G(X) = X_+X_+^TX_-X_-^T + X_-X_-^TX_+X_+^T > 0$$

(2.40)

Before stating the main result of this section we would like to elaborate on this definition. The matrix $Q_G(X)$ is positive definite exactly if $X_+X_+^TX_-X_-^T$ is positive definite in the sense that $v^T X_+X_+^TX_-X_-^Tv > 0 \forall v \in \mathbb{R}^{n_b}$. Compared to the regular assumption for linear regression inputs that $[XX^T]^{-1}$ exists, it is not that much stronger to also require the existence of $[X_+X_+^T]^{-1}$ and $[X_-X_-^T]^{-1}$. But we would like to underline the fact that this is not enough to fulfill the requirements of Definition 2.29. The exact condition for the product of the two PD matrices $X_+X_+^T$ and $X_-X_-^T$ to also be PD is that the product should be a normal matrix,
Chapter 2. Data Perturbation Methods

i.e. it should commute with its transpose [Meenakshi and Rajian, 1999]. There is a large set of inputs that fail to meet this criterion.

**Example 2.30.** Consider the case when

\[
X^+ X^T = \begin{bmatrix} 5 & 2 \\ 2 & 1 \end{bmatrix} \quad X^- X^T = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}
\]

Both matrices are positive definite, however their product has \([2 \ 1][-1 \ 1]^T = -1\) in the bottom right corner. This means that for \(v = [0 \ 1]^T\) the quadratic form will be negative. Indeed \(Q_G(X) = \text{diag}(16, -2)\), thus the condition is not satisfied.

Since this example was given in terms of the products \(X^+ X^T\) and \(X^- X^T\) there is a continuum number of ways how this product can be obtained by selecting \(X^+\) and \(X^-\) appropriately. Just as we have seen when elaborating on Definition 2.27, we see here as well that the conditions on the input for DP methods is significantly stronger than usually.

**Theorem 2.31.** *If the joint distribution of the noise vector \(E\) is symmetric then a data perturbation method can be constructed for the parameter vector \(\theta_0\) of a linear regression problem that will result in connected confidence regions containing the least squares estimate. These confidence regions will also be bounded if the input is symmetrically exciting with respect to any perturbation matrix \(G_k^{\Pi}\) for \(k \geq 2\) in the used DP setup \(\Pi\) and the confidence level is less than \(1 - 1/m\).*

The performance measure for this test is

\[
Z(\theta, X, Y) = (Y - X^T \theta)^T X^T X (Y - X^T \theta) \tag{2.41}
\]

and the accepted set of orderings \(\mathcal{A}\) is selected such that permutations are included in \(\mathcal{A}\) in increasing order with respect to the position of 1 in them until \(\alpha(\mathcal{A})\) reaches the desired confidence level.

**Proof.** With the exception of the boundedness of the confidence regions every statement in the theorem is proved exactly as in the proof of Theorem 2.28, so we do not repeat them here.
The only thing that is left to prove is that the limit $Z_1 - Z_i \to \infty$ as $\|\theta_0 - \theta\| \to \infty$ holds for this performance measure as well. Similarly to (2.36) and (2.37) $Z_i$ can be written as

$$Z_i = (Y - X^T \theta)^T G_i^T X^T X G_i (Y - X^T \theta)$$

$$= (Y - X^T \theta_0)^T G_i^T X^T X G_i (Y - X^T \theta_0) +$$

$$+ 2 (Y - X^T \theta_0)^T G_i^T X^T X G_i X^T (\theta_0 - \theta) +$$

$$+ (\theta_0 - \theta)^T X G_i^T X^T X G_i X^T (\theta_0 - \theta)$$

Without loss of generality it can be assumed that the diagonal of $G_i$ contains all +1-s grouped in the first columns and all −1 values are grouped to the last columns. Using the corresponding decomposition of $X$ to $X_+$ and $X_-$, the coefficient of the quadratic term and $XX^T XX^T$ can be written as

$$X G_i^T X^T X G_i X^T =$$

$$X_+ X_+^T X_+ X_+^T - X_+ X_+^T X_- X_-^T - X_- X_-^T X_+ X_+^T + X_- X_-^T X_- X_-^T$$

$$XX^T XX^T = X_+ X_+^T X_+ X_+^T + X_+ X_+^T X_- X_-^T + X_- X_-^T X_+ X_+^T + X_- X_-^T X_- X_-^T$$

This shows that the coefficient of the quadratic term in $Z_1 - Z_i$ depends on

$$Q_i = 2 \left( X_+ X_+^T X_- X_-^T + X_- X_-^T X_+ X_+^T \right) \quad i \geq 2 \quad (2.42)$$

The sufficient excitation condition (2.40) is the exact condition needed to ensure that $Z_1 - Z_i \to \infty$ as $\|\theta_0 - \theta\| \to \infty$.

If we were to generalize Theorem 2.31 from symmetric noises to noises invariant under a subgroup of O(N), the matrix in the middle of the quadratic form (2.39) would become $X^T X - G_i^T X^T X G_i$. The positive definiteness of this matrix is what determines if a confidence region is bounded or not. Since the product with matrix $G_i$ is nothing more than a rotation of the eigenspaces, this difference can become indefinite easily, as illustrated in Example 2.30. This is a major setback with respect to the performance measure shown in Theorem 2.17, as in that case the worst case scenario was that (2.39) might become semidefinite, but never indefinite. It was already mentioned in [Csáji et al., 2012a] that the weighting with $[XX^T]^{-1}$
helps to obtain better shaped confidence region. The previous argument should server as additional argument for using that weighting.

Beside providing the proof of the properties of a different performance measure, this result should also serve as justification for thinking about the performance measure (2.35) as the “natural” one.

As the only difference between the methods defined in Theorem 2.28 and Theorem 2.31 is the performance measure, the same DP setup can be used in both cases. This justifies the comparison of confidence regions belonging to these methods on the same dataset using the same DP setups. The difference is visible only in at least two dimensional problems, because in one dimension the two performance measures will result in the same orderings as they are a constant multiple of each other, depending on the regressors \( \mathbf{X} \). Figure 2.5 shows confidence regions constructed using the DP methods corresponding to Theorems 2.28 and 2.31, where we refer to the confidence region belonging to Theorem 2.28 as “proper”, while those belonging to Theorem 2.31 are referred to as “unweighted”. This figure depicts the borders of 0.75 confidence regions constructed using 25 samples and
Gaussian noise. The inside of the confidence regions is the part of the plane that is on the same side of the curves as the least squares estimate $\hat{\theta}_{\text{LS}}$. This can be considered as a visual proof for why we consider performance measure (2.35) to be superior, as in three out of four cases the unweighted performance measure results in unbounded confidence regions.

### 2.4.3 The Price of Assumptions

Before moving on from linear regression models to linear dynamical systems, we would like to highlight some properties of DP methods on a simple test case example, to further help grasp the essence of these methods. We are going to consider the one dimensional linear regression problem with white Gaussian noise to show how different assumptions about the noise distribution affect the hypothesis tests. We would like to emphasize four items on the bill that we have to pay for moving away from the strict assumptions of the classical theory.

The first item on this list is the condition given in Definition 2.27 showing that these methods require richer input compared to the case when the noise distribution is fully assumed.

Second, choosing the right performance measure for a given model structure is by no means evident. Since the confidence level of the tests is not influenced by the choice of the performance measure, the fitness of a performance measure needs to be evaluated based on the corresponding confidence regions or power functions, which is difficult for more complex model structures.

The third important characteristic of DP methods is that we trade in assumptions about the noise distribution for uncertainty in the outcome of the hypothesis test, even if it is repeated on the same dataset. This uncertainty is contained in the DP setup generated for the test. We illustrated this trade-off between assumptions and uncertainty in the confidence regions already in Figure 2.4.

We want to illustrate how the power function of these tests look like. To this end, we estimate the power function of different hypothesis tests on a sample linear regression problem with i.i.d. Gaussian noise. This noise distribution falls into all three categories that we discussed in Section 2.3.4, so the power functions belonging to different assumptions about the noise can be compared. The shape of
the Gaussian power function depends only on $XX^T$, but the shapes of the power functions belonging to different assumptions about the noise also depend on the particular values of the input. They are shown in Figure 2.6, where the power functions correspond to inputs with significantly different shape, but normalized in $XX^T$. As expected, tests not relying on the exact knowledge of the distribution are less powerful than the Gaussian test.

What is more interesting is that the order of these methods is not defined by the amount of information assumed.

The amount of the assumed information can be interpreted as a quantity proportional to the size of the set of invariant transforms $\mathcal{G}$. In case of power defined distributions this is a continuum set, for exchangeable noises its size is $N!$ while for symmetric noises it is $2^N$. Suppose that there is a group of transformations $(\mathcal{G}_0, \cdot)$ for which the noise distribution is invariant. If we assume full knowledge about the noise, we can rely on $(\mathcal{G}_0, \cdot)$, if we assume less, we have to resort to
subgroups of \((\mathcal{G}_0, \cdot)\). In case of Gaussian noise we see that the group of permutations and that of sign changes are subgroups of the orthogonal group. This might lead to the conclusion that if we can assume power defined noise, we should use a DP method corresponding to that. Similarly if we assume that the noise is both coordinatewise symmetric and exchangeable, than the DP method corresponding to the exchangeability assumption should be used. Figure 2.6 illustrates that this is not the case. The power functions for different inputs are shown, and the SPS method relying on symmetry outperforms the DP method relying on exchangeability on Figure 2.6a, while it is the other way around on Figure 2.6b. The reason for this has to do with the sufficient excitation condition in Definition 2.27. The power of the tests depend on how well the condition \(Q_G(X) > 0\) is satisfied for a uniformly distributed \(G\) over the group corresponding to a specific assumption. It can very well happen that, on average, the elements of a smaller group perform better on this input dependent test than those of a larger group. In case of our example the group of sign changes and that of permutations are disjoint groups and by carefully selecting two different input matrices the order of the power functions belonging to exchangeable and symmetric noise distributions could be exchanged. As we already pointed out after Definition 2.27, the DP method for exchangeable noises does not work at all for constant input. A small deviation from constant makes the input sufficiently exciting, but it is intuitive that permuting an almost constant input makes less of an impact than randomly changing signs of the inputs. This is the reason why the power function belonging to the symmetric assumption is above the power function belonging to the exchangeability assumption in Figure 2.6a. On the other hand if the input is taken from equidistant points of a straight line going through the origin, then the order of power functions change, as shown in Figure 2.6b.

One might argue that the order of these two DP methods could be exchanged because they rely on disjoint symmetry groups (apart from the unit element) but for proper subgroups there would be a definite ordering between DP methods. To illustrate that the problem is more complicated than this we consider a fourth DP method, relying on the symmetry group that is the closure of sign changes and permutations. This is a finite subgroup of \(SO(N)\) and it translates to assuming exchangeable noise distribution with symmetric marginals. Figure 2.7 shows the
Figure 2.7: Difference of power functions for different assumptions with different input

difference between power function corresponding to different assumptions compared to that corresponding to the power defined assumption as baseline. What should be emphasized about these figures is that the difference $\Delta$ Exch-Symm goes bellow zero in both cases, meaning that assuming less about the noise distribution (only symmetry of marginals and exchangeability) seems to offer better statistical power than assuming power definedness. Although this may seem counterintuitive, it underlines the importance of examining the behavior of power functions offered by different possible choices of assumptions.

We consider this phenomenon important enough, to summarize it in the following remark. The consequences of this remark could also lead to territories of input design that we are not going to pursue in this dissertation.

**Remark 2.32.** When more than one DP method is applicable to a problem, the best one is not obviously the one that relies on the most restrictive assumptions. The order of methods depends on the actual input $X$ as well.
Another important lesson that can be learned from Figure 2.6 is that if the input is sufficiently exciting then relaxing on the assumptions does not cost that much in terms of statistical power. The power function did drop when loosening the assumption from known Gaussian distribution, to any power defined distribution, but this loss might be acceptable if the reliability of the Gaussian assumption is questionable.

The fourth and last item that needs to be emphasized is the computational complexity of these methods. We talked about testing a single model parameter $\theta$ whether it is accepted by the test or not. Although this membership test characterizes the confidence regions completely it is not a very practical way of defining a set. If an algorithm is designed to take uncertainty of the estimates into account, it seldom relies on the characteristic function of the confidence set, but much more frequently on some handy description of that. Ellipsoidal region descriptions, or polytopes are the most common [Boyd et al., 1994]. As it can be seen on Figure 2.4, the shape of DP confidence regions can be quite exotic. There are different ways to explore sets with such complex surfaces, the easiest one being simple gridding that is offered in [4], using interval analysis as in [2] or [Kieffer and Walter, 2014] or by finding and following the surface starting from an internal point as it was done when the graphs in this dissertation were created. The common problem of all these approaches is that they are computationally heavy and their output is very complex. This is the main motivation behind the contents of Chapter 4.

### 2.5 DP Methods for Linear Dynamical Systems

So far in this chapter we have introduced DP methods and developed insight into their behavior through the linear regression model structure. The goal of this section is to extend our understanding to generalized Box-Jenkins models, as defined in Section 1.2.1. First, we are going to take a look at the performance measure given in Theorem 2.28 and we provide different interpretations of it in Section 2.5.1. This will allow us to define performance measures for dynamical systems that are natural extensions from the linear regression case. One of these interpretations will be based on the gradient of the prediction error cost function, that was used when
the SPS method was generalized in [Csáji et al., 2012b]. We are going to elaborate on this choice in Section 2.5.2. We take a different approach in Section 2.5.3 for output error models and we further that approach in Section 2.5.4 to more general model structures. The computational complexity and implementation details will be discussed in 2.5.5. We conclude with some comments about the defined DP methods that are important in case of dynamical system identification.

2.5.1 Different Interpretations of the Performance Measure

Let us remind ourselves of a few facts about the least squares estimate for linear regression problems, where the LS estimate $\hat{\theta}^{LS}_N$ is defined as

$$\hat{\theta}^{LS}_N = \arg\min_{\theta} V_N(\theta) \quad V_N(\theta) = \frac{1}{2N} \sum_{k=1}^{N} (y_i - x_k^T \theta)^2$$

(2.43)

and the solution is given as $\hat{\theta}^{LS}_N = [XX^T]^{-1}XY$. The gradient of the cost function can be written as

$$\frac{\partial V_N(\theta)}{\partial \theta} = -\frac{1}{N} \sum_{k=1}^{N} (y_i - x_k^T \theta) x_k = -\frac{1}{N} X(Y - X^T \theta)$$

(2.44)

and the covariance of the estimate is

$$\mathbb{E} \left( (\hat{\theta}^{LS}_N - \theta_0)(\hat{\theta}^{LS}_N - \theta_0)^T \right) = [XX^T]^{-1}$$

(2.45)

As we have argued in the previous sections, the performance measure (2.35) defined in Theorem 2.28 is an appropriate choice for linear regression problems. We give three equivalent formulae for this performance measure

$$Z(\theta, X, Y) = (Y - X^T \theta)^T X^T [XX^T]^{-1} X (Y - X^T \theta)$$

(2.46)

$$= (Y - X^T \theta)^T X^T [XX^T]^{-1} [XX^T][XX^T]^{-1} X (Y - X^T \theta)$$

(2.47)

$$= (X^T [XX^T]^{-1} X \theta - \theta)^T [XX^T]^{-1} [XX^T]^{-1} X Y - \theta$$

(2.48)

$$= (X^T [XX^T]^{-1} X Y - X^T \theta)^T (X^T [XX^T]^{-1} X Y - X^T \theta)$$

(2.49)
Let $\hat{Y}(Y, X) = X^T \theta$ denote the predicted output of the model $\theta$ on the current dataset, $\hat{\theta}(Y, X, V) = [XX^T]^{-1}XY$ denote the estimate belonging to the dataset and the cost function $V$, $\hat{\Sigma}_N(Y, X) = [XX^T]^{-1}$ denote the estimated covariance of the estimate $\hat{\theta}$ on the given data, $\frac{\partial V(\theta)}{\partial \theta}(Y, X, V) = -\frac{1}{N}X(Y - X^T \theta)$ denote the derivative of the cost function $V$ corresponding to the given data and $H_V(\theta) = \frac{1}{N^2}XX^T$ denote the Hessian of the cost function at $\theta$. We will not denote the arguments of these quantities when it is obvious from the context. Using these notation equations (2.46), (2.48) and (2.49) can be rewritten as

$$Z(\theta, X, Y) = \left(\frac{\partial V(\theta)}{\partial \theta}\right)^T H_V^{-1}(\theta) \frac{\partial V(\theta)}{\partial \theta}$$

$$= (\hat{\theta} - \theta)^T \hat{\Sigma}_N^{-1}(\hat{\theta} - \theta)$$

$$= (\hat{Y} - \hat{Y})^T (\hat{Y} - \hat{Y})$$

As we move away from linear regression problems with least squares estimation to other model structures and other estimation criteria, the notions of estimate, cost function derivative and estimate covariance will get a different meaning. For more complex model structures equations (2.50), (2.51) and (2.52) will no longer define the same performance measure. As Theorem 2.17 states that the confidence level of a DP method is independent of the used performance measure we are free to take any of these performance measures when we create DP methods for dynamical systems but we should bear in mind that using a “wrong” performance measure will lead to confidence regions that might not be the best, as we have illustrated it in Section 2.4.2.

Let us briefly discuss the implications of these performance measures, keeping in mind that we are aiming for good confidence regions for parameters of generalized Box-Jenkins models. These approaches are going to be discussed more in later sections, but we mention here some of their most important properties, to give a quick overview of their relation.

Starting with (2.50), this can be interpreted as the weighted norm of the gradient of the cost function, where the weighting matrix is the inverse of the Hessian of the cost function. The biggest advantage of the gradient based performance measure is that it can be evaluated efficiently, as the gradient and the Hessian can
be calculated from the data directly. The disadvantage of this approach is that
the confidence regions belonging to this performance measure will have unwanted
structural properties due to the complex nature of the considered cost function.
The measure is parametrization dependent and it is hard to give intuitive justifi-
cation for its use.

Looking at (2.51) the performance measure is defined as the weighted distance
between the tested model $\theta$ and the estimate $\hat{\theta}$ belonging to the given dataset,
using the inverse of the estimated covariance matrix of $\hat{\theta}$ as weighting. This is a
quantity that appears naturally around confidence regions for dynamical systems,
as in the asymptotic regime this value (evaluated with the actual measurement
data) is $\chi^2_{n_{\theta}}$ distributed and that is used to create the confidence regions. Although
this is already seems to be a more natural choice for a performance measure than
(2.50), there are two things that need to be emphasized about it. The first is
that evaluating this performance measure involves finding $\hat{\theta}$. In the context of DP
methods this means $m-1$ estimation problems on the perturbed datasets whenever
testing a given parameter $\theta$. Second, the estimated covariance $\hat{\Sigma}_N$ is calculated
on a perturbed dataset. As we test models away from $\theta_0$ these perturbed datasets
have decreasing signal to noise ratio. As the reliability of the covariance matrix
estimate depends on the SNR, this makes the estimated covariance matrix less and
less reliable. To sum up, we have a more justified system theoretic interpretation
for the performance measure but this comes with a high increase in computation
complexity.

The last interpretation, given in (2.52), can be described in words as the dis-
tance between the predictions of the tested model $\theta$ and the estimated model $\hat{\theta}$
on the current measurement input. As in the previous case, evaluation of this per-
formance measure requires solving an estimation problem on the sample dataset.
Compared to that case it is an improvement that only the parameter vector $\hat{\theta}$ is
estimated, the Hessian does not factor into the performance measure. Another
beneficial property is that if the input is such that it does not differentiate be-
tween models in a subset $A \subset \theta$, the performance measure will make no distinction
between these models either.

From now on we will refer to the performance measures defined by equations
(2.50), (2.51) and (2.52) as “gradient”, “estimate distance” and “prediction dis-
Table 2.5: Pros and cons of different performance measures

<table>
<thead>
<tr>
<th>Performance measure based on</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>gradient</td>
<td>low computational complexity</td>
<td>no clear system theoretic interpretation</td>
</tr>
<tr>
<td>estimate distance</td>
<td>better system theoretic interpretation</td>
<td>high computational complexity</td>
</tr>
<tr>
<td>prediction distance</td>
<td>good system theoretic interpretation</td>
<td>high computational complexity</td>
</tr>
</tbody>
</table>

...tance” based performance measures. The pros and cons of these are summarized in Table 2.5. Relying on this insight, we will pursue the investigation of the gradient based performance measure because of its computational efficiency and the prediction distance based performance measure for its good system theoretic properties.

2.5.2 Gradient Based Performance Measure

When results for dynamical systems LS estimates are proved that generalize corresponding results from linear regression problems, the first attempt is usually to replace the regressors with the gradients of the prediction errors. This is one way to prove most of the asymptotic results for LTI estimates and this was also the approach of [Csáji et al., 2012b] when the SPS method was extended from linear regression models to dynamical systems.

As we already mentioned, this corresponds to the performance measure

$$Z(\theta, X, Y) = \left( \frac{\partial V(\theta)}{\partial \theta} \right)^T H_V^{-1}(\theta) \frac{\partial V(\theta)}{\partial \theta}$$  \hspace{1cm} (2.53)

that can be efficiently evaluated but it does not have a well justified system theoretic interpretation. The goal of this section is to illustrate problems that arise when this performance measure is used to create a DP method.

Suppose that $\hat{\theta}_V$ corresponds to the estimate minimizing the twice differentiable cost function $V$. In this case $Z$ is a positive quadratic form and $\frac{\partial V}{\partial \theta}(\hat{\theta}_V) = 0$. When $\hat{\theta}_V$ is tested for membership $Z_1$ is going to be zero. Since the cost functions
belonging to the perturbed datasets are a bit different (due to their dependence on the perturbed output $Y^{(i)}$) it is highly unlikely that the other $Z_{k\geq 2}$ values will also be zero, so the estimate $\hat{\theta}_V$ will always be in the confidence set if the gradient based performance measure is used. The problem that arises when complex cost functions are used is that the previous argument is also true for every parameter vector $\theta$ for which $\frac{\partial V}{\partial \theta}(\theta) = 0$, including local maxima and inflection points of the cost function. There is no apparent system theoretic reason why these points should be included in a confidence region, it seems to be an artifact of the particular performance measure. If we consider the least squares prediction error cost function on a sample of length $N$, the cost function is an approximately $2N$ degree polynomial of the model parameters where the coefficients depend on the actual measurement data. This means that there are approximately $N$ points that will be included in the confidence region and we know nothing about their position in the parameter space.

We illustrate this effect using a first order output error model. It is known (see [Ljung, 1999]) that the gradient of the LS cost function can be given as

$$\frac{\partial V}{\partial \theta}(\theta) = \frac{1}{N} \sum_{k=1}^{N} \varepsilon(\theta)[k] \frac{\partial \varepsilon(.)[k]}{\partial \theta}(\theta)$$  \hspace{1cm} (2.54)$$

where the derivative of the prediction errors $\varepsilon(\theta)[k]$, with respect to coefficients in $B(\theta, z)$ and $F(\theta, z)$, are given as

$$\frac{\partial \varepsilon(\cdot)}{\partial b_k}(\theta) = \frac{1}{F(\theta, z)} z^{-k} u$$  \hspace{1cm} (2.55)$$

$$\frac{\partial \varepsilon(\cdot)}{\partial f_k}(\theta) = -\frac{B(\theta, z)}{F^2(\theta, z)} z^{-k} u$$  \hspace{1cm} (2.56)$$

Since $B(\theta, z) = 0$ implies $\frac{\partial V}{\partial f_k}(\theta) = 0$ for every $k \in [n_{\mathcal{F}}]$, we know that there is going to be at least one value $f_1$ where $\frac{\partial V}{\partial \theta}(\theta_1) = 0$ (at least if $n_{\mathcal{F}} = 1$). Finding this value makes it possible to explore connected components of a confidence region around $\theta_1$ and $\hat{\theta}_L$ generated by the gradient based DP method. Figure 2.8a shows the connected component of the confidence region around $\hat{\theta}_L$. The sample data was created with high SNR to allow good estimation. The inside of the marked polygon was tested using a fine grid and the blue region consists of points that passed the
Figure 2.8: Not connected confidence region from gradient based DP method

(a) Connected component around $\hat{\theta}^{LS}$
(b) Connected components around $\hat{\theta}^{LS}$ and an inflection point

This example shows that, unlike in the case of linear regression, the gradient based confidence regions can be not connected. One consequence of this fact is that exploring a connected component of a confidence set and finding its contour (as shown in Figure 2.8a) is not enough to find the whole set. If the uncertainty of an estimate is measured by the volume of the generated confidence set, then evaluating such a measure is difficult, as guaranties are needed to ensure that all components of the confidence set were explored. Currently we can offer no guaranties that Figure 2.8b is the entire confidence region, it very well may be that there are other components as well. This is only going to become worse in higher dimensional situations. As the values $f_1$ in the connected component around the inflection point are positive, it is evident that this model cannot belong to a properly sampled system, as such parameters correspond to models with oscillation between samples. This shows that there is no good reason for including that part of the parameter space into the confidence region.

As it is shown in [Garatti et al., 2004], it is not necessarily a problem if a confidence region is not connected in a given parameterization, as the Euclidean distance in the parameter space has almost nothing to do with any meaningful...
system theoretic notion of distance. The property that good confidence regions should aim for is to contain system parameters that are close to each other in a system theoretic sense.

Although it is clear to us, from a purely theoretic point of view, that gradient based performance measures should be discarded when DP methods are created, they still need to be considered because of practical reasons concerning computational complexity. This is going to be elaborated in Section 2.5.5.

### 2.5.3 The Output Error Model

Although there are dynamical system models that can be reformulated as linear regression, from the point of view of DP methods, the output error (OE) model structure is the closest to linear regression, so we leap forward in this direction. From this point on we will consider the prediction distance as performance measure. This section will introduce the notion of output space that plays a key role in DP methods relying on prediction distance. We will give a visual interpretation of how prediction distance based DP methods work. Before doing that, we need to check what are the consequences of Assumptions 2.3 and 2.5 in the OE case, where the model is defined by taking the polynomials \( A(z) = C(z) = D(z) = 1 \).

Assuming that the parameter vector \( \theta \) contains the coefficients of the polynomials \( B(z) \) and \( F(z) \), the state space description of the model can be given as

\[
Y = \Gamma_{B/F} x_{B/F}[0] + H_{B/F} U + E
\]

(2.57)

where \( U \) is the measured input, \( E \) is the unobserved noise, \( Y \) is the observed output, all \( \mathbb{R}^N \) vectors. \( x_{B/F}[0] \) is the initial condition of the system with dimensions \( \max(n_B, n_F) \) and \( \Gamma_{B/F} \) and \( H_{B/F} \) are the matrices defined in (1.14) and (1.15).

In order to meet the invertibility assumption, we need to express the noise \( E \) as

\[
E = Y - \Gamma_{B/F} x_{B/F}[0] - H_{B/F} U
\]

(2.58)

which immediately shows us that the initial conditions \( x_{B/F}[0] \) either need to be assumed known, or they have to be included in \( \theta \) along with the coefficients of the polynomials. As in the case of OE models the subscript \( \cdot_{B/F} \) is redundant, we will omit it in the rest of the section.
Definition 2.33 (Estimated output). The estimated output of a model $\theta$ for a given input $U$ is denoted by $\hat{Y}_\theta$ given as

$$\hat{Y}_\theta(U) = \Gamma_{B/F}x_{B/F}[0] + H_{B/F}U$$  \hspace{1cm} (2.59)$$

In case of the output error model the estimated output is the same as the predicted output, but for more general models this is no longer true. The estimated output is defined as considering the noise as zero $E = 0$, while predictions usually incorporate estimates of the past noise values.

Definition 2.34 (Output space). Let $\Theta$ denote the whole parameter space for a specific choice of $n_B$ and $n_F$. The output space corresponding to an input $U$ is

$$\mathcal{O}_U = \left\{ Y \in \mathbb{R}^N \mid \exists \theta \in \Theta : Y = \hat{Y}_\theta(U) \right\}$$  \hspace{1cm} (2.60)$$

Definition 2.35 (Model distance). The distance of two models $\theta_1, \theta_2 \in \Theta$ with respect to the input $U$ is defined as

$$\|\theta_1 - \theta_2\|_U = \|\hat{Y}_{\theta_1}(U) - \hat{Y}_{\theta_2}(U)\|$$  \hspace{1cm} (2.61)$$

where any norm over $\mathcal{O}_U$ can be used to induce a norm over $\Theta$.

Since the input $U$ is constant throughout the whole process of hypothesis testing and confidence region generation, we will no longer denote the dependence on $U$ to ease notation.

Simply put, the output space is the subset of $\mathbb{R}^N$ that contains the possible noise free outputs of all models from the considered model class with respect to the given input. The main difficulty of system identification is that this subspace of $\mathbb{R}^N$ does not have a simple structure for finite order polynomial models. What makes the estimation of $N$ order finite impulse response (FIR) models simple is the fact that for a given input $U$, the output space $\mathcal{O}_U$ is a linear subspace of $\mathbb{R}^N$. This is no longer true for fixed order polynomial models. Nonetheless, in order to show how prediction distance based DP methods work, we are going to examine them assuming that $\mathcal{O}_U$ is a linear subspace of $\mathbb{R}^N$.

The visualized argument that we present in the sequel applies only to the case when the noise distribution is assumed to be invariant under a subgroup of $O(N)$.  

Chapter 2. Data Perturbation Methods

The property that connects OE models and linear regression models is that the noise $E$ is added directly to the nominal output. This fact, in combination with assuming orthogonal perturbation matrices $G_k^H$, means that when the perturbed data sets are created then the perturbed noise vectors have the same length. Let us consider how membership of a model $\theta$ in a $\frac{1}{2}$ confidence set is decided. The first perturbed output $Y_{\theta}^{(1)} = Y$ for every tested model $\theta$, while $Y_{\theta}^{(2)} = \hat{Y}_{\theta} + G_2^H (Y - \hat{Y}_{\theta})$.

Let $\hat{\theta}^{(1)}$ and $\hat{\theta}^{(2)}$ denote the estimates obtained using the common input $U$ and the perturbed outputs $Y_{\theta}^{(1)}$ and $Y_{\theta}^{(2)}$ respectively. With this notation, the prediction distance based performance measure defines the value $Z_i$ as

$$Z_i = \|\hat{Y}_{\theta}^{(i)} - \hat{Y}_{\theta}\|^2 = \|\hat{\theta}^{(i)} - \theta\|^2_U.$$

**Remark 2.36** (Estimation procedure dependence). It is important to note that the prediction distance based performance measure depends on the estimation procedure used to determine $\hat{\theta}^{(i)}$, but any prediction error or subspace method based estimate can be used.

The goal of every estimation procedure is to find a suitable parameter vector $\theta$ such that its predicted output $\hat{Y}_{\theta}$ is as close to the measurements $Y$ as possible. The advantage of using a least squares criterion and having $O_U$ as a linear subspace of $\mathbb{R}^N$ is that the estimate can be calculated analytically and $\hat{Y}_{\theta}$ is simply the orthogonal projection of $Y$ to $O_U$. Figure 2.9 illustrates the two typical scenarios when the hypothesis test of a parameter vector $\theta$ is performed. If the perturbation matrix $G_2^H$ is orthogonal then the distance of the original measurement $Y = Y^{(1)}$ and the perturbed measurement $Y^{(2)}$ from $\hat{Y}_{\theta}$ are the same. It is the distance of their projection to the subspace $O_U$ from $\hat{Y}_{\theta}$ what determines the outcome of the hypothesis test.

Once we move away from the situation where the least squares prediction error criterion is used, $\hat{Y}_{\theta}^{(i)}$ is no longer the closest orthogonal projection of $Y^{(i)}$ onto $O_U$, but if $\hat{\theta}^{(i)}$ is obtained as a meaningful estimate, it is a good approximation. As the output for a given fixed input $U$ is linear in the nominator coefficients and the initial conditions, we observe that that predicted outputs of models with the same denominator $F(z)$ from an $n_B + \max(n_B, n_F)$ dimensional linear subspace in $\mathbb{R}^N$. Since the predicted output of these models is a continuous function of the
Chapter 2. Data Perturbation Methods

coefficients in $\mathcal{F}$, these subspaces also have a nonlinear parametrization with $n_\mathcal{F}$ parameters. This means that in case of the OE model structure $\mathcal{O}_U$ is the union of a continuum number of linear subspaces of $\mathbb{R}^N$.

In order to transform the heuristic argument outlined above into a rigorous proof some major difficulties need to be handled. The relation between $Y^{(i)}$ and $\hat{Y}_{\hat{\theta}(i)}$ is complicated due to the complicated nature of finding the estimate $\hat{\theta}^{(i)}$. This fact combined with the structure of $\mathcal{O}_U$ makes us believe that confidence regions accepted by prediction distance based DP methods are not connected. A rigorous proof is not given for this, but a numerical experiment in Section 2.5.5 will show that this is the case up to numerical precision.

We have discussed the performance measures of prediction distance based DP methods and conjectured that the corresponding confidence regions are not connected but we did not discuss their boundedness property. In case of linear regression models we have seen that this property depends on the measured input to the system $X$. Similarly, for linear dynamical systems the boundedness of the confidence regions will depend on the input $U$, more precisely on $\mathcal{O}_U$. If the input is rich enough to differentiate between any two models $\theta_1$ and $\theta_2$ in the sense that $\hat{Y}_{\theta_1} \neq \hat{Y}_{\theta_2}$ then the prediction distance based DP methods will result in bounded confidence regions (in the sense that $\hat{Y}_C$ is a bounded subset of $\mathbb{R}^N$).
2.5.4 General Noise Models

The goal of this section is to move away from the OE model structure to handle more complex noise models. Since we were not able to rigorously prove structural results for the OE models, this will be the case also for more complex models. The goal of this section is to show that the same performance measure that was used for OE models is good for other noise models as well. We will emphasize the points that need to be considered when more general noise models are handled.

The first aspect that needs to be considered is that of parameterization. The output of the generalized Box-Jenkins model can be given as

\[ Y = \Gamma_{1/A}x_{1/A}[0] + H_{1/A} \left( \Gamma_{B/F}x_{B/F}[0] + H_{B/F}U + \Gamma_{C/D}x_{C/D}[0] + H_{C/D}E \right) \]

In order to meet the invertibility assumption, we need to be able to express \( E \) using the measurement data and model parameters. This means that along with the model parameters and plant model initial conditions, the initial conditions of the noise model \( x_{C/D}[0] \) also need to be considered as parameters that are to be estimated.

Adding the initial conditions to the parameter space is a major difference compared to the classical asymptotic theory. Of course initial conditions do not matter in the asymptotic sense, this is the reason why the asymptotic confidence regions are defined in the space of model parameters only. Nonetheless, the estimate for the asymptotic covariance used in the description of these confidence regions depends on the obtained estimate for the initial conditions, it allows a better alignment of the covariance matrix. So in the case of asymptotic confidence regions we estimate both initial conditions and model parameters and we construct a confidence region only for the model parameters. While in the case of DP methods the confidence region is constructed in the joint space of model parameters and initial conditions.

We propose to use formally the same performance measure for general noise models that was defined for OE models, namely

\[ Z_i = \| \hat{Y}_{\theta^{(i)}} - \hat{Y}_\theta \|^2 = \| \hat{\theta}^{(i)} - \theta \|^2_{U_i} \quad (2.63) \]
Although formally this is the same as for OE models, mind that there is an underlying estimation step to find $\hat{\theta}^{(i)}$ that takes into account the modified model structure.

In case of the OE model the perturbed datasets were created as $Y_{\theta}^{(i)} = \hat{Y}_{\theta} + G_{\Pi}^{\Pi}(Y - \hat{Y}_{\theta})$. Irrespective of the tested model $\theta$, the perturbation was a constant transformation with $G_{\Pi}^{\Pi}$ of the vector $Y - \hat{Y}_{\theta}$ around the center $\hat{Y}_{\theta}$. In case of higher order noise models the perturbed datasets are defined as

$$Y_{\theta}^{(i)} = \Gamma_{1/A}x_{1/A}[0] + H_{1/A}(\Gamma_{B/F}x_{B/F}[0] + H_{B/F}U + \Gamma_{C/D}x_{C/D}[0]$$

$$+ H_{C/D}G_{i}^{\Pi}H_{C/D}^{-1}$$

$$\left(\Gamma_{1/A}(Y - \Gamma_{1/A}x_{1/A}[0]) - \Gamma_{B/F}x_{B/F}[0] - H_{B/F}U - \Gamma_{C/D}x_{C/D}[0]\right)$$

where the most notable part is the core perturbation of the noise $H_{C/D}G_{i}^{\Pi}H_{C/D}^{-1}$ which depends on the parameters $\theta$. Because of this fact, the intuitive geometrical interpretation presented in case of OE models is no longer valid.

There is a particular situation that needs special attention when $U = 0$ or $B(z) = 0$, i.e. when parameters of an auto regressive moving average time series are estimated. The proposed performance measure results in $Z_i = 0$ for all $i \in [m]$, as $\hat{Y}_i = 0$ in this case. This degrades the hypothesis test to a coin toss as in Theorem 2.24. If one wants to create a non-degenerate DP method for time series analysis, a significantly different performance measure is needed. The possible choices include using the gradient or estimate distance based performance measures, bearing in mind the consequences of these choices.

### 2.5.5 Computational Complexity

The goal of this section is to present how can a confidence region be discovered, or approximated, when this region belongs to a DP hypothesis test. There are two aspects of this questions that need to be discussed. The first is to discuss the computational complexity required to test the membership of a single model $\theta$. In case of dynamical systems evaluation of the proposed performance measure involves solving $m - 1$ full blown estimation problems on the perturbed datasets. On top of this, when parameters are tested that are far away from the estimate
belonging to the original dataset these estimation problems will have a bad signal to noise ratio, that usually has a detrimental effect on the reliability of the used underlying estimation procedure.

The second aspect is the need to find the entire confidence region. As we have seen the gradient based DP methods can produce disconnected confidence regions. There is also no guarantee that the confidence region of prediction distance based DP methods will be connected in the parameter space. This shows that exploration of confidence regions should be done such that it provides guarantees that no parameter value outside the discovered set will pass the hypothesis test.

Interval analytic methods make it possible to create such an algorithm, as shown in [2] or [Kieffer and Walter, 2014]. However, there are difficulties transforming this theoretical possibility into practically applicable solutions in general.

We would like to mention that set membership or bounded error identification methods operate with the same interval analytic principle. Based on the assumed bounds there is a set of possible error values that are unknown. These identification methods aim at finding the set of all parameters that are consistent with some value of the errors adhering to the assumed bounds. When it comes to characterizing a set of identified parameters or a DP confidence region, the same curse of dimensionality is faced by set membership or bounded error methods as by DP methods.

Without going into the details we present the main ideas of interval analysis, so we are able to present the main difficulty of this technique. The basic ideas of interval analysis are described in [Moore, 1966], while [Hansen, 1992] describes how this framework can be used efficiently in global optimization. This tool was used earlier in system identification context as well (see [Kampen et al., 2011] or [Jaulin and Walter, 1993]) and there are software packages available to facilitate implementation of interval analytic algorithms [Rump, 1999].

The main idea is that operands are intervals instead of numbers and the result of an operation is again an interval with the property that whatever combination of the operand values is used from the operand sets, the resulting value will be inside the result interval. It is up to the implementation how this is achieved. For example the product with two operands $[-1, 2] \cdot [-1, 2]$ can be defined to have the result set $[-2, 4]$, which is the tightest possible result. However $[-5, 5]$
would also be acceptable, as all possible results are still covered. One of the main difficulties when designing efficient interval analytic algorithms is the minimization of decoupling. Decoupling is the phenomenon, when the bounds of the result set are calculated such that different appearances of the same variable are not considered as one value. It is best illustrated in an example, lets consider $x^2$, where $x \in [-1, 2]$. A result set can be calculated using the fact that $x^2 = x \cdot x$, and the bounds of the result set are calculated as $x \cdot y$, where both $x, y \in [-1, 2]$. $x$ and $y$ are decoupled instances of the same variable. The resulting set obviously fits the criteria of interval analytic definitions, as all possible results are covered with the case $x = y$, but lots of other values are also included in vain. Evaluating $x \cdot x$ gives $[-2, 4]$ as the result set. In this case we know that tighter bounds can be calculated, and an interval analytic algorithm can be implemented to evaluate $x^2$ in a way to result in bounds $[0, 4]$. When a formula needs to be evaluated with interval variables, it is crucial to minimize the decoupling. In case of linear regression problems the effects of decoupling can be kept to a minimum such that the tightest bounds are used when the confidence regions are generated. This is presented in [Kieffer and Walter, 2014]. But this cannot be done for dynamical systems models, as the evaluation of performance measures involves summation of different powers of the model parameters. For this reason efficient (or even remotely usable) interval analytic implementations of DP methods for dynamical systems are not possible. This obstacle is there irrespective of the chosen performance measure.

Due to the fact that discovering the entire confidence region currently seems computationally unfeasible, we propose the use of its principal connected component. As the estimate $\hat{\theta}$ belonging to the original dataset is always accepted by DP methods based on any discussed performance measure, the connected component around $\hat{\theta}$ can be used as an approximate confidence set. Although this is only a lower approximation, its approximation error is expected to be smaller that that of the asymptotic confidence set.

In order to illustrate the usability of the principal connected component approximation we try to evaluate the whole parameter space for a piece of synthetic data in case of a first order OE model. We do this by evaluating the membership test on a fine grid over $(-\pi/2, \pi/2)^2$ and the model is parametrized such that
Chapter 2. Data Perturbation Methods

Figure 2.10: Justification of the principal connected component heuristic

\[ b_1 = \tan(\theta_1) \text{ and } f_1 = \tan(\theta_2). \] Figure 2.10a shows the result of this evaluation.

The part of the parameter space between the red lines corresponds to stable models and the principal component of this confidence set is marked with the circle. The first thing to be noted is that there appear to be many nonconnected components on the figure. Although this seems convincing, we would like to emphasize that all of the nonprincipal components lie close to the border or outside the stability region and the obtained \( \hat{\theta}^{(i)} \) estimates in these cases might not be numerically reliable. We are inclined to believe that confidence regions of prediction distance based DP methods are not connected, not even for the OE model class, but this figure is not a solid evidence to that.

For every model \( \theta \) that was accepted by the test, we can calculate the corresponding noise sequence \( E(\theta) \). Figure 2.10b shows a scatter plot of the standard deviation of these noise sequences against their mean, with the points corresponding to the principal connected component circled. This figure shows that noise sequences in the nonprincipal connected component have significant bias and considerably higher variance. As a result, if the distribution of the noise would actually be known, these noise sequences should have a really low density value. This means that removing the corresponding points from the confidence regions should not have a big impact on the confidence level of the remaining set. Nonetheless, it should be kept in mind that the use of the principal connected component instead of the full set is just a heuristic.
Finding the surface of a connected set can be done relatively efficiently, even if the performance measure used to define the membership test is the prediction distance based measure. It needs to be acknowledged that although the gradient based performance measure has less theoretical foundation, it is computationally more efficient than other performance measures relying on the solution of the perturbed estimation problems. This computational gain can justify the use of the gradient based performance measure when finding the principal connected component of the confidence set.

There is an implementation detail that we would like to emphasize in connection with gradient based DP methods. As given in [Csáji et al., 2012b], the gradient of the prediction error needs to be calculated for the tested model \( \theta \). When PEM methods are implemented the same expressions are used to calculate the gradient of the cost function in many different points [Ljung, 1999]. The main difference between the two scenarios is that while in case of the PEM methods these gradients need to be calculated for many different models, the membership test in a DP method requires the evaluation of the gradient only at the tested model. This allows a more efficient calculation of the gradients. We illustrate this using the partial derivatives with respect to the coefficients of \( C(z) \), that is given as

\[
\frac{\partial}{\partial c_k} \hat{y}[t] = -\frac{D(z)B(z)}{C(z)C(z)F(z)}u[t - k] + \frac{D(z)A(z)}{C(z)C(z)}y[t - k] = \frac{1}{C(z)} \varepsilon[t - k] \quad (2.64)
\]

where \( \varepsilon[\cdot] \) denotes the prediction errors. Mind, that the \( y[\cdot] \) in this formula comes from the measurements when \( Z_1 \) is calculated, and it corresponds to the perturbed datasets when calculating \( Z_{i \geq 2} \). Since the generation of the perturbed datasets requires explicit calculation of \( E(\theta) \), which is exactly \( \varepsilon[\cdot] \), we can choose between the two previous formulas. Since the filter \( \frac{1}{C(z)} \) is smaller order (less than half) than either \( \frac{D(z)B(z)}{C(z)C(z)F(z)} \) or \( \frac{D(z)A(z)}{C(z)C(z)} \), computing the gradient as \( \frac{1}{C(z)} \varepsilon[t - k] \) costs less than 25% compared to the other formula. In general it can be stated that instead of just using \( Y \) and \( U \) in the gradient calculation, we can also rely on formulas depending on \( \varepsilon[\cdot] \) (as DP methods cannot avoid their calculation). Comparing the order of the involved filters and selecting formulas involving lower order filters will reduce the cost of gradient calculation and the amount of reduction depends on the model structure. For OE models there is no possible gain, but for auto
regressive models we have a more than 75% gain. The achievable gain for other model structures lies somewhere between these two extremes.

2.6 Place in the Identification Workflow

The goal of this chapter was to provide means of confidence region generation for a parameter estimate $\hat{\theta}$ based on knowledge of the model structure, a set of finite samples and mild assumptions about the contaminating noise. Although within these frames we tried to give a picture as full as possible, we have to remind ourselves that this is just a small piece in the puzzle that is the full identification workflow.

Before getting to the starting point of the presented theory, we have to make certain choices that will influence the outcome of the presented algorithms. After having the results any self-respecting practitioner would want to validate these results. The goal of this section is to briefly comment on those steps of the workflow that precede or follow the use of the data perturbation methodology.

There are two assumptions that are crucial for providing the exact confidence level hypothesis tests. The first one is the exact knowledge of the model structure, given by Assumption 2.3 and the selected model structure should also be invertible with respect to the noise as in Assumption 2.5. The second such assumption is formulated in the conditions of Theorem 2.17, namely that the distribution of the noise should have some symmetries.

First and foremost we have to choose what assumptions are we going to make about the noise distribution (i.e. the noise is an i.i.d. sequence). Once this assumption is made, we can search for a specific class of model structures and an appropriate model order $n_{\theta}$. The regular machinery for model selection can be used, see for example Chapter 16 in [Ljung, 1999]. These methods provide ways for selecting a model order based on the measured data $X$ and $Y$ and assumptions about the noise, so the later use of DP methods has no impact on the outcome of these decisions. If we decided on a specific class of model structures (i.e. linear regression, or linear dynamical systems) and model order $n_{\theta}$ then Assumption 2.3 about perfect knowledge of $f$ can be considered fulfilled. It might be that data
perturbation ideas can be used for designing model selection algorithms as well, but currently such theory is not yet available.

It is important to remark that Assumption 2.3 about the full knowledge of the function \( f \), that describes the model structure, does not put any restriction on the model structure. Every imaginable model (linear or nonlinear) fits into this frame. We deliberately chose a nonlinear model in Section 2.2 to introduce the framework, stressing the point the method can be applied to nonlinear models as well. The real restriction comes from the invertibility assumption, Assumption 2.5, but the nature of that assumption is such that it can be fulfilled or violated by both linear and nonlinear models.

Once we have Assumption 2.3 fulfilled, we still have to check whether the selected model structure is invertible or not with respect to the noise, that is, is Assumption 2.5 fulfilled or not. As we already mentioned in Section 2.5.4, this assumption implies that the initial conditions of the noise model need to be known or they have to be included into the set of estimated parameters in order to fulfill the invertibility assumption.

There can be other circumstances that might prohibit the applicability of DP methods in cases where it could be applicable under ideal conditions. One such condition is where we have corrupted or missing measurements. In case of linear regression models, where observations can be left out from the identification process this is not a serious issue, as dropping the measurements that are suspicious for being corrupted makes the dataset still consistent. However, having a missing or corrupted measurement in the case of dynamical system models, where expressing the contaminating noise relies on all past data, is detrimental. Since the noise cannot be expressed, the invertibility condition cannot be satisfied. This situation falls in the class of errors-in-variables problems which in general involves not invertible model structures.

If we made our assumptions about the noise distribution, selected the model class and model order then we can perform the estimation and use a DP method to create a confidence region for the estimate. What remains is the validation of the obtained results. Here we can follow the same approach as we did for model selection, that is to rely on already established theory for validation. At this point we can divide validation methods into two categories. The first one contains
validation techniques that rely only on the estimate $\hat{\theta}$ and the data $(X,Y)$, the second category consists of techniques that involve the confidence region around the estimate as well. The use of DP methods has no impact on the outcome of validation techniques belonging to the first group. It remains to be checked whether DP confidence regions can be used or not in the case of methods from the second group. A simple example for such a rule is that if the confidence region for a parameter that represents additional model complexity (such as a time delay parameter) contains zero, than this parameter can be dropped to decrease the model complexity. For this rule it is not important weather the asymptotic or a DP confidence region is used to make the decision.

Some of the classical validation rules are intended to find the correct model order. So both low and high model orders will trigger an alarm. In case of modeling errors it is expected that the estimate $\hat{\theta}$ will not pass the validation step. Although this is good news, we would also like to comment on the statistical properties of DP confidence regions constructed in case of both under- and overmodeling. In case of undermodeling the confidence level of any set in the smaller parameter space is zero, but performing the DP confidence region generation will result in some set. There are no flags that can be raised inside the data perturbation framework that would signal an undermodeling situation, this has to be done using classical validation techniques.

In case of overmodeling the situation changes in the sense that the confidence level of the generated confidence sets will be correct. Though, we expect that overmodeling will generally alter the shape of the confidence region. This change is expected to be visible even if the generated confidence region is restricted to the correct model order, but at this point we do not have precise understanding of these effects.

### 2.7 Remarks and Contribution Summary

At the end of the chapter we would like to explicitly state our contribution to the topic and summarize the main conclusions that should be drawn.

The SPS method paved the way to the framework of data perturbation methods. Our main contributions are that we formulated these methods in a way
that the independence of the performance measure from the noise assumptions is clearly stated, and the precise conditions are identified that allow definition of such methods. We identified that the key condition for these methods is the group transformation invariance condition for the joint distribution of the noise. We have also shown that the noise distribution assumptions define how the perturbed datasets are generated, but they have no impact whatsoever on what performance measure can be used. This is formalized in Theorem 2.17. We have shown that although any performance measure can be used to have a desired confidence level, it is desirable to choose a meaningful one. Some performance measures that degenerate the hypothesis test to a coin toss are identified in Theorem 2.24. As shown in Section 2.4.2 the performance measure should be chosen carefully, even if it does not degrade the test to a coin toss, it should be examined thoroughly. We concluded that the theoretically most supported performance measure to be used is the prediction distance based measure. In case of dynamical systems we have to cope with the fact that the resulting confidence regions can be disconnected. This lead to the approximation by the principal connected component. We have also shown that the gradient based performance measure has significant computational advantages, so when approximations are done, this principal connected component of the gradient based methods can also be used.

We have emphasized that sufficiently exciting input is needed for these methods to work. One result in this direction is the statement about boundedness of linear regression confidence regions in Theorem 2.28. Another illustration for this point was that when DP methods are created for parameters of stochastic time series (no deterministic input at all), then DP methods degrade to a coin toss.

One of the important features of these methods that should be kept in mind is their random nature. A hypothesis test repeated on the same data might result in a different outcome. This is best illustrated by Figure 2.4.

In the context of system identification another aspect that we would like to mention is the proper applicability of these techniques to error in variables (EIV) problems. The most significant difficulty in this case is the fulfillment of Assumption 2.5 about the noise invertibility. The whole point of EIV problems is that there are usually twice as many noise variables compared to the number of ob-
servations that need to be considered. This makes it theoretically impossible to construct a DP method for EIV problems.

We would like to make a final remark about prediction distance based DP methods which concerns their invariance with respect to model parameterization. As the performance measure depends on the model through its predicted output, the actual parameterization of the model is not important. This means that singular model structures can be used (as long as there is an estimation algorithm that delivers $\hat{\theta}^{(i)}$) and these singular models are accepted consistently (when two models produce the same output, they are accepted or rejected together). This is not true for the gradient or estimate distance based performance measures.
Chapter 3

Globally Optimal Estimation of Polynomial Models

As we have seen in the previous chapters, optimization plays a pivotal role in system identification since almost every identification problem can be cast as an optimization problem. Additionally, we have seen that evaluation of the prediction distance or estimate distance based performance measures in data perturbation (DP) methods rely on solutions of perturbed identification problems. If these performance measures are defined using an estimate that is the solution of an optimization problem, the efficiency of such DP methods rely on the capability of delivering such estimates. Apart from the simplest cases, these optimization problems have multiple local minima. This makes estimation a challenging task, as a poor estimation procedure might return a local minimum as the estimate. For this reason, initialization of the identification tools is still an actively researched topic (e.g. [Ljung, 2003], [Tohme et al., 2007], [Tohme et al., 2009]).

In the context of DP methods it is interesting to note, that relying on a “poor” estimation routine (that does not provide the exact estimates, sometimes has numerical difficulties or it gets stuck in a local minimum) does not degrade the exact confidence level of the corresponding hypothesis test. A performance measure building upon such an estimation routine will satisfy the conditions of Theorem 2.17, as long as separate executions of the estimation routine are independent of each other. It is the shape of the confidence regions belonging to the test that will show the consequences of these imperfections.
Estimation procedures that return local optimum points of a cost function will result in asymptotic confidence regions that can be either smaller or larger than those created based on global minimizers. This is due to the fact that second order derivatives at a local minimum have no relation to the covariance of the estimates. Getting trapped in a local optimum can also cause bias or the loss of consistency.

A series of results in the optimization literature led to the fact that global solutions to polynomial optimization problems can now be approximated to arbitrary precision, or even obtained exactly. This chapter is devoted to exploring the usability of this methodology on the identification problem of generalized Box-Jenkins models using the least squares prediction error cost function. The conclusions drawn at the end will show that, although this method has potential and the identification problem is nicely aligned with the requirements of the method, its computational complexity restricts its applicability to very small scale problems. The usage of these methods appeared in the control literature earlier, a detailed survey is given in [Chesi, 2010], but they were not used for identification purposes until [7] and [Feng et al., 2010]. Publications related to the content of this chapter are [7], [8], [9], [10], [11], [6] and [12].

We will survey those results that are relevant to reaching the aforementioned conclusions in Section 3.1 by citing the relevant theorems and definitions from the literature. As the computational complexity of these methods is high, there were attempts to exploit sparsity in the problems to reduce it. One particular choice of such reduction, based on correlative sparsity, is also going to be presented. The least squares identification problem will be presented in Section 3.2 in two different forms that comply with the introduced notation. Both of these formulations will be evaluated from the perspective of the general solution in Section 3.3 and from the perspective of correlative sparsity in Section 3.4. The main result is going to be that the computational cost of these methods is exponential in the model complexity and linearly depends on the sample count. Numerical experiments are presented in Section 3.5.
3.1 Lasserre Hierarchies for Polynomial Optimization Problems

The class of polynomial optimization problems (POPs) consists of constrained optimization problems where the objective function is a polynomial, and the constraints are given in the form of polynomial inequalities.

\[ p^* = \min_{x \in K} p_0(x) \quad X^* = \{ x \in \mathbb{R}^n : p_0(x) = p^* \} \]

\[ K = \{ x \in \mathbb{R}^n : p_i(x) \geq 0, i \in [r] \} \]

The first thing that should be mentioned about this problem class is that it contains NP-complete problems, so we do not expect to have efficient algorithms for solving such problems in general. What we are going to present in this section is an approximation algorithm that relies on the solution of a series of nested SDP problems of increasing size. From the solutions of these SPDs one can obtain lower bounds for \( p^* \), and in certain cases it can be identified if the obtained lower bound actually equals \( p^* \). In such cases the set of optimum points \( X^* \) can also be identified.

The presented method was developed in a series of papers [Lasserre, 2001], [Lasserre, 2006], [Waki et al., 2006], [Lasserre, 2007], with the given algorithms having available implementations as well ([Waki et al., 2008], [Henrion et al., 2009]). In order to be able to present our analysis we will first give an introduction to this Lasserre hierarchy of SDPs. A thorough description of most of the relevant material can be found in [Laurent, 2009].

Let \( \mathbb{N}^n \) denote vectors of length \( n \) with non-negative integer entries and \( \mathbb{R}[n] \) denote the set of real valued polynomials in \( n \) variables. For \( \alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}^n \) and \( x \in \mathbb{R}^n \), the monomial \( x^\alpha \) should be understood as \( x_1^{\alpha_1} \cdots x_n^{\alpha_n} \). The total number of monomials in \( n \) variables of degree at most \( m \) is \( s(m) = \binom{m+n}{m} \). These \( s(m) \) monomials can be enumerated such that whenever \( \sum_{i=1}^n \alpha_i < \sum_{i=1}^n \beta_i \), then \( x^\alpha \) is before \( x^\beta \) in this enumeration. An \( m \) degree polynomial \( p \in \mathbb{R}[n] \) can be written as

\[ p(x) = \sum_{\alpha \in \mathbb{N}^n, \sum \alpha_i \leq m} p_{\alpha} x^\alpha \]
so the vector $p = \{p_\alpha\} \in \mathbb{R}^{s(m)}$ is the coefficient vector of polynomial $p$ in the monomial bases given by the previously mentioned enumeration.

The numerical solution to SDPs became efficiently reachable with the appearance of primal-dual path following interior point methods ([Nesterov and Nemirovskii, 1993]) that inherently rely on the dual of the optimization problem to be solved. The main idea behind the Lasserre hierarchy of SDPs is that instead of solving problem (3.1) some relaxation of it is solved, where the relaxation can be tuned with a relaxation order $m$. There are two relaxation approaches, both having an intuitive interpretation, and they both translate to SDPs that form a primal-dual pair. As we increase the relaxation order $m$ we get sharper approximations at the price of an increase in the size of the SDPs that need to be solved. These two relaxations are called the “sum of squares” and the “moment” relaxations, with detailed description in [Laurent, 2009]. For the purpose of the coming discussion it is enough to explain one of these, and we choose the moment based approach.

Instead of optimizing in the variable $x$, we search for a probability measure $\mu$, supported inside $K$ which minimizes the integral

$$p^* = \min_{\mu: \sup(\mu) \subseteq K} \int_{x \in \mathbb{R}^n} p(x) \, d\mu(x) = \min_{\mu: \sup(\mu) \subseteq K} \sum_{\alpha \in \mathbb{N}^n} p_\alpha \int_{x \in \mathbb{R}^n} x^\alpha \, d\mu(x)$$  

If we introduce the moment variables $y = \{y_\alpha\}$, such that $y_\alpha = \int_{x \in \mathbb{R}^n} x^\alpha \, d\mu(x)$ then the objective to be minimized becomes linear in these variables $p^T y$. What is left to be ensured is that the optimization considers only moment sequences $y$ that do have a representing measure $\mu$, moreover, that measure is supported on $K$.

### 3.1.1 Moment and Localizing Matrices

**Definition 3.1** (Moment matrix). Given a vector $y \in \mathbb{R}^{s(2m)}$, with first element $y\alpha = 1$, let $M_m(y) \in \mathbb{R}^{s(m) \times s(m)}$ denote its moment matrix. With the rows and columns indexed by the ordering of degrees, the $(\alpha, \beta)$ element of this matrix is defined as

$$M_m(y)(\alpha, \beta) = y_{\alpha + \beta} = \int_{x \in \mathbb{R}^n} x^{\alpha + \beta} \, d\mu(x)$$  

Using the definition of the moment matrix it is easy to see that for two polynomials \( p \) and \( q \) of degree at most \( m \), and \( M_m(\mu) \), the moment matrix of the moment sequence corresponding to measure \( \mu \), the quadratic form

\[
p^T M_m(\mu) q = \int_{x \in \mathbb{R}^n} p(x)q(x) \, d\mu(x)
\]

which immediately shows that moment matrices of positive signed measures are positive definite (as the quadratic form \( p^T M_m(\mu) p \) corresponds to integrating \( p^2(\cdot) \)).

**Definition 3.2** (Localizing matrix). Given a polynomial \( q \) with degree \( w \), represented by coefficients \( q \in \mathbb{R}^{s(w)} \), and a vector \( y \in \mathbb{R}^{s(m)} \), with first element \( y_0 = 1 \), let \( M_m^q(y) \) denote the \( s(\lfloor \frac{m-w}{2} \rfloor) \times s(\lfloor \frac{m-w}{2} \rfloor) \) localizing matrix created from \( y \), corresponding to polynomial \( q \). The \((\alpha, \beta)\) element of this matrix is defined as

\[
M_m^q(y)(\alpha, \beta) = \sum_{\gamma} q_\gamma y_{\alpha+\beta+\gamma}
\]

Using the definition of the moment matrix it is easy to see that for polynomials \( q \) of degree at most \( w \) and \( p \) of degree at most \( s(\lfloor \frac{m-w}{2} \rfloor) \), and \( M_m^q(\mu) \), the localizing matrix of the moment sequence corresponding to measure \( \mu \) and polynomial \( q \), the quadratic form

\[
p^T M_m^q(\mu) p = \int_{x \in \mathbb{R}^n} p^2(x)q(x) \, d\mu(x)
\]

showing that if the measure \( \mu \) is supported only on the set where \( q(x) \geq 0 \), then \( M_m^q(\mu) \geq 0 \) also holds.

The notes given after the previous definitions show that if the moment and localizing matrices are built using the moment sequence of a measure \( \mu \), up to order \( m \), then these matrices are positive semidefinite. It is important that, apart from some special cases in one dimension, these conditions are necessary but not sufficient for the reverse statement to hold. That is, given a vector \( y \), the semidefiniteness of these matrices (for any chosen order \( m \)) is necessary but not sufficient to ensure that there actually exists a probability measure \( \mu \) for which \( y \) is its moment sequence and and if it exists, its support lies within the positive levelset of the polynomials used in the localizing matrices.
3.1.2 Reformulation as SDP

As the polynomials are fixed in (3.1), using the new decision variables to be the elements of the moment sequence $y$, the matrices involved in the constraints depend linearly on $y$. The resulting optimization problem can be written as

$$
\mathcal{O}_m \rightarrow \min_{y \in \mathbb{R}^{s(2m)}} p_0^T y
$$

$$
M_m(y) \geq 0
$$

$$
M_{p_i}^m(y) \geq 0 \quad i \in [r]
$$

where the constraints are linear matrix inequalities and the objective function is also linear. This is an SDP in the dual form, and its primal pair could be interpreted in terms of sum of squares polynomials. There is a minimal relaxation order $m_{\text{min}}$ above which all appearing matrices are defined, and as we increase the relaxation order we get SDPs that are nested in the sense that if $y \in \mathbb{R}^{s(2m_1)}$ is a feasible point of $\mathcal{O}_{m_1}$, then it is also a feasible point of $\mathcal{O}_m$ for any $m_{\text{min}} \leq m \leq m_1$. This series of SDPs is called the Lasserre hierarchy and it has the property that, under some conditions on the involved polynomials $p_i$, $p^*_m$ converges to $p^*$ as $m \rightarrow \infty$ [Lasserre, 2001].

3.1.3 Convergence of the Hierarchy

**Assumption 3.3 (Representability assumption).** The set $K$ is compact and there exists a real-valued polynomial $u(x) : \mathbb{R}^n \rightarrow \mathbb{R}$ such that $\{x \in \mathbb{R}^n : u(x) \geq 0\}$ is compact and

$$
u(x) = u_0(x) + \sum_{i \in [r]} p_i(x) u_i(x) \quad \forall x \in \mathbb{R}^n \quad (3.9)$$

where the polynomials $u_i$ are all sum of squares polynomials for all $i = [r]$.

**Theorem 3.4.** If the optimization problem is defined as in (3.1) and Assumption 3.3 holds then

(a) as $m \rightarrow \infty$ we have $\inf \mathcal{O}_m \uparrow p^*$ and if $K$ has a nonempty interior, then there is no duality gap between $\mathcal{O}_m$ and its dual $\mathcal{O}_m^*$. **
(b) if \( p_0(x) - p^* \) can be written as

\[
p_0(x) - p^* = t_0(x) + \sum_{i \in [r]} p_i(x)t_i(x) \tag{3.10}
\]

for some polynomial \( t_0 \) of degree at most \( 2m \) and some polynomials \( t_i \) of degree at most \( 2m - \deg(p_i) \) such that all polynomials \( t_i \) are sum of squares for \( i = [r] \), then the duality gap is zero in the \( m \) order relaxation, \( p^\star = \min \mathcal{O}_m = \max \mathcal{O}_m^\star \).

It is important to emphasize that Theorem 3.4 provides only that the solutions of \( \mathcal{O}_m \) are lower bounds to \( p^\star \) and these lower bounds actually reach \( p^\star \) under some conditions. It does not guarantee that at the relaxation \( m \), where \( p^\star = \min \mathcal{O}_m \) is reached, \( X^\star \) can also be extracted from the obtained part of the moment sequence \( y \). Under some conditions on \( X^\star \), and maybe requiring the solution of higher order relaxations, the optimizer set can also be recovered [Laurent, 2004].

### 3.1.4 Sparsity

As it can be seen from the definition of \( \mathcal{O}_m \), the SDP to be solved involves as many linear matrix inequality (LMI) constraints as many constraints are in the original problem, and the size of these LMIs increases with \( m \) combinatorially through \( s(m) \). Different heuristics appeared trying to decrease the computational complexity of the method, by defining sparsity of a POP problem in a way that it will result in sparsity of the SDPs corresponding to the problem (see [Kim et al., 2005] and [Kojima et al., 2005]). These attempts resulted in the definition of correlative sparsity ([Waki et al., 2006]) and a sparse version of Theorem 3.4 was proved in [Lasserre, 2006]. We will discuss correlative sparsity in Section 3.4, but we give the sparse versions of \( \mathcal{O}_m \) here for completeness of the introduction.

**Definition 3.5** (Running intersection property). A collection \( \{I_1, \ldots, I_p\} \) of sets \( I_k \subseteq [n] \) has the running intersection property if \( [n] = \bigcup_{k=1}^p I_k \) and

\[
\forall k \in [p-1] : (\exists s \leq k : I_{k+1} \cap (\bigcup_{j=1}^k I_j) \subset I_s) \tag{3.11}
\]

is satisfied.
For a collection \( \{I_1, \ldots, I_p\} \) of sets \( I_k \subseteq \{1, \ldots, n\} \) such that \( \{1, \ldots, n\} = \bigcup_{k=1}^{p} I_k \) let us denote \( X(I_k) \) the subset of variables \( x_i \) for \( i \in I_k \).

**Assumption 3.6** (Separability assumption). The objective function \( p_0(X([n])) \) can be written as a sum of terms supported on the variables in \( X(I_k) \), that is

\[
p_0(X([n])) = \sum_{k=1}^{p} p_{0,k}(X(I_k)) \tag{3.12}
\]

and every constraint depends only on variables from exactly one of these sets

\[
\forall i \in [r] : \exists k_i : \left( \forall \alpha \in \{\beta : (p_i)_{\beta \neq 0}\} : \sum_{j \notin k_i} \alpha_j = 0 \right) \tag{3.13}
\]

**Definition 3.7** (Reduced moment and localizing matrix). Let \( \mu \) be a probability measure over \( \mathbb{R}^n \), with moment sequence \( y_{\mu} \) and let \( I \subseteq \{1, \ldots, n\} \) be an arbitrary index set. The reduced moment matrix \( M_m(y_{\mu}, I) \) and reduced localizing matrix \( M^0_m(y_{\mu}, I) \) are the submatrices of \( M_m(y_{\mu}) \) and \( M^0_m(y_{\mu}) \) respectively, obtained such that the \((\alpha, \beta)\) entry of the matrix is preserved if \( \text{sup}(\alpha) \subseteq I \) and \( \text{sup}(\beta) \subseteq I \).

We note that a localizing matrix corresponding to polynomial \( q \) can only be reduced to an index set \( I \subseteq \{1, \ldots, n\} \) if \( q \) depends only on variables from \( X[I] \).

If a collection \( \{I_1, \ldots, I_p\} \) of sets is available such that Assumption 3.6 holds for (3.1) then a sparse version of (3.8) can be defined as

\[
p^*_{m,I} = \min_{y \in \mathbb{R}^{|I|}} p^0_I y
\]

\[
\exists k \in [p] : M_m(y, I_k) \geq 0
\]

\[
\forall k \in [p] : M^0_m(y, I_k) \geq 0 \quad i \in \{j : p_i \text{ depends only on } X[I_k]\} \tag{3.14}
\]

The difference between \( S_m \) and \( O_m \) is that instead of one moment matrix LMI of size \( \left( \begin{array}{c} m+n \cr m \end{array} \right) \), we have \( p \) smaller ones of size \( \left( \begin{array}{c} m+|I| \cr m \end{array} \right) \) and every localizing LMI of size \( \left( \begin{array}{c} \frac{m-w_i}{2}+n \cr \frac{m-w_i}{2} \end{array} \right) \) is replaced with a smaller one of size \( \left( \begin{array}{c} \frac{m-w_i}{2} + |I| \cr \frac{m-w_i}{2} \end{array} \right) \). If the collection \( \{I_1, \ldots, I_p\} \), fulfilling Assumption 3.6 can be found, such that \( |I_k| \ll n \) then the size of the SDP in \( S_m \) is much smaller than that in \( O_m \). The main result in [Lasserre, 2006] is that if the index sets \( I_k \) have the running intersection property,
and some additional technical conditions are satisfied, then a sparse version of Theorem 3.4 holds with similar statements.

3.2 The Identification Problem as POP

Identification problems, when formulated as an optimization problem, can be given with many different cost functions. The least squares cost function defined in (1.18) is one of these, but prediction error methods can be constructed using different ones as well. I.e. when for prediction errors \( \varepsilon \) that are large, only \( |\varepsilon| \) is penalized instead of using \( \varepsilon^2 \). As Lasserre hierarchies can be applied only to purely polynomial problems, the analysis in this chapter is restricted to the least squares cost function. Maximum likelihood identification can lead to other polynomial cost functions in case of some special noise distributions. In these cases the analysis presented in the following sections has to be revised by taking into account the actual cost function at hand.

Although we have settled on the least squares cost function, for a given dataset \((Y,U)\) the identification cost function can be represented by different POPs. The parametrization of the model can be chosen in many different ways, but it also offers choices how the model structure is represented in the problem. We present two of the standard choices in this section and their analysis is going to be presented throughout the chapter; other choices will only appear on the level of comments.

3.2.1 Parameterization

3.2.1.1 Monic Parameterization

We are going to analyze the standard parameterization of extended Box-Jenkins models introduced in Section 1.2.1

\[
y[k] = \frac{1}{A(z)} \left( \frac{B(z)}{F(z)} u[k] + \frac{C(z)}{D(z)} e[k] \right)
\]  

(3.15)

where the initial conditions for systems \( 1/A(z) \), \( B(z)/F(z) \) and \( C(z)/D(z) \) are denoted by \( x_{A}, x_{B} \) and \( x_{C} \). For each of these vectors we are also introducing
the corresponding polynomials $X(z) = x_1z^{-1} + \cdots + x_nz^{-1}$ where $x_i$ is the \(i\)-th element of $x$.

At this point we have to make the choice of how to parameterize the model. We are going to focus on the choice where $A$, $F$, $C$ and $D$ are monic, $b_0 = 0$ and the noise model $C(z)/D(z)$ is written in the form $\left(1 + \tilde{D}(z)/C(z)\right)^{-1}$. The parameter vector $\theta$ is composed of the unknown coefficients of the polynomials $A, B, F, C, \tilde{D}$ and the elements of the initial conditions. We will refer to this as the monic parameterization when comparing with other choices.

Normalizing the parameter space, by choosing monic polynomials, is needed in order to eliminate manifolds in the parameter space that correspond to the same model and differ only in parameterization. This normalization is not sufficient to ensure that the identification cost function will have a single global minimum. If $z$ is changed to $z^{-1}$ in the noise model then inverting the (infinite) prediction error sequence will result in the same least squares error, thus providing a second global minimizer.

### 3.2.1.2 Norm Constrained Parameterization

Another possible way of parameterizing the model would be to treat all polynomial coefficients as unknown and constrain the model set in this space. This leads to a model formulation, a bit different from (3.15),

$$y[k] = \frac{a_0}{A(z)} \left( \frac{B(z)}{F(z)} u[k] + \frac{C(z)}{D(z)} c[k] \right) \quad (3.16)$$

where all polynomial coefficients (including $a_0$, $c_0$, $f_0$, $d_0$) are unknown and $c_0$ and $d_0$ are represented with the same decision variable. The model set is normalized such that $\|a\|^2 = 1$, $\|b\|^2 + \|f\|^2 = 1$ and $\|c\|^2 + \|d\|^2 = 1$. We will refer to this as the norm constrained parameterization. In some respect this parameterization has advantages over the monic choice, from the Lasserre hierarchies point of view, but we cannot use this to get better performance out of the method. For this reason we are going to concentrate on the monic parameterization in the sequel; nonetheless, we are going to comment on the differences.
3.2.2 Reformulations of the LS Optimization Problem

We are going to formulate the LS optimization problem in two different ways. In the first case the cost function $V_d$ is going to be expressed in terms of the measurement data and problem parameters directly; while in the second case $V_d$ is defined using auxiliary variables $\varepsilon[k]$ and model consistency is going to be ensured using equality constraints in the form of difference equations (1.2).

3.2.2.1 Unconstrained Formulation

The direct formulation of the least squares cost function can be deduced simply from one of the canonical state space models of the involved systems.

**Definition 3.8 (Observable canonical realization).** The observable canonical realization of an impulse transfer function

\[
\frac{N(z)}{D(z)} = \frac{n_1z^{-1} + \cdots + n_mz^{-m}}{1 + d_1z^{-1} + \cdots + d_mz^{-m}}
\]  

is given by the matrices

\[
A = \begin{bmatrix}
-d_1 & 1 & 0 & \cdots & 0 \\
0 & 1 & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & 1 & 0 \\
\vdots & & \ddots & \ddots & \ddots \\
-d_m & 0 & \cdots & \cdots & 0
\end{bmatrix}, \quad c^T = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix}, \quad b = \begin{bmatrix} n_1 \\ 0 \\ \vdots \\ \vdots \\ n_m \end{bmatrix}, \quad d = 0
\]  

When we refer to the extended observability matrix $\Gamma_N$ or to the matrix with the Markov parameters $H_N$ of a system $\frac{N(z)}{D(z)}$ we implicitly consider these matrices as defined by (1.14) and (1.15) respectively, based on the observable canonical realization of $\frac{N(z)}{D(z)}$. When we use this notation with systems where $n_0 = 1$, then $d = 1$ should be considered and the other three components of the state space description should be constructed with the remainder nominator $\tilde{N} = N - D$. We will only use this notation in the monic parameterization, so the objects will be well defined in all cases.
Using the notation introduced above, the direct cost function can be written as

\[ V_d(\theta) = \frac{1}{2N} \left\| H_{v}^{-1} \left[ H_{\lambda}^{-1} \left( Y - \Gamma_{\lambda} x_{\lambda} \right) - \Gamma_{v} x_{v} - H_{v} U - \Gamma_{\lambda} x_{\lambda} \right] \right\|^2 \] (3.19)

Because of the matrix inversion this is not yet a finite degree polynomial of the decision variables, but using the fact that \( H_{v}^{-1} = H_{v} \) and \( H_{\lambda}^{-1} = H_{\lambda} \) this can be written as

\[ V_d(\theta) = \frac{1}{2N} \left\| H_{v} \left[ H_{\lambda} \left( Y - \Gamma_{\lambda} x_{\lambda} \right) - \Gamma_{v} x_{v} - H_{v} U - \Gamma_{\lambda} x_{\lambda} \right] \right\|^2 \] (3.20)

which is a finite degree polynomial with variables corresponding to the elements of the monic parameterization and the coefficients of the polynomial depend on the measurement data \((Y, U)\). It is important to note that it is a special property of the monic parameterization that the least squares cost function can be expressed as a polynomial involving only positive degrees of the decision variables. This is not the case for the norm constrained parameterization.

### 3.2.2.2 Constrained Formulation

The other possible way of formulating the estimation problem is through the use of auxiliary variables \( \varepsilon[k] \) that play the role of prediction errors, and model consistency is kept by constraining the optimization space. The constraints come from the difference equation (1.2). The output of the generalized Box-Jenkins model can be given using impulse transfer functions as

\[ y[k] = \frac{X_A(z)}{A(z)} \delta[0] + \frac{1}{A(z)} \left( \frac{X_F(z)}{F(z)} \delta[0] + \frac{B(z)}{F(z)} u[k] + \frac{X_D(z)}{D(z)} \delta[0] + \frac{C(z)}{D(z)} \varepsilon[k] \right) \]

Multiplying with the common denominator and using the decision variables \( \varepsilon[k] \) in the place of the noise variables \( e[k] \), model consistency is ensured by the equality constraints

\[ A(z) F(z) D(z) y[k] = F(z) D(z) A(z) \delta[0] + \] (3.21)
+ \mathcal{D}(z) (\mathcal{X}_F(z) \delta[0] + B(z) u[k]) + \mathcal{F}(z) (\mathcal{X}_D(z) \delta[0] + C(z) \varepsilon[k])

These constraints are polynomials in the model and initial condition variables of the monic parameterization, but a similar difference equation can be given also if the norm constrained parameterization is used. The least squares optimization problem can be given as the minimizer of the cost function

\[ V_a(\theta, \varepsilon) = \frac{1}{2N} \sum_{k=1}^{N} \varepsilon^2[k] \] (3.22)

under constraints (3.21) for \( k \in [N] \).

The rest of the chapter contains the analysis of these two optimization problems from the perspective of Lasserre hierarchies.

### 3.3 General Method and the Identification POPs

In order to understand the properties of the Lasserre hierarchies in the context of the identification POPs we are going to examine two questions. One of these questions is about the convergence properties of the relaxations \( \mathcal{O}_m \), as defined in (3.8), to see if the conditions for convergence are fulfilled or not. The other aspect that we examine is the size of the SDPs appearing in \( \mathcal{O}_m \), which is determined by the number of variables, the degree of the cost function and that of the constraints.

#### 3.3.1 Convergence of Relaxations

##### 3.3.1.1 Additional Constraints for Convergence

Assumption 3.3 needs to be fulfilled by the optimization problem in order to have convergence of the solutions of \( \mathcal{O}_m \) to the global minimum of \( V_d(\theta) \) or \( V_a(\theta, \varepsilon) \). In order to fulfill this assumption, the feasible set \( K \) of the optimization problem needs to be compact, which implies that it should be closed and bounded.

In case of the monic parameterization and cost function \( V_d(\theta) \) we see that the optimization space is in fact \( \mathbb{R}^{n_\theta} \). Although the coefficients of the denominator can be constrained into a compact set by imposing constraints such as stability of the model, this cannot be done for the initial condition variables and the coefficients
in $\mathcal{B}$. Looking at the cost function $V_a(\theta, \varepsilon)$, it is also clear that the optimization space is not bounded. This is due to the same reasons with the model parameters as before, and the prediction error variables $\varepsilon[k]$ cannot be bounded either.

The result of these considerations is that the optimization problems corresponding to identification problems cannot be tackled as they are with the Lasserre hierarchy approach. Additional constraints will be added to the formulated optimization problems to ensure convergence relying on available results in the literature. We cannot prove that these additional constraint are needed, in fact it could happen that for the specific class of problems that we are targeting these are not necessary.

The question, how to relax the compactness assumption of $K$, was in the focus of intensive research ([De Klerk and Laurent, 2011], [Jeyakumar et al., 2014]), but it seems that boundedness of $K$ cannot be relaxed.

Least squares problems are special in the sense that their cost function is the sum of polynomials with only even degrees. This cannot be exploited to loosen the conditions for convergence [Kim and Kojima, 2010]. This means that in order to have convergence of the solutions of $\mathcal{O}_m$ to $v_d^*$ and $v_a^*$, the optimization space needs to be bounded. Adding a constraint $\|\theta\|^2 \leq R^2$ ensures that Assumption 3.3 is satisfied, however $R$ should be chosen such that the global optimizer should be inside the ball with radius $R$.

Adding this extra constraint may seem to be just a technical tweak to adhere to the conditions of Theorem 3.4, but this is not the case. The necessity of this additional constraint can be illustrated easily on the unconstrained optimization of $V_d(\theta)$. Although $V_d(\theta)$ is a sum of squares (SOS) polynomial, this cannot be stated about $V_d(\theta) - v_d^*$ is general. When $V_d(\theta) - v_d^*$ is not SOS, then the optimal value $p_m^*$ of the SDP hierarchy $\mathcal{O}_m$ corresponding to the problem will stop to increase after some relaxations in the sense that $p_m^* = p_{\text{max}}^*$ for all $m \geq m_{\text{max}}$, such that $p_{\text{max}}^* < v_d^*$. This means that there is a duality gap between $\mathcal{O}_m$ and its dual $\mathcal{O}_m^*$. In such cases extracting information about the minimizers from the moment sequence candidate $\mathbf{y}^*$ is difficult. Adding the extra norm constraint eliminates this phenomenon that solving larger and larger SDPs does not offer more information.
3.3.1.2 Rate of Convergence

An interesting fact is that when the constraints are not defining a convex set $K$, then finite convergence is not expected ([Helton and Nie, 2010], [De Klerk and Laurent, 2011]). The fact that the solution of $\mathcal{O}_m$ does not reach $p^*$ for a finite value of $m$ is not a real problem in practice, as numerical errors would appear anyhow. What is more important is the rate of convergence of $p^*_m$ to $p^*$. There is not much known about this convergence rate, but it is known to depend on the problem size (the degree of the polynomials and the number of variables). More interestingly, it also depends on the optimal value of the cost function $p^*$, such that the convergence rate exponentially depends on $1/p^*$ ([Nie and Schweighofer, 2007]).

It is easy to see that the constrained optimization formulation with the prediction error variables $\varepsilon[k]$ has a non-convex set $K$ defined by the equality constraints that ensure model consistency. For this reason we don’t expect finite convergence of the Lasserre hierarchies.

3.3.1.3 Extracting Optimizers

The last note that we want to make before turning our attention to the size of the SDPs related to the identification POPs concerns extraction of the system parameter estimates. Detecting if finite convergence occurred is possible in cases where there are finitely many global minimizers to the problem ([Curto and Fialkow, 2000]), which is true for all of the presented identification POPs if the measurement data is not degenerate. The algorithm for extracting these minimizers is given in [Henrion and Lasserre, 2005]. This algorithm can be executed on moment sequence candidates $y^*$ obtained in any relaxation $m$ to extract one minimizer $\theta^*_m$. If there are only finitely many global minimizers of the identification problem, then $\theta^*_m$ will converge to one of these. This allows extraction of useful information from the solution of $\mathcal{O}_m$, even if the relaxation order was not high enough to reach the global minimum.
3.3.2 Size of the SDPs

To be able to discuss the complexity of the SDPs appearing in $\mathcal{O}_m$ let us introduce some notation. Let the indicator variable $1_P$ show whether the polynomial $P$ contains unknowns or not

$$1_P = 1_P(\theta) = \begin{cases} 1, & \text{the polynomial } P \text{ contains unknowns} \\ 0, & \text{otherwise} \end{cases} \tag{3.23}$$

where we will not mark the dependence on $\theta$ explicitly. We define the degree of a matrix $M$ with polynomial entries to be the maximum of the degrees of its elements

$$\deg(M) = \max_{i,j}\{\deg(M(i,j))\} \tag{3.24}$$

3.3.2.1 Unconstrained Formulation

The degrees of different terms appearing in the cost function $V_d(\theta)$ can be calculated as

$$\deg\left(H^{\mathcal{N}}\right) = 1_{\mathcal{N}} + (N-2)1_{\mathcal{D}} \quad \deg\left(\Gamma^{\mathcal{N}}\right) = (N-1)1_{\mathcal{D}}$$

In general the SDP based solution in the $m$-th order relaxation involves at least an LMI of size $\binom{n_\theta + m}{m}$, where $m$ needs to be greater than $\frac{1}{2}\deg(V_d)$.

**Statement 3.9 (Unconstrained general SDP sizes).** The degree of the least squares cost function $\deg(V_d)$ can be calculated as

$$\frac{1}{2}\deg(V_d(\theta)) = 1_{\mathcal{C}} + (N-2)1_{\mathcal{D}} + \max\left\{1_{\mathcal{A}} + 1_{\mathcal{X}}^1 (1 + (N-1)1_{\mathcal{A}}) \right. \\
\left. 1_{\mathcal{X}}^2 (1 + (N-1)1_{\mathcal{F}}) , \\
1_{\mathcal{B}} + (N-2)1_{\mathcal{F}} , \\
1_{\mathcal{X}}^3 (1 + (N-1)1_{\mathcal{D}}) \right\}$$

Depending on the model structure, the size of the SDP corresponding to the $m$ order relaxation of the optimization problem

$$\hat{\theta} = \arg\min V_d(\theta)$$
can be calculated based on the number of unknowns and the degree of the polynomial cost function. There are two, full sized LMI constraints in the SDP formulation. Based on model structure complexity there are three different classes of problem difficulty.

1. **Model complexity related difficulty class:** it is characterized by \( n_f = n_d = 0 \) and no initial condition variables. The minimal relaxation order is \( m_{\text{min}} = 1 \). The model complexity is \( n_\theta = n_a + n_b + n_c \). The combinatorial factor of the SDP size growth depends only on \( n_\theta \).

2. **Model complexity and sample count related difficulty class:** Contains problems where the minimal relaxation order depends on \( 1N \). This class covers models with \( D = 1 \). The minimal relaxation order increases with the sample count \( N \).

3. **Model complexity and double sample count related difficulty class:** Contains problems where the minimal relaxation order depends on \( 2N \). Parameter combinations not fitting into the previous two categories belong here. The minimal relaxation order increases with the double of the sample count \( 2N \).

Since the degree of the polynomial determines the minimal relaxation order \( m \) from which \( \mathcal{O}_m \) is a proper relaxation of the identification problem, the categorization of model structures should be based on that degree. The three categories identified in Statement 3.9 align nicely with the more general difficulty based categorization of these problems if we consider methods tailored to a specific model structure.

We would like to emphasize that the difference between the second and the third model category in Statement 3.9 is significant. Since the size of the SDP relaxations depends combinatorially on the degree, the difference between \( 1N \) and \( 2N \) is an exponential jump in the SDP size.
3.3.2.2 Constrained Formulation

A statement, similar to Statement 3.9, can be made considering the optimization problem formulation with auxiliary variables, where the cost function is \( V_a(\theta, \varepsilon) \) and the constraints are equality constraints of the form (3.21).

\[
\hat{\theta} = \text{argmin } V_a(\theta, \varepsilon) = \frac{1}{2N} \sum_{k=1}^{N} \varepsilon^2[k] \\
A(z)F(z)D(z)y[k] = F(z)D(z)X_A(z)\delta[0] + \\
+D(z)(X_F(z)\delta[0] + B(z)u[k]) + F(z)(X_D(z)\delta[0] + C(z)\varepsilon[k]) \quad \forall k \in [N] \\
\|\theta\|^2 + \|\varepsilon\|^2 \leq R^2
\]

This is in an optimization problem with quadratic cost function, polynomial constraints, and variables whose number is linearly increasing with the sample count \( N \), similarly to the number of constraints.

**Statement 3.10** (Constrained general SDP sizes). *Depending on the model structure, the size of the SDP corresponding to the \( m \) order relaxation of the optimization problem (3.25) can be calculated based on the degree of the equality constraints (both polynomial coefficients, initial condition and error variables should be counted). There are \( 2N + 2 \), full sized LMI constraints in the formulated SDPs.*

Based on model structure complexity there are three different classes of problem difficulty.

1. **Linear constraints:** The minimal relaxation order \( m_{\text{min}} = 1 \).
2. **Quadratic constraints:** The minimal relaxation order \( m_{\text{min}} = 2 \).
3. **Cubic constraints:** The minimal relaxation order \( m_{\text{min}} = 3 \).

In every case the number of variables in the optimization problem is \( n_\theta + N \). This results in SDP sizes exponentially growing both in the sample count \( N \) and the relaxation order \( m \).

Apart from the simplest problem (that is in fact linear regression) both approaches rely on solving SDPs with sizes that exponentially depend on the sample
count $N$. This limits the applicability of these SDP relaxations to very simple systems and extremely short data records. It depends on the model structure which approach can be started with a smaller SDP size, but neither of these approaches can be applied in practice due their exponential complexity in the sample count.

### 3.4 Sparsity of the Identification POPs

As we have seen in the previous section, the general framework offered by the Lasserre hierarchies is not applicable to identification problems in practice due to the fact that the size of the resulting SDPs depends exponentially on the sample count. To be able to reduce the size of the SDP sequence we are going to rely on the sparsity of the problems in the sense of correlative sparsity as defined in [Waki et al., 2006].

#### 3.4.1 Correlative Sparsity

The correlative sparsity (CSP) graph of a POP is defined in a way that the nodes of the graph correspond to the decision variables of the POP. Two nodes are connected if they appear in one monomial term of the cost function or if they appear in the same constraint polynomial.

**Definition 3.11** (Chordal graph). A simple graph with vertex set $V$ and edge set $E$ is chordal if every circle $C$ on the vertices $V_C$ using edges $E_C$, that is not a triangle, has a chord in the graph, i.e. $\exists v_1, v_2 \in V_C$ such that $(v_1, v_2) \notin E_C$ but $(v_1, v_2) \in E$.

As it is shown in [Blair and Peyton, 1993], the maximal cliques of a chordal graph satisfy Assumption 3.6. One way to find the collection of sets $\{I_1, \ldots, I_p\}$ that satisfy Assumption 3.6 is to find a chordal extension of the CSP graph and use the maximal cliques in the extension. This is the procedure suggested in [Waki et al., 2006], with the main advantage that it can be automated for every problem.

We are going to show in this section that the chordal extension of the CSP graph belonging to the directly formulated POP minimizing $V_d(\theta)$ is a fully connected graph, while the CSP graph of the POP formulation with equality constraints
can be extended to a chordal graph that has approximately $N$ maximal cliques of constant size. As a result we will be able to show that for a fixed model complexity the sparse relaxations $S_m$ will still grow combinatorially with the model complexity, but only linearly with respect to the sample count $N$.

### 3.4.2 Unconstrained Formulation

We formulate a statement about the sparsity of the unconstrained formulation of the identification POP.

**Statement 3.12** (Sparsity of the unconstrained optimization problem). *If the sample count is greater than the number of identifiable parameters, then the CSP graph of the unconstrained optimization problem is fully connected, thus there is no sparsity in the problem.*

The previous statement is the direct consequence of the fact that there is going to be a monomial in the expansion of $(x_1 + \cdots + x_n)^k$ that contains $x_1\cdots x_n$ if $k \geq n$. Since the unconstrained optimization problem used the minimal number of decision variables, it was not expected that there is any sparsity left in the problem.

### 3.4.3 Constrained Formulation

In order to reveal the CSP structure of the identification POP with equality constraints, we are going to consider (3.25) without the norm constraint $\|\theta\|^2 + \|\varepsilon\|^2 \leq R^2$. It would make the CSP graph a fully connected graph and it is not an integral part of the problem formulation.

The schematic structure of the minimal chordal extension of the CSP graph is shown in Figure 3.1. The unknown coefficients of the model polynomials and the initial conditions form a fully connected subgraph. Note that this is not true in the CSP graph, as initial condition variables from the same system might not appear in one constraint. We will refer to this clique as the core of the graph. Each of the $N$ equality constraints connect at most $n_C + n_F$ consecutive noise variables to the core (we note again that the edges to the initial condition variables are extensions compared to the CSP graph). This shows that selecting $p = N - n_C - n_F + 1$, the
collection of sets \( \{I_1, \ldots, I_p\} \) can be defined as

\[
I_k = C \cup \{\varepsilon[k], \ldots, \varepsilon[k + n_C + n_F - 1]\} \quad k \in [N - n_C - n_F + 1] \\
C = \{A, X_A, B, C, D, X_D, F, X_D\}
\]  

(3.26)

It is not difficult to see that the graph defined as the union of fully connected graphs over the sets \( I_k \) is an extension of the CSP graph of the original problem. It is also easy to check that the identification POP together with this collection of sets \( \{I_1, \ldots, I_p\} \) satisfies Assumption 3.6. Let \(|I|\) denote the size of each set \( I_k \), as they are of the same size. Replacing the one global norm constraint in (3.25) with \( p \) localized constraints of the form

\[
\|\theta_C\|^2 + \|\varepsilon_{I_k}\|^2 \leq R_k \quad k \in [p]
\]  

(3.27)

allows definition of sparse Lasserre hierarchies \( \mathbb{S}_m \), such that convergence of the solutions to the global minimum is guaranteed according to Theorem 3.6 in [Lasserre, 2006]. \( \theta_C \) and \( \varepsilon_{I_k} \) denote the parts of the parameter vector belonging to the model parameters in the clique \( C \) and to the prediction error variables in set \( I_k \) as defined in (3.26)

\[
\hat{\theta} = \arg\min V_a(\theta, \varepsilon) = \frac{1}{2N} \sum_{k=1}^{N} \varepsilon^2[k] \\
\mathcal{A}(z)\mathcal{F}(z)\mathcal{D}(z)y[k] = \mathcal{F}(z)\mathcal{D}(z)\mathcal{X}_A(z)\delta[0]+
\]  

(3.28)
\[ + \mathcal{D}(z) (\mathcal{X}_\mathcal{F}(z) \delta[0] + \mathcal{B}(z) u[k]) + \mathcal{F}(z) (\mathcal{X}_\mathcal{D}(z) \delta[0] + \mathcal{C}(z) \varepsilon[k]) \quad \forall k \in [N] \]
\[ \| \theta \|_C^2 + \| \varepsilon \|_I^2 \leq R_k \quad k \in [p] \]

**Statement 3.13** (Constrained sparse SDP sizes). Depending on the model structure, the size of the SDP corresponding to the m order relaxation of the optimization problem (3.28) can be calculated based on the degree of the equality constraints (both polynomial coefficients, initial condition and error variables should be counted). There are \( 2N + 2p \) reduced size sized LMI constraints in the formulated SDPs, where the size of each LMI is \( \binom{|I|+m}{m} \).

Based on model structure complexity there are three different classes of problem difficulty, this classification is the same as in Statement 3.10.

To compare Statements 3.10 and 3.13, in the first case \( 2N \) LMI constraints of size \( \binom{n_\theta+N+m}{m} \), while in the second case \( 2N + 2p \approx 4N \) LMI constraints of size \( \binom{n_\theta+|I|+m}{m} \) appear. This shows that if the identification problem is formulated with equality constraints as minimization of the cost function \( V_a(\theta, \varepsilon) \), with the sparse norm constraints (3.27), then the size of the SDPs to be solved in the relaxations \( S_m \) no longer depends combinatorially on the sample count \( N \).

### 3.5 Numerical Experiments

We illustrate the capabilities and limitations of the sparse Lasserre hierarchy corresponding to problem (3.28) on a sample output error system. The plant model is the sampled equivalent of the continuous transfer function \( G(s) = \frac{0.8s}{s^2+1.2s+1} \) sampled with sampling time \( T_s = 1 \). The input is a step at time 0 and the output noise is zero mean white Gaussian noise with standard deviation 0.05. 20 samples are used for the identification and zero initial condition is assumed to be known.

The parameters of this experiment are chosen such that the sparse Lasserre SDPs are still solvable up to a few relaxations and also the problematic nature of the estimation problem can be illustrated by the fact that the system identification toolbox in Matlab returns locally optimal results.

The 50 experiments differ only in the noise realization contaminating the measurements. In every experiment we calculated estimates in three different ways. The Matlab oe routine was parameterized to obtain an estimate, and the least
Chapter 3. Globally Optimal Estimation of Polynomial Models

Figure 3.2: Difference between cost function values obtained using oe and $S_3$ squares error corresponding to this estimate is denoted by $oe^*$. The third sparse relaxation $S_3$ was solved, and the corresponding estimate was approximated with the first order moments of the obtained $y^*$ moment sequence, the least squares error corresponding to this estimate is denoted by $S_3^*$. The Matlab oe routine was executed again with specified starting values obtained from $S_3$, the least squares error corresponding to this estimate is denoted by $oe^*_{S_3}$. The results of the experiments are summarized in Figure 3.2. When reading these figures it should be kept in mind that the variance of the added noise was $5 \cdot 10^{-4}$, so asymptotically we expect the optimum value of the cost function to be around $5 \cdot 10^{-4}$.

It can be seen in Figure 3.2a that, due to the bad signal to noise ratio, the LS values corresponding to the auto initialized oe estimates are off by orders of magnitude, and in these instances the oe routine got stuck in a local minimum. We would like to mention that in these figures $S_3^*$ is not the lower limit for the cost function obtained during the relaxation, but the LS value corresponding to an estimate extracted from the moment sequence. This is the reason why $oe^* - S_3^* < 0$ is possible. Figure 3.2b shows how much the oe routine could improve on the estimate obtained from $S_3$. Since $S_3$ did not converge, it was expected that improvement is still possible and considerable improvement is achieved in almost half of the experiments. It is an implementation problem of the Matlab oe routine that it can return estimates that are worse than the given initial estimate. This illustrates, that although the first relaxations are not able to provide very good estimates, they can provide good initial values for the plant models. It is important to note that this statement is only true to the plant models because their value is expected to be unique. Since there are two global minimizers corresponding to the 'flipped
time’ noise models, the first order moments of the noise model variables in the moment sequence $y^*$ will not be good estimates of the corresponding parameters.

In order to illustrate the computational complexity of these problems we mention that the average execution time for the $S_2$ approximations was around 7 seconds, while solving $S_3$ took almost eleven minutes on average. Solving $S_4$ was not possible on a 32 bit machine due to memory limitations. It is a straightforward calculation to compare the size growth in the SDPs to see what happens if the model order or the relaxation order is increased. The conclusion can be drawn that for low relaxation orders (we cannot really go above $S_4$) it is cheaper to increase model complexity. Indeed $S_3$ could be solved for higher order models as well as for larger sample counts.

### 3.6 Concluding remarks

We have discussed the applicability of Lasserre hierarchies for system identification and we draw the following conclusions.

Due to the fact that this approach transforms general NP-hard problems into a series of SDPs, and SDPs can be solved in polynomial time of their size, it is expected that the size of the converted SDPs depends exponentially on the original problem size. We were able to show that the sample count dependence of the converted SDPs can be kept linear, this is done using the equality constrained formulation of the LS identification problem and exploiting its sparsity pattern. The size growth remains exponential in the model complexity parameters.

Discovering the actual computational time and memory requirements of solving the SDPs of the Lasserre hierarchy showed a rather consistent behavior that we discuss now. Second and third order relaxations can be solved in general for moderately sized problems. Solving the fourth order relaxations is practically impossible on 32 bit machines, due to memory limitations (2GB on Windows). One extra meaningful relaxation order can be obtained using 64 bit machines. Solving the fifth order relaxation runs into numerical issues and it is not possible to extract better estimates from it compared to previous relaxations.

We think that these methods can prove useful for finding initial estimates for local optimization algorithms in low SNR problems. As we have seen in Sec-
tion 3.5, the estimates extracted from the third relaxation sometimes lead to bet-
ter estimates than the built-in initialization algorithm of the System Identification
Toolbox.

To summarize, the main conclusion that we can draw is that the Lasserre
hierarchy can be used to generate initial estimates in small scale problems, but its
computational complexity makes it unfit for larger scale problems. Also solving
enough relaxations to get close to convergence to the global solution is out of reach
with the currently available computational power.
Chapter 4

Effect of Finite Measurement Length on Uncertainty Estimation

As it was illustrated in Section 1.3, the asymptotic confidence regions, constructed based on the Gaussian limiting distribution of the parameter estimates, are not accurate when the sample count is small. A more detailed discussion about the reasons of this deficiency and of cases when these confidence regions are reliable can be found in [Garatti et al., 2004].

One way to address this issue is to avoid the use of asymptotic theory completely and such an approach was given in Chapter 2. Beside the papers mentioned in that chapter there exist other possibilities as well to quantify uncertainty estimates relying on finite samples. The reason why these were not cited in Chapter 2 is that these results follow a different line of thinking, namely, they do rely on the asymptotic methodology. The common factor of these methods is that they arrive to a hypothesis test that boils down to checking coverage by an ellipsoid. In the limiting case, as $N \to \infty$, these methods result in the same ellipsoids, but their finite sample behavior is different from the classical asymptotic theory. [Douma and Van den Hof, 2005] and [Douma and Van den Hof, 2006] defines exact hypothesis tests using the classical asymptotic theory, and the confidence level is kept exact by using special filtering of the input and output. The validity of asymptotic confidence regions, along with definition of other confidence regions based on likelihood ratio tests is given in [Quinn et al., 2005]. Comparison of
the finite sample properties of different asymptotically equivalent methods can be found in [Dekker et al., 2007] and [Dekker et al., 2008].

The classical ellipsoidal confidence regions are based on the assumption that the distribution of the estimates is Gaussian, and an estimate of the corresponding covariance matrix is available. It is a byproduct of the Gaussian distribution that these ellipsoids are also the tightest confidence regions that can be constructed for any confidence level. Even if the distribution of the estimates is not Gaussian, one can still aim for ellipsoidal confidence regions. The axes directions of these ellipsoids can be selected based on the estimated covariance of the estimates. What is left to be chosen is the length of these axes.

This idea of using ellipsoidal sets to reduce computational complexity appears in other identification approaches as well. Outer ellipsoid approximations for the SPS method are defined in [Weyer et al., 2013]. The same idea is behind optimal bounding ellipsoids algorithms [Joachim and Deller, 2006] which also appears in the case of set membership identification methods [Cheung et al., 1991].

The main goal of this chapter is to quantify the reliability of the asymptotic confidence regions. Unreliability of the confidence regions can come from many different sources, but we are going to focus only on one of these, the imprecision of the variance estimate. The illustration of the full-blown problem is given in Section 4.1. Section 4.2 describes the concept of small sample variance loss, that was introduced and elaborated in [Schoukens et al., 2013], [3] and [13]. A correction heuristic is given in Section 4.3 that suggests rescaling of the asymptotic confidence regions by taking the finite sample variance loss into account.

## 4.1 Illustration of Finite Sample Effects

As the asymptotic confidence regions are constructed using the estimates of the second order characteristics of the distribution of \( \hat{\theta} \), the behavior of these confidence regions with respect to different sample counts is captured best by the behavior of the variance estimates belonging to the estimate \( \hat{\theta} \). In order to be able to compare the outcome of identification scenarios with different sample counts, the problem setup needs to be such that the infinite measurement horizon is homogeneous in time in terms of both input and noise. To this end we could consider stationary
Chapter 4. Effect of Finite Measurement Length on Uncertainty Estimation

Figure 4.1: Normalized variances of parameters against sample count.

stochastic processes for both of these signals, but for the sake of simplicity we are going to assume white noise signals. We acknowledge that white noise is not the best input signal for identification, to say the least. The consequences drawn throughout the chapter also extend to other choices of stationary signals; we will mention the consequences of the use of colored signals in notes.

In the example we are going to use the ZOH discrete time step equivalent of the continuous transfer function $G(s) = \frac{1+2s}{(2.5s^2+2.0.52.5s+1)(1+1.5s)}$ with output error noise model and zero initial condition. This results in the parameter vector $\theta = [b_1 \ b_2 \ b_3 \ f_1 \ f_2 \ f_3]^T$, with nominal values $\theta_0 = [0.0880 \ 0.0195 \ -0.0442 \ -2.0536 \ 1.4611 \ -0.3442]$. The dominant time constant of the system is less than five samples. We simulated the system for sample counts ranging from 10 to 1000 samples with different noise and input realizations. Our goal is to illustrate the difference between the asymptotic variance estimate and the finite sample behavior. To this end Figure 4.1 shows the expected value of the variance of the estimate $\hat{\theta}$, compensated for the $1/N$ decay.

Figure 4.1a shows the normalized variances of the numerator coefficient, while Figure 4.1b shows the same for the denominator coefficients. What can be seen on these figures is that as the sample count increases all of these curves converge to a constant value, to the normalized asymptotic variance. The asymptotic theory
approximates these curves with a horizontal line corresponding to the limiting value.

Another feature is that, for small values of the sample count, the normalized variance differs significantly from that predicted by the asymptotic theory. We are going to focus on this effect. Before doing that, there is another fact that should be emphasized and it is visible in Figure 4.1b, where in the beginning of the scale the variance seems to increase with the sample count. This is due to the fact that the PEM estimate is biased. For small sample counts most of the estimate variance is “traded” for bias error. Once the bias becomes negligible, the variance starts to decrease with growing sample count, as expected. The bias error depends heavily on the signal-to-noise ratio (SNR). For large SNR values the bias is small and vanishes quickly. Whereas for small SNR values it might take more samples to overcome the bias than for the normalized variance estimates to approach the asymptotic values. We are going to limit our attention to the case where the bias effects can be neglected, i.e. the SNR is high enough.

### 4.2 Finite Sample Variance Loss

The goal of this section is to quantify the difference between the variance of the estimates compared to what is predicted based on the asymptotic theory. We are going to rely on the fact that prediction error estimates minimize a cost function in the form $\sum_{k=1}^{N} \ell(\varepsilon[k])$ and that they are asymptotically unbiased and Gaussian distributed

$$\sqrt{N}(\hat{\theta}_N - \theta_0) \xrightarrow{d} N(0, \Sigma_{\infty}) \quad \text{Var}(\hat{\theta}_N - \theta_0) = \frac{1}{N} \Sigma_{\infty}$$

(4.1)

**Definition 4.1** (Finite sample variance loss). For a scalar parameter $\theta$, let $\sigma^2_{N|U}(\hat{\theta}_N)$ denote the variance of the parameter estimate $\hat{\theta}_N$ obtained using $N$ samples, conditioned on the realization of the input $U$. The asymptotic variance is defined as $\sigma^2_{\infty,U}(\hat{\theta}) = \lim_{N \to \infty} N \sigma^2_{N|U}(\hat{\theta}_N)$. The finite sample variance loss for $N$ samples is defined as

$$\alpha_{N|U}(\hat{\theta}_N) = \frac{N \sigma^2_{N|U}(\hat{\theta}_N)}{\sigma^2_{\infty,U}(\hat{\theta})}$$
The *expected finite sample variance loss* for $\theta$ is defined as

$$\bar{\alpha}_N(\hat{\theta}_N) = \mathbb{E}_U \left( \alpha_{N|U}(\hat{\theta}_N) \right)$$

Since we assumed that both the input and the noise is a stationary signal, the asymptotic variance $\sigma_{\infty|U}^2$ is independent of the particular realization of the input. However, the finite sample version of this quantity $N\sigma_{N|U}^2(\hat{\theta}_N)$ depends on the actual realization of $U$ heavily. In the case when multiple different realizations of the input $U$ correspond to the same asymptotic properties (i.e. multi-sine signal with random initial phase but common spectra, realizations of a stationary stochastic process) the expected variance loss $\bar{\alpha}_N$ captures the part of the variance loss that is characteristic to the asymptotic properties of the input, while $\alpha_{N|U}$ depends on the chosen realization of the input $U$.

**Distribution of the variance loss**

We want to emphasize that the finite sample variance loss $\alpha_{N|U}(\hat{\theta}_N)$ is a quantity depending on the distribution of the estimate. In case of LS linear regression problems the variance loss depends only on the regressors $X$ and the second order moment of the noise. However in the case of dynamical systems, since the estimate is a nonlinear function of the measurement data, its distribution is not accessible.

It follows from the definition of the variance loss that $\alpha_{N|U}(\hat{\theta}_N) \to 1$ as $N \to \infty$, and so does its expectation $\bar{\alpha}_N(\hat{\theta}_N)$. For linear regression problems with random regressors it is a direct consequence of Jensen’s inequality that the expected variance loss $\bar{\alpha}_N(\hat{\theta}_N)$ is always greater than one.

**Assessing validity of asymptotic results**

The reason for paying attention to the finite sample variance loss is that it provides means to assess if the asymptotic theory is reliable at a given sample count or not. If the variance loss is significantly greater than one, than we can say for certain that the asymptotic theory will underestimate the uncertainty of the estimate. Of course, the exact value of the variance loss $\alpha_{N|U}(\hat{\theta}_N)$ is not available, but it is always possible to generate synthetic measurement data using the obtained estimate $\hat{\theta}_N$. 
Chapter 4. Effect of Finite Measurement Length on Uncertainty Estimation

The variance loss can be estimated based on these simulated measurements and this estimate should serve as an indication.

Choosing experiment length

The previous paragraph shows how the variance loss can be used once a measurement is already available. Other use cases can be found in the field of experiment design, when the identification procedure starts even before the input signal $U$ is fixed. Suppose that an identification protocol is being defined, where the input signal will contain random components (i.e. a random phase multisine or pseudorandom binary sequence) and the end result is expected to be an estimate with uncertainty guarantees. If the uncertainty estimate is delivered using asymptotic results, then the measurement length of the protocol needs to be selected such that the variance loss $\alpha_N(U(\hat{\theta}_N))$ is small in some sense. The sample count can be selected based on different criteria, such as

$$
N_{\varepsilon} = \min_N \left\{ \bar{\alpha}_N(\hat{\theta}_N) \leq 1 + \varepsilon \right\} \quad N_{P,\varepsilon,\delta} = \min_N \left\{ \mathbb{P}_U(\alpha_N(U(\hat{\theta}_N) < 1 + \varepsilon) > \delta \right\}
$$

(4.2)

Figure 4.2 shows the evolution of the estimation error for $f_1$ in the example problem. The log-log scatter plot shows $(\hat{f}_1 - f_1)^2$ for independent simulations of the estimation problem. The straight orange line shows the predicted value based on the asymptotic variance. The blue curve shows the average of the found values, this is the expectation of $(\hat{f}_1 - f_1)^2$ with respect to the finite sample count.
Chapter 4. Effect of Finite Measurement Length on Uncertainty Estimation

The purple curve shows the 0.9 percentile of the distribution of the finite sample squared estimation error. As the sample count tends to infinity, these three curves converge to each other, but for finite samples the difference is significant.

The purpose of the finite sample variance loss is to quantify the excess of the two higher curves with respect to the asymptotic theory based prediction.

This figure shows nicely that if we prescribe record lengths based on thresholds of these different quantities then the asymptotic estimate will be the first to go below the threshold, resulting in the shortest record length. The expected variance criterion will suggest a slightly longer record length, while the percentile based criterion requires the highest number of samples to meet the same threshold.

An interesting feature of the variance loss, that we have encountered in many simulation experiments, is that even if the bias of the estimates was already negligible for a given sample count, the variance loss was still significantly greater than one. And the variance loss remained significant for unexpectedly high values of the sample count. In the above example the dominant time constant of the system is 2.5 and the 2% settling time is 21 samples, yet it takes almost 400 samples to push the expected variance loss below 1.1, $N_{E,0.104} \approx 383$. This shows that hundreds of dominant timeconstants worth of measurement is needed to get into the asymptotic range.

Remark 4.2. The variance loss remains significant even for sample counts for which the estimates themselves are accurate.

4.3 Approximation of Expected Variance Loss in LTI Systems

Being able to calculate the variance loss exactly is equivalent in difficulty to being able to characterize the distribution of the estimates exactly. This is a difficult and analytically intractable problem in most cases; this is one of the reasons why asymptotic estimates are used for finite sample variance estimation. My goal is to present an approximation procedure for the variance loss of dynamical systems parameter estimates that can be used in the case when the estimate $\hat{\theta}_N$ is unbiased.
This approximation is computationally more tractable than the exact problem and provides reasonable results.

Introducing the row vectors $\phi[k] \in \mathbb{R}^{1 \times n_\theta} = \frac{\partial y[k]}{\partial \theta} \bigg|_{\theta=\theta_0}$, the matrix $\Phi \in \mathbb{R}^{N \times n_\theta} = (\phi[k])_{k=1}^N$ and using the assumption that the estimate $\hat{\theta}_N$ is unbiased, it is known that its covariance is bounded by the Cramér-Rao lower bound, that is

$$\mathbb{E} \left( (\hat{\theta}_N - \theta_0)(\hat{\theta}_N - \theta_0)^T \bigg| U \right) \geq M_U^{-1} \quad (4.3)$$

where

$$M_U = \frac{1}{\sigma_y^2} \mathbb{E} \left( \Phi^T \Phi \big| U \right) \quad (4.4)$$

In view of this setup the effect of using the asymptotic theory can be interpreted as follows. When the variance loss is calculated, $M_U$ is no longer treated as a random variable, but it is considered as a deterministic object determined by the asymptotic properties of the input and noise. So the asymptotic theory replaces the lower bound $\mathbb{E} \left( M_U^{-1} \right)$ with $(\mathbb{E} (M_U))^{-1}$. This is the reason why the asymptotic theory always underestimates the uncertainty of unbiased estimates.

Since exact calculation of the variance loss is not feasible, we are presenting a computationally tractable approximation algorithm that is able to capture the characteristics of the variance loss. The algorithm is presented in Section 4.3.1 and it is executed in a detailed example in Section 4.3.2.

### 4.3.1 Approximation Algorithm

The rest of this section presents an approximation algorithm that can be used to approximate the lower bound $\mathbb{E} \left( M_U^{-1} \right)$ more accurately compared to the $1/N$ rescaling of the asymptotic results. This approximation is exact for linear regression problems with i.i.d. Gaussian regressors, and it is only an approximation for dynamical systems. Algorithm 4.1 is a moment matching based approximation method that can be used to evaluate the expected variance loss $\bar{\alpha}_N((\hat{\theta}_N)_i)$.

The moment matching in the third step of Algorithm 4.1 can be realized by finding the root of a scalar function $f(X^T X, \Phi^T \Phi)$. For one dimensional problems the trivial choice is the solution of the equation $\mathbb{E}(X^T X - \Phi^T \Phi) = 0$. For higher
Algorithm 4.1 Approximating the expected variance loss

1: Calculate the covariance matrix \( \Psi = \mathbb{E}(\phi^T[N]\phi[N]) \).
2: Let \( X \in \mathbb{R}^{\kappa N \times n_\theta} \) be a matrix whose rows are independent Gaussian random vectors with covariance \( \Psi \).
3: Determine the degrees of freedom \( \kappa N \) of the matrix \( X \) by means of moment matching between \( X^TX \) and \( \Phi^T\Phi \).
4: Approximate the expected finite sample variance loss as
\[
\bar{\alpha}_N((\hat{\theta}_N)_i) \approx \frac{1}{\sigma_y^2} \frac{N}{\kappa N - n_\theta - 1} (\Psi^{-1})_{i,i} \sigma_y^2
\]

dimensional problems matching the expectation of the \( i \)-th diagonal term is a viable choice, or matching the determinants as \( \mathbb{E}(|X^TX|) = \mathbb{E}(|\Phi^T\Phi|) \).

The approximation factor \( \frac{N}{\kappa N - n_\theta - 1} \) in the fourth step is the result of the Gaussian approximation in the second step. As \( X \) is a Gaussian matrix, \( XX^T \) is an \( n_\theta \times n_\theta \) random matrix with Wishart distribution \( \mathcal{W}_{n_\theta}(\Psi, \kappa N) \) defined by the covariance matrix \( \Psi \) and \( \kappa N \) degrees of freedom. The expected value of the inverse of a \( \mathcal{W}_n(\Sigma, \nu) \) distributed matrix is \( \frac{1}{\nu-n-1} \Sigma^{-1} \). For details about the Wishart and inverse Wishart distributions see [Mathai and Provost, 1992].

Figure 4.3 shows the behavior of the expected variance loss with respect to sample count for different parameters of the example system, together with the approximation provided by Algorithm 4.1. The empirical values are marked with

![Figure 4.3: Empirical and approximate expected variance loss of parameters.](image-url)
circles, while the approximate curves have square markers. Lines with the same color correspond to the same parameter.

Since the approximation in Algorithm 4.1 replaces $\Phi$ with $X$, where the correlation between the rows of $\Phi$ is forgotten and the distribution of the rows is approximated with Gaussian distribution. The lost correlation information is supposed to be captured by parameter $\kappa$ that is obtained via the moment matching step, while information from the distribution of individual rows is kept through the covariance matrix of the row distribution.

### 4.3.2 Concrete Approximation Example

The goal of this section is to present the proposed approximation algorithm on an example, by calculating the approximate expected variance loss of $\hat{f}_1$ in an output error model, so

$$y[k] = \frac{b_1 z^{-1}}{1 + f_1 z^{-1}} u[k] + e[k]$$

(4.5)

We are going to assume that the input $u[k]$ is zero mean white noise with variance $\sigma_u^2$ and the output noise is zero mean white Gaussian noise with variance $\sigma_y^2$.

#### 4.3.2.1 A Closer Look on the Approximation Algorithm

Before going through the steps of Algorithm 4.1 we take a closer look at the calculation of the Cramér-Rao lower bound, which can be calculated using the derivatives of the log-likelihood function (see [Ljung, 1999, p. 215])

$$M_U = \mathbb{E}_E \left( \left[ \frac{\partial}{\partial \theta} \log f(\theta, Y, U) \right] \left[ \frac{\partial}{\partial \theta} \log f(\theta, Y, U) \right]^T \right|_{\theta = \theta_0} \right)$$

(4.6)

where $f(\theta, Y, U)$ denotes the joint density of the input $U$ and the measurements $Y$ for a model $\theta$, while $\mathbb{E}_E$ denotes the expectation taken over the random variables of the noise $e[k]$. 
Assuming Gaussian output error noise model we can explicitly write the derivative of the log-likelihood function as

$$\frac{\partial}{\partial \theta} \log f(\theta, Y, U) \bigg|_{\theta = \theta_0} = \frac{1}{\sigma_y^2} \sum_{k=1}^{N} e[k] \frac{\partial}{\partial \theta} \hat{y}[k]$$  \hspace{1cm} (4.7)

where $\hat{y}[k]$ is the one step ahead prediction of the model. Introducing the row vectors $v[k] = e[k] \frac{\partial}{\partial \theta} \hat{y}[k]$ we can write $M_U$ as

$$M_U = \frac{1}{\sigma_y^2} \mathbb{E}_E (V^TV)$$  \hspace{1cm} (4.8)

where $V \in \mathbb{R}^{N \times n_\theta}$. Using the fact that the pair $e[k]$ and $\frac{\partial}{\partial \theta} \hat{y}[k]$ are independent for every $k$ and applying the law of iterated expectations for $e[N], e[N-1], \ldots$ we get that

$$M_U = \frac{1}{\sigma_y^2} \mathbb{E}_E (\Phi^T \Phi)$$  \hspace{1cm} (4.9)

where $\phi[k] \in \mathbb{R}^{1 \times n_\theta} = \left. \frac{\partial}{\partial \theta} \hat{y}[k] \right|_{\theta = \theta_0}$ and the matrix $\Phi \in \mathbb{R}^{N \times n_\theta} = (\phi[k])_{k=1}^{N}$. Calculating the expectation in (4.9) is difficult due to the dependence between the rows of $\Phi$.

If the input and the noise are assumed to be (filtered) Gaussian signals then the distribution of $\phi[k]$ is also Gaussian with an appropriate covariance matrix. Since the Cramér-Rao bound is given in terms of $M_U^{-1}$, the expected Cramér-Rao bound depends on $\mathbb{E}_U(M_U^{-1})$. The main difficulty in calculating the expected Cramér-Rao bound is caused by the matrix inversion.

The inverse Wishart distribution $W_n^{-1}(\Psi, \nu)$ is the distribution of random matrices of the form $[X^TX]^{-1}$, where $X$ has $\nu$ independent rows that are Gaussian vectors with covariance $\Psi$ ([Mathai and Provost, 1992]). The goal of Algorithm 4.1 is to approximate $M_U$ with a Wishart matrix $W_n(\Psi, \nu)$ which allows the approximate calculation of $\mathbb{E}_U(M_U^{-1})$, since the expectation of $W_n^{-1}(\Psi, \nu)$ can be given as $\frac{1}{\nu - n_\theta - 1} \Psi^{-1}$ where $n_\theta$ is the dimension of $\Psi$. This approximation boils down to finding the appropriate covariance matrix $\Psi$ and degree of freedom $\nu$. These are steps 1 and 3 of Algorithm 4.1.
4.3.2.2 Approximating the Wishart Parameters

Covariance matrix. If both the input and the noise signal are stationary stochastic signals, then \( \Psi[k] = \mathbb{E}\left(\phi^T[k]\phi[k]\right) \) converges as \( k \) tends to infinity. Algorithm 4.1 chooses the covariance matrix \( \Psi[N] \) to be used in the Wishart approximation.

For the sake of the example, we are going to calculate the variance of \( \frac{\partial}{\partial f_1} \hat{y}[k] \).

As it is given in [Ljung, 1999, p. 330], the partial derivative of the prediction with respect to the model parameter \( f_1 \) can be calculated as

\[
\frac{\partial}{\partial f_1} \hat{y}[k] = -\frac{\mathcal{D}(z)\mathcal{B}(z)}{\mathcal{C}(z)\mathcal{F}^2(z)} u[k-1] = -\frac{\mathcal{B}(z)}{\mathcal{F}^2(z)} u[k-1]
\]

(4.10)

where we used the fact that \( \mathcal{B} = \mathcal{C} = \mathcal{D} = 1 \) for output error models. If \( u[k] \) is a (filtered) sequence of i.i.d. zero mean Gaussian variables then the variance parameter \( \Psi[N] = \mathbb{E}\left(\left(\frac{\partial}{\partial f_1} \hat{y}[k]\right)^2\right) \) of the approximation can be calculated using Statement B.1. Let \( (g_k)_{k \geq 0} \) denote the impulse response of the system \( \frac{\mathcal{B}(z)}{\mathcal{F}^2(z)} z^{-1} \) with the polynomial coefficients having their nominal value from \( \theta_0 \). If \( u[k] \) is a white signal, this results in

\[
\Psi[N] = \sigma_u^2 \sum_{k=0}^{N} g_k^2
\]

(4.11)

In case of the first order example model (4.5) we have that the impulse response is

\[
\frac{\mathcal{B}(z)}{\mathcal{F}^2(z)} z^{-1} = \frac{b_1 z^{-2}}{(1 + f_1 z^{-1})^2} = \frac{b_1}{f_1^2} \sum_{k=1}^{\infty} k(-f_1)^{k+1} z^{-(k+1)}
\]

(4.12)

\[
g_k = (k-1)(-f_1)^{k-2} \quad k \geq 2
\]

(4.13)

Summing the squares of these values, according to (4.11) we get that

\[
\Psi[k] = \sigma_u^2 \frac{b_1^2}{f_1^2(1 - f_1^2)^3} \left[ (-(k - 1)^2 + 2(k - 1) - 1) f_1^{2(k+1)} \\
+ (2(k - 1)^2 - 2(k - 1) - 1) f_1^{2k} \\
- (k - 1)^2 f_1^{2(k-1)} + f_1^4 + f_1^2 \right]
\]

(4.14)
Chapter 4. Effect of Finite Measurement Length on Uncertainty Estimation

Degrees of freedom. The goal of the moment matching step in Algorithm 4.1 is to find a suitable value for the degrees of freedom parameter \( \nu = \kappa N \) such that the matrix \( \Phi^T \Phi \) matches a \( \mathcal{W}_{n_\theta}(\Psi[N], \kappa N) \) matrix in some sense. In case of a scalar parameter \( f_1 \) the obvious choice is to match their expected value, \( \mathbb{E}(\Phi^T \Phi) = \kappa N \Psi[N] \). This results in the relation

\[
\sum_{k=3}^{N} \Psi[k] = \kappa N \Psi[N] \tag{4.15}
\]

with the solution being

\[
\kappa = \frac{\frac{4f_1^6 - f_1^4(N-2) - f_1^{6+2N}(N-2)^2 + f_1^8 N - f_1^{2+2N} N^2 + 2f_1^{4+2N}(N^2 - 2N - 1)}{N}}{(f_1^4 - 1)(f_1^4 + f_1^6 - f_1^{4+2N}(N-2)^2 - f_1^{2N}(N - 1)^2 + f_1^{2+2N}(3 - 6N + 2N^2))}
\]

4.3.2.3 Deducing the Expected Variance Loss

Since we determined the parameters of the Wishart approximation \( \Phi^T \Phi \approx \mathcal{W}_{n_\theta}(\Psi[N], \kappa N) \), we have that \( M_U \approx \mathcal{W}_{n_\theta}(\frac{1}{\sigma_y^2} \Psi[N], \kappa N) \), providing the expected Cramér-Rao bound \( \frac{1}{\kappa N - n_\theta - 1} \Psi^{-1}[N] \sigma_y^2 \). Compensating this bound against the \( 1/N \) decay we get that the finite sample expected covariance estimate

\[
\frac{N}{\kappa N - n_\theta - 1} \Psi^{-1}[N] \sigma_y^2 \tag{4.16}
\]

is to be compared with the asymptotic covariance

\[
\lim_{N \to \infty} \frac{N}{\kappa N - n_\theta - 1} \Psi^{-1}[N] \sigma_y^2 = \frac{\sigma_y^2}{\sigma_u^2} \frac{(1 - f_1^2)^3}{b_1^2(f_1^2 + 1)} \tag{4.17}
\]

This expression is fairly complicated due to the complex form of \( \kappa \). In order to get a more revealing expression, we can neglect those terms in \( \kappa \) that are sample count dependent powers of \( f_1 \) (assuming a stable system with \( |f_1| < 1 \)). If we assume \( b_1 \) to be known, thus \( n_\theta = 1 \) we get to a simplified approximate value of the expected variance loss of \( f_1 \)

\[
\tilde{\alpha}_N(f_1) \approx 1 + \frac{\frac{4+4f_1^2-2f_1^4}{1-f_1^2}}{\frac{4+4f_1^2-2f_1^4}{1-f_1^2}} \tag{4.18}
\]
This expression shows that the expected variance loss decays roughly as $1/N$ and the details of the decay depend on the nominal values of the system.

### 4.3.3 Properties of the Approximation

Before discussing the specific properties of the approximations we show the behavior of the variance loss in Figure 4.3. This figure shows the expected variance loss for all parameters of the sample system used in Section 4.1. The expected variance loss $\bar{\alpha}_N(\cdot)$ is estimated empirically using 10000 independent realizations of the input and output signals. The approximations $\approx \bar{\alpha}_N(\cdot)$ resulting from Algorithm 4.1 are calculated using the nominal values of the parameters for calculating $\Psi$ in the first step of the algorithm.

One particular consequence of the proposed algorithm is that the variance loss of parameters belonging to specific parts of the model behave similarly. This follows from the form of the derivatives of the predictions with respect to different coefficients in the same model polynomial. Namely, that they are given as the output of the same filters on delayed versions of the input and measured output (see [Ljung, 1999, p. 330]). If these signals are stationary stochastic processes, then so are their filtered version. This behavior is nicely visible in Figure 4.3b for both the simulated and the approximate expected variance losses. For some reason the empirical curves in Figure 4.3a are not aligned as predicted by this principle and we have no explanation for that.

Another consequence of replacing the rows of $\Phi$ with independent random vectors during the approximation is that this allows more variability in $X^T X$ compared to $\Phi^T \Phi$. This is the reason why the approximate curves are below the empirical ones. Although this will ensure that if experiment lengths are chosen according to (4.2), then we are underestimating the actual record length that would suffice to achieve our goals, but the estimates are still going to be better than that given by the asymptotic theory.

Multiple consequences can be drawn from the behavior of the expected variance loss. The distribution of the variance loss is not concentrated on a single value for small values of $N$, thus it can be above or under the expected value. This shows that, when choosing input signals for short experiments, the emphasis should shift...
from designing asymptotic properties of a signal towards the design of the actual realization.

The variance loss remains significant for quite long, in the range of hundred dominant time constants in the illustration example.

Before conducting an identification experiment, the length of the data record can be selected such that either the approximate variance of the estimates or their variance loss goes below a prescribed bound. This will result in longer experiments compared to the case when the same condition is ensured using only the asymptotic variance.

### 4.3.4 Initial Conditions and Variance Loss

We considered both the simulation example of Section 4.1 and the analytic example of Section 4.3.2 with known zero initial conditions for the sample systems. Now we are going to intuitively discuss the relationship between the initial conditions and the variance loss.

The first point that we would like to make is that known nonzero initial conditions are expected to decrease the variance loss. The practically more relevant situation is to consider unknown initial conditions. If we constrain our attention to the case where the input $u[k]$ is still a zero mean (filtered) Gaussian sequence, then it is expected that the variance loss is going to be increased compared to the case of known zero initial condition.

Algorithm 4.1 can consider initial conditions implicitly through the derivatives $\dot{\phi}[k] = \frac{\partial}{\partial \theta} \dot{y}[k]$, however we wish to emphasize that the impact of initial conditions on the variance loss is not particularly significant. This can be attributed to two facts.

Firstly, the approximation is governed by the decay of the impulse response of the filters providing the derivatives. These filters are always significantly slower compared to the original system. In case of the output error example we have that the denominator of the plant is $\mathcal{F}(z)$, while the denominator of the filter for $\psi[k]$ is $\mathcal{F}^2(z)$. This shows that the effects of initial conditions die out faster compared to how the calculations involving the impulse response of the filters for $\psi[k]$ converge to their stationary values.
Chapter 4. Effect of Finite Measurement Length on Uncertainty Estimation

The second argument for not taking initial conditions into consideration more explicitly is that the approximations provided by Algorithm 4.1 make practical sense only after the bias error of the estimates is negligible. The effects of a stationary random initial condition are negligible by this time.

Numerical simulations have shown that the difference between the empirical expected variance loss for the known zero initial condition and the unknown stationary initial condition case are not visible.

4.3.5 Filtered Noise Input

Since white noise is not the ideal identification input signal we can also consider filtered white noise. Statement B.1 is formulated such that it provides means to calculate $\psi[k]$ also if the input signal $u[k]$ is filtered white noise.

Due to the complicated nature of the calculations we investigated only numerically the effect of filtered white noise excitation to see how filtering affects the variance loss. The results of a typical experiment result are shown in Figure 4.4.

Estimations were carried out for different values of the sample count $N = kT_0$ where $T_0$ is the dominant time constant of the nominal system in samples, and different filters providing the input signal. The input filter was a first order filter.
Chapter 4. Effect of Finite Measurement Length on Uncertainty Estimation

Figure 4.5: The estimated expected variance loss $\tilde{\alpha}_N(\hat{f}_1)$ for filtered white noise input against different filtering frequencies and sample counts $N/T_0$

with time constant $\tau$. Figure 4.4 shows the normalized variance of the estimate $\hat{f}_1$ when the spectrum of the input signal is varied. The input signal was normalized in power after the filtering, so the experiments are comparable. As we have seen in the case of white noise, the normalized variance converges as $N$ grows but for small values of $N$ it is significantly higher than the limiting value. This behavior is common regardless of the spectrum of the input signal.

The limiting value of the normalized variance and the concrete behavior of the variance loss depends on the cutoff frequency of the filter. It is known that the best input signal to estimate $\hat{f}_1$ is a sine wave with matching frequency. We see a similar phenomenon here as well. If the input signal is white noise passing through the same filter that we are identifying then the limiting normalized variance is small. As the cutoff frequency of the input filter moves away from that of the system to be identified we see this quantity degrade.

Figure 4.5 shows the behavior of the variance loss estimate in a color coded plot. As it can be seen, the variance loss is higher around the characteristic frequency of the nominal system. This is because the base limiting value is small in this range.
and the finite sample variance estimate from the first few time constants of data cannot differentiate based on the spectrum of the input signal. This difference in the variance loss disappears as the sample count is increased. In this example the values of variance loss behave similarly after $N/T_0 \sim 7$.

Nonetheless, the observations about the variance loss mentioned in Section 4.3.3 remain valid in case of filtered white noise excitation as well.

### 4.3.6 Adjusting Asymptotic Confidence Sets for Variance Loss

After the data is collected and the estimation is carried out, the corresponding asymptotic variance estimates can be adjusted to account for the small sample variance loss. This can be done either by relying on simulations using $\hat{\theta}$, or by using approximate expressions supplied by Algorithm 4.1. Computational time can be saved by using the same value for coefficients of the same model polynomial.

Let us assume that the variance loss $\alpha_{N,U}(\hat{\theta}_N)_i$ of each parameter $\theta_i$ is approximated and the asymptotic covariance matrix of the estimate is $\Sigma_{\infty}$. A rescaled finite sample covariance matrix can be obtained by transforming the eigenvectors of $\Sigma_{\infty}$ such that the axes lengths of the corresponding ellipse are adjusted. Let $\alpha \in \mathbb{R}^{n_\theta}$ be the vector containing the finite sample variance losses of the individual parameters $\left(\alpha_{N,U}(\hat{\theta}_N)_i\right)_{i=1}^{n_\theta}$ and let the rescaling operator $r(v, \alpha)$ be defined as $r(v, \alpha) : \mathbb{R}^{n_\theta} \times \mathbb{R}^{n_\theta} \rightarrow \mathbb{R}^{n_\theta}; (r(v, \alpha))_i = v_i \sqrt{\alpha_i}$. The finite sample covariance matrix is obtained as

$$\Sigma_N = N \sum_{i=1}^{n_\theta} \lambda_i r(v_i, \alpha) r^T(v_i, \alpha)$$

where $\lambda_i$ and $v_i$ are the eigenvalues and eigenvectors of $\Sigma_{\infty}$. This method provides a multidimensional rescaling of the asymptotic covariance matrix. This rescaling keeps the covariance structure provided by the asymptotic approximation while trying to correct for the finite sample variance loss. This finite sample covariance matrix can be used instead of the asymptotic one to obtain a confidence region, thus maintaining a simple description for the uncertainty region while incorporating the finite sample variance loss.
Since we are focusing on finite sample identification scenarios it is important to explicitly mention how initial conditions fit into this picture. As the finite sample variance loss is defined in terms of the asymptotic variance, it is not even defined for initial conditions (or it is infinite). The previously outlined correction algorithm relies on the estimate of the asymptotic covariance matrix and the use of initial condition variables in the estimation helps to improve the covariance estimate.

Another conclusion that we could draw from our simulation studies is that the variance loss remains significant for surprisingly long. This shows that, even though the variance loss does not account for the initial conditions, if the experiment length is chosen such that the variance loss becomes small, then the influence of the initial condition will also be small in the experiments. This is the reason why Algorithm 4.1 cannot be applied to initial condition variables in $\theta$.

So far we have approached the problem from a direction where we assumed a given parametric model of the system. This allowed the definition of the approximation Algorithm 4.1 and the correction procedure (4.19). In order to offer some insight into the behavior of the variance loss across different model structures, we calculated approximate expressions for the $n$-th order output error model. The details of the calculation can be found in [13] where we found that the moment matching results in $\kappa = \frac{T_s}{\tau}$, where $\tau$ is the dominant time constant of the system and $T_s$ is the sampling time. The diagonal terms of $\Psi^{-1}$ were further approximated by $\kappa N^2 \sigma^2$, which gave a simpler formula for the expected variance loss

$$\bar{\alpha} \approx \frac{\kappa N}{\kappa N - n_\theta - 1}$$

(4.20)

Expressing the measurement length $N$ in terms of the model complexity and variance loss we get that

$$N \cdot T_s \approx \tau (n_\theta + 1) \frac{\alpha}{\alpha - 1}$$

(4.21)

This expression shows that with fixed variance loss the required measurement length grows linearly with model complexity and with the dominating time constant. Decreasing the variance loss induces a more significant increase in the required sample count.
4.4 Conclusions

The main conclusion that can be drawn from the studies related to this chapter is that obtaining reliable uncertainty estimates based on the asymptotic theory requires really long measurements. If the experiments are carried out on a shorter time scale then the actual realization of the input signal has an increased importance.

The notion of small sample variance loss makes it possible to quantify if the record length is in the asymptotic regime or not. The algorithm proposed for approximating this quantity shows that, on top of the $1/N$ decay of the variance estimate, the variance loss also decays roughly as $1/N$, but the details of the decay for the variance loss differ from that of the variance. This can be attributed mostly to the fact that the derivatives of the predictions are produced by filtered version of the input and measurement noise and these filters are slower than the plant and noise models under identification.

For long experiments it is a good choice to design asymptotic properties of the input signals, while for short ones the focus should be on the actual realization. The proposed concept of variance loss can help to decide which of these two approaches should be taken.

The variance loss can also be used to adjust asymptotic confidence regions by re-scaling their axes using the the value of the variance loss.
Chapter 5

Concluding Remarks

The focus of the presented work was concentrated around problems arising when identification is carried out using a restricted number of samples. As usual, this restriction is relative to the complexity of the model to be identified and the information content of the data record.

After a brief introduction in the first chapter, the second chapter described a generic family of hypothesis testing methods that can be used to construct finite sample exact confidence regions for a large variety of estimation problems. The main features of data perturbation methods are that: they rely only on mild assumptions about the noise while still being able to guarantee exact confidence levels; they are randomized algorithms; their efficiency depends on the information content of the data record. Some of the practically relevant noise distribution assumptions were discussed and the algorithm was illustrated on a detailed example. We have derived conditions to quantify when the defined hypothesis tests reduce to a simple random coin tosses. This can happen in case of wrongly chosen performance measures and also in case of poor input excitation.

Special attention was devoted to linear regression problems as they offer deep insight and are analytically tractable. The gained insight was used to extend the methodology to estimation of linear dynamical systems parameters. This is done in a way that we consider to be more natural compared to previous choices in the literature.

The algorithm proposed for constructing confidence regions for LTI systems relied on the existence of an estimator. Although the exact confidence level was
independent of the quality of this estimator, the shape of the resulting confidence regions depended on it heavily. This motivated the search for estimation procedures that are able to provide globally optimal estimates. The content of Chapter 3 investigated the applicability of Lasserre hierarchies for least squares estimation. Although the sparsity structure of the identification optimization problem allows a significant reduction in computational complexity compared to the general complexity of the approach, we had to come to the conclusion that identification is still practically unfeasible with the currently available computational power. Initialization of local search methods may be the single relevant use case where the Lasserre hierarchies can be used with good results.

Since computing exact finite sample confidence regions turns out to be computationally demanding, we turned our attention to trying to utilize asymptotic results in the finite sample regime. With the definition of finite sample variance loss in Chapter 4, we were able to define conditions to quantify the record length needed in order to be able to safely apply the asymptotic theory for specific models. Since calculating the variance loss is difficult, we proposed an approximation algorithm for its expected value. The approximate formulae that were derived for the expected variance loss offered insight into the behavior of this quantity. The most significant finding was that the variance loss remains significant for unexpectedly long measurement lengths. Beside assessing the appropriateness of a given record length for identification, the variance loss can also be used to adjust the confidence regions obtained using asymptotic theory.

We hope that the results and discussion contained in the thesis helps the identification community to better understand problems appearing in identification scenarios where the amount of available information cannot be considered infinite.
Appendix A

Statement of contribution

A.1 Preliminaries and objectives

System identification as a research field continues to evolve in many directions. The focus has moved from identification of the now classical SISO linear models and it is shared now between different model structures and identification scenarios. Linear and non-linear models in block structures or even large networks are considered. The same topics are also approached from an experiment design perspective. Among many directions the field has taken one is concerned with identification problems where the measurement time is relatively short, meaning that the amount of collected data is barely enough to facilitate meaningful estimation. The uncertainty analysis of the estimates in such a scenario requires different tools than what can be used when practically an infinite amount of data is available. Distribution free methods form a substantial part of the applicable toolkit, their importance was also emphasized in a keynote of the 16th IFAC Symposium on System Identification [Campi et al., 2012]. The philosophy of the distribution free theory captured my attention, this contributed a great deal when selecting the topic.

The objective of my research was to understand the already available tools and push the boundaries of existing methods where possible. This thesis contains the results of this endeavor.
A.2 New scientific results

The results are organized into three theses. These approach the system identification problem from different perspectives where the common denominator is the lack of some information that is assumed available in the mainstream system identification literature.

It is usually assumed that the distribution of the noise that contaminates our measurements is known, this enables us to calculate uncertainty certificates for the parameter estimates. These uncertainty quantifications may also rely on asymptotic theory. The first thesis describes a randomized uncertainty quantification method inspired by [Csáji et al., 2012a] and related works. I describe a family of hypothesis testing methods called data perturbation methods (DP methods) and the requirements for the estimation problem that need to be fulfilled so that these methods can be applied. I characterized a family of noise distributions for which connected and bounded confidence regions can be obtained under certain excitation conditions for linear regression problems. I also described data perturbation methods for identification of linear dynamical systems.

In many situations the identification problem can be cast as an optimization problem with multiple local minima. The second thesis is concerned with finding global minima of such optimization problems. I investigate possible reformulations of the linear dynamical system identification problem into a convex optimization problem using the so called Lasserre hierarchy. I give explicit formulae for the growth of the semidefinite programming problem size in different parametrizations of the identification problem. I conclude that the computational complexity of these methods restrict their usability to relatively short data records.

The third thesis considers identification problems with (partially) random excitation. I investigate the influence of random input on the uncertainty of the estimates. More precisely, how this effect vanishes as the measurement time increases. Based on this analysis, I formulate rules for the measurement length in order to ensure that, with high probability, the estimates will have good uncertainty properties.
The rest of this section lists all three theses in order, citing my publications corresponding to the topic, followed by the general overview of the work, and summary of the results.

**Thesis I:**

**Data perturbation methods for identification**

The following publications are related to this thesis: [1], [2], [5], [4].

Distribution free statistical methods have been studied in the statistics literature mainly in the first part of the twentieth century. These methods made it to the system identification community through the works [Campi and Weyer, 2005, Csáji et al., 2012a] and references therein.

The simplest parametric statistical example may be when a hypothesis about the mean of a sample is tested when the sample is drawn from a Gaussian population with known variance. When the u-test is constructed, we rely on the assumption that the population is Gaussian (the distribution assumption). Given the samples, the u-test delivers a deterministic answer whether to reject the null hypothesis or not.

The aim of distribution free methods is to facilitate hypothesis tests without exact knowledge about the distribution of the measurements. The SPS method introduced in [Csáji et al., 2012a] is an example to this with two main characteristics. On one hand it is a distribution free method, on the other hand the delivered result is not deterministic, but it is random.

Motivated by the results in [Csáji et al., 2012a] I described a more general family of hypothesis tests that has the same properties as the SPS method, called data perturbation methods, and I investigated its properties for linear regression problems and parameter estimation problems for linear dynamical systems.

**Subthesis I.1: Data perturbation methods**

I described a family of hypothesis testing methods that can deliver decisions on an exact, user chosen level, without relying on full distribution assumptions on the contaminating noise, although partial assumptions on the distribution are used.
The decisions delivered by methods from this family are random even with respect to the fixed sample. This family includes the SPS method.

In order to be able to use data perturbation methods on a given data set a number of assumptions and conditions need to be satisfied.

**Assumption A.1** (Model structure assumption). *We assume that the measurements come from a fixed model structure described in the following form*

\[
Y = f(\theta_0, X, E)
\]

\(X\) contains known data, not contaminated by noise, \(E\) contains the randomness contaminating the measurements, \(Y\) contains the observables, \(f : \Theta \times X \times E \rightarrow Y\) is a known mapping from the model parameters, inputs and noises to the observables.

Assumption A.1 means that we know perfectly how noise contaminates our measurements. I describe a family of methods with the goal to create a hypothesis test for the null hypothesis \(H_0 : \theta_0 = \theta\) and alternative hypothesis \(H_1 : \theta_0 \neq \theta\).

**Assumption A.2** (Invertibility with respect to noise). *A model \(f : \Theta \times X \times E \rightarrow Y\) is invertible with respect to noise if there exists a mapping \(f^* : \Theta \times X \times Y \rightarrow E\) such that*

\[
\forall \theta \in \Theta, \forall X \in X, \forall Y \in Y : \quad (\exists! E \in E : Y = f(\theta, X, E)) \Rightarrow E = f^*(\theta, X, Y)
\]

This invertibility assumption together with the model structure assumption means that for every set of data \(D = (Y, X)\) we can explicitly and uniquely calculate the contaminating noise \(E\) for every possible model parameter \(\theta\), if the data can be generated with that model.

**Assumption A.3** (Transformation invariance). *Let \(P : \mathcal{E} \rightarrow [0, 1]\) be the joint distribution of a random vector. Let \(G\) be a set of transformations such that together with their natural composition operation of these transformation \((G, \cdot)\) forms a compact group. The distribution is invariant with respect to \(G\) if \(\forall E \subset \mathcal{E}, \forall g \in G \ F(E) = F(gE)\).*

The immediate consequence of the invariance assumption is that the Haar measure over \(G\) is finite, so uniform distribution can be defined over \(G\).
Definition A.4 (Well defined decreasing ordering with respect to a permutation).
The well defined decreasing ordering of values $Z_1, \ldots, Z_m$ with respect to a permutation $\pi$ of values $\{1, \ldots, m\}$ is another permutation $O$ of values $\{1, \ldots, m\}$ defined such that $i$ precedes $j$ in $O$ if 1. $Z_i > Z_j$ or if 2. $Z_i = Z_j$ and $i$ precedes $j$ in $\pi$.

The well defined decreasing ordering of real values is the usual ordering if there are no equal values to be ordered. If there are ties, then these ties are uniquely solved by the given permutation $\pi$.

Definition A.5 (Performance measure). A function can be considered as a performance measure if $Z : \Theta \times \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$.

Let $D = (Y, X)$ be the dataset available for parameter estimation with an assumed model structure $f$ that is invertible with respect to the noise. Moreover let $G$ be a set of transformations such that the noise distribution is invariant with respect to $G$. Let $\mathcal{A}$ be an arbitrary subset of all permutations of values $\{1, \ldots, m\}$ and let $\alpha(\mathcal{A}) = \frac{|\mathcal{A}|}{m!}$. Let $\theta$ be the model under test. Let $Z$ be an arbitrary performance measure.

Theorem A.6. Let the dataset $D = (Y, X)$ be generated by a model that satisfies Assumptions A.1 and A.2 with underlying model parameter $\theta_0$. Moreover the joint distribution of the contaminating noise is transformation invariant with respect to a group of transformations $G$. Let $\mathcal{A}$ be an arbitrary subset of all permutations of values $\{1, \ldots, m\}$.

Under these conditions

$$\mathbb{P}(\text{TestModel}(\theta_0, X, Y, \mathcal{A}) = \text{Accept}) = \alpha(\mathcal{A}) = \frac{|\mathcal{A}|}{m!}$$

where TestModel is defined in Algorithm A.1.

The SPS method is a data perturbation method where the joint distribution of the noise is assumed to be symmetric and the group of transformations consists of elementwise sign changes acting on vectors. The performance measure is defined using the norm of the gradient of the squared prediction error cost function.
Algorithm A.1 The data perturbation method

1: procedure TestModel($\theta, X, Y, A$)
2: if $f^*(\theta, X, Y)$ is not defined then
3: Reject $\theta$.
4: else
5: Calculate $E(\theta) = f^*(\theta, X, Y)$.
6: Select $g_2, \ldots, g_m$ random elements uniformly from $G$.
7: Create $m$ perturbed datasets
   $D^{(1)} = (Y, X)$
   $D^{(i)} = (f(\theta, X, g_iE(\theta)), X) \forall i = 2, \ldots, m$
8: Define the values $Z_i$ as the performance of the model $\theta$ on the different datasets
   $Z_i = Z(\theta, X, Y^{(i)}) \forall i = 1, \ldots, m$
   $Y^{(i)} = f(\theta, X, g_iE(\theta))$
9: Select a random permutation $\pi$ uniformly from all permutations of
   $\{1, \ldots, m\}$.
10: Calculate the well defined ordering $O$ of the values $Z_1, \ldots, Z_m$ with
    respect to $\pi$.
11: if $O \in A$ then
12: Accept $\theta$.
13: else
14: Reject $\theta$.
15: end if
16: end if
17: end procedure

Subthesis I.2: Data perturbation methods for linear regression problem

I studied data perturbation methods corresponding to the linear regression model structure. I have characterised a large class of distributions that facilitate creation of data perturbation methods that result in connected confidence regions with the appropriate performance measure. I have given sufficient conditions for the regressors that will ensure that the confidence regions are bounded.

Linear regression problems have the model structure

$$Y = f(\theta_0, X, E) = X^T\theta_0 + E, \quad Y, E \in \mathbb{R}^N, \quad X \in \mathbb{R}^{n_\theta \times N}, \quad \theta_0 \in \mathbb{R}^{n_\theta}$$
This model structure is clearly invertible with respect to noise

\[ E = f^*(\theta, X, Y) = Y - X^T \theta \]

**Definition A.7 (Unitary group).** The unitary group of order \( n \), denoted by \( U(n) \), is the group of \( n \times n \) unitary matrices with the regular matrix multiplication.

**Theorem A.8.** If the joint distribution \( F \) of the noise vector \( E \) is invariant under a subgroup \( G \) of \( U(n) \) then data perturbation methods can be constructed for the parameter vector \( \theta_0 \) that will result in connected confidence regions.

The performance measure corresponding to these tests is

\[ Z(\theta, X, Y) = ([XX^T]^{-1}X^TY - \theta)^T [XX^T] ([XX^T]^{-1}X^TY - \theta) \]

The accepted set of orderings \( A \) is selected such that permutations are included in \( A \) in increasing order with respect to the position of 1 in them until \( \alpha(A) \) reaches the desired level.

The SPS method is defined with two different performance measures in [Csáji et al., 2012a]. One of these corresponds to the performance measure given in Theorem A.8. I have given sufficient conditions on the regressors for both measures given in [Csáji et al., 2012a] that ensure boundedness of the confidence regions. I have shown that the other performance measure reported in [Csáji et al., 2012a] is not general in the sense that it results in bounded confidence regions only for a very restricted family of inputs.

**Subthesis I.3: Data perturbation methods for linear dynamical systems**

I defined data perturbation (DP) methods corresponding to linear dynamical time invariant systems parameter estimation. I suggested appropriate performance measures that can be seen as natural extensions of the performance measures used in the linear regression case. I have shown that the resulting DP methods, due to the non-linear dependence on the parameters, result in non-connected confidence regions. I examined the properties of the “central” component of the confidence region and suggested to use it as a good lower approximate confidence set.
The model structure is assumed to be of the following form

\[ A(z)y[k] = B(z)u[k] + C(z)e[k] \]

where \( A, B, C, D \) and \( F \) are finite order polynomials of the shift operator \( z^{-1} \) with orders \( n_A, n_B, n_C, n_D \) and \( n_F \) respectively. The measured part \( X \) of the dataset \( D \) contains the \( u[k] \) inputs of the plant model \( G(z) = B(z)A(z)F(z) \), the observables \( Y \) are the measured output values \( y[k] \) and the non-measured contaminating noise \( E \) is represented by the input \( e[k] \) of the noise model \( H(z) = C(z)A(z)D(z) \).

**Statement A.9 (Initial conditions and invertibility).** In order for this model to fulfil the invertibility Assumption A.2, the initial conditions for both the plant model \( G \) and noise model \( H \) need to be available.

The meaning of Statement A.9 is that the initial conditions are needed to be known, or the parameter vector \( \theta \) needs to contain those initial conditions that are not assumed in order to be able to construct DP methods for such models. This results in a parameter vector \( \theta \) that contains the coefficients of the polynomials \( A, B, C, D \) and \( F \) augmented with initial conditions that are not assumed.

Once the parameter vector \( \theta \) and our assumptions facilitate the creation of DP methods (Assumptions A.2 and A.3 are fulfilled) the only thing left to do is to provide suitable performance measures for the given model structure. I propose a more natural extension of the performance measure from the linear regression case than what is given in [Csáji et al., 2012b].

The performance measure in the linear regression case was

\[ Z(\theta, X, Y) = (XX^T)^{-1}X^TY - \theta \]

which can be interpreted in three different ways. One being the weighted norm of the gradient of the cost function, as used in [Csáji et al., 2012a], the second interpretation being the weighted norm of the difference between the least-squares estimate and \( \theta \). The third possible interpretation is the norm of the difference between the predictions given by the estimated and the tested model

\[ Z(\theta, X, Y) = (X^T[XX^T]^{-1}X^TY - X^T\theta)^T (X^T[XX^T]^{-1}X^TY - X^T\theta) \]
I extended this third interpretation of the linear regression performance measure to parameter estimation of the plant model. Let $D$ denote a perturbed dataset and $\hat{\theta}$ denote a parameter estimate corresponding to this dataset (it can be obtained by any means). Let $\hat{Y}_\theta$ denote the simulated output of the plant model from $\hat{\theta}$ for the input in $D$ and similarly let $\hat{Y}_\theta$ be the output of the plant model from $\theta$ on the same input.

In the linear regression case, when the least squares estimate is used as $\hat{\theta} = [XX^T]^{-1}X^TY$, the performance measure can be written as

$$Z(\theta, X, Y) = \| \hat{Y}_{[XX^T]^{-1}X^TY} - \hat{Y}_\theta \|^2$$

Statement A.10 (Performance measure for system models). Let the data set be composed of the input values $u[k]$ and observations $y[k]$. Let $\hat{\theta}$ denote an estimate of the parameter vector corresponding to this data. Using the performance measure

$$Z(\theta, U, Y) = \| \hat{Y}_\theta - \hat{Y}_\theta \|^2$$

in the 6th step of a data perturbation method results in confidence regions that contain $\hat{\theta}$. The connected component of the confidence region containing $\hat{\theta}$ is a good approximate confidence region for the plant model in the output error framework.

The performance measure defined above results in non connected confidence regions. This is true for all performance measures that come to mind, also for [Csáji et al., 2012b], and the root cause for this is that the problem is not linear in the parameters.

Statement A.10 defines a performance measure in case of output error models. This performance measure depends only on the plant model (as the noise model is assumed to be known $H(z) = 1$). I specified a performance measure for linear dynamical systems with more general noise models $H(z)$.

The performance measure suggested in the previous paragraph is the most natural choice but other meaningful choices are also presented in the thesis. These behave similarly in many respects. The resulting confidence regions are disconnected and are similar as long as the performance measure is sensible.
As given in the previous statements, I suggest the connected component around \( \hat{\theta} \) as a good lower approximate confidence region. I backed up this suggestion with different heuristic arguments showing that the other components of a confidence region that do not contain \( \hat{\theta} \) contribute a negligible part of the confidence.

**Thesis II:**

**Globally optimal estimation of polynomial models**

The following publications are related to this thesis: [6], [7], [8], [9],[10],[11],[12].

In many cases the identification problem corresponds to an optimization problem with more than one local extrema. Obtaining the globally optimal model is the goal in these situations. Although good enough local solutions are obtained in most cases but it is beneficiary to have theoretically certified global solutions if possible.

As described in the previous thesis, performance measures corresponding to DP methods for linear dynamical systems parameter estimation, require finding the point estimates \( \hat{\theta} \) corresponding to the perturbed data sets. This can be carried out using different techniques (such as prediction error minimizing, subspace or instrumental variable methods) and the approximate confidence regions around \( \hat{\theta} \) are constructed in a way that the estimate belonging to the primary data set will always be contained in the resulting set. This makes it important to obtain the estimates in a reliable way. For example, in case of prediction error minimization, if the optimization stops at a local minimum and the confidence set is constructed around that value, then the confidence of the method will be compromised.

For the reasons mentioned above I investigated ways to get global solutions to the identification problem of polynomial system models using the Lasserre hierarchy of semidefinite programming (SDP) relaxations [Lasserre, 2001].

This method can be used to solve general polynomial optimization problems with polynomial constraints. I investigated the properties of this approach with respect to the special properties of the optimization problems arising from identification of polynomial models. It is known that the general approach is computationally intractable for high order problems or problems in many variables, so different solutions were proposed to profit from the problem structure in order
to decrease the size of the resulting SDPs. I have shown that the optimization problem connected to system identification is sparse according to the definition of correlative sparsity. Based on this sparsity property, the resulting SDPs become numerically more tractable [Waki et al., 2006].

In general, the sparsity based problem reduction comes at the cost of losing convergence to the global solution. However, sufficient and necessary conditions are known for the correlative sparsity based reduction to preserve convergence to the solution of the original problem [Lasserre, 2006]. I have shown that these conditions are met in case of the identification problem.

The combined result is that a hierarchy of SDPs can be defined to approximate the global solution of the identification problem. The size of these SDPs can be reduced due the correlative sparsity structure of the problem in a way that the solution will still approximate the global solution of the original optimization problem. For every relaxation level, the size of the corresponding SDPs will grow only linearly with the sample count $N$.

The actual size of the solved SDP may vary with the used software package as these packages apply special tricks to enhance their precision or computational performance, but this difference is not significant.

**Subthesis II.1: Optimization without sparsity**

There are two basic parametrizations of the identification problem of polynomial models. First, when only the parameters of the model are used as unknowns. In this case the identification is a high order unconstrained optimization problem. Second, when auxiliary variables are introduced for the driving noise of the noise model resulting in a quadratic cost function with equality constraints. I gave explicit formulae to determine the size of the resulting SDPs in terms of the model complexity, sample count and relaxation number.

The model structure is assumed to be of the following form

$$A(z)y[k] = \frac{B(z)}{F(z)}u[k] + \frac{C(z)}{D(z)}e[k] + \frac{\lambda(z)}{D(z)}\delta[0] + \frac{\lambda(z)}{A(z)}\delta[0] \quad k \geq 1$$
where $A$, $B$, $C$, $D$ and $F$ are finite order polynomials of the shift operator $z^{-1}$ with orders $n_A$, $n_B$, $n_C$, $n_D$ and $n_F$ respectively. The polynomials $X$ are polynomials representing the initial conditions of the system. The polynomials with the exception of $B(z)$ and the initial conditions are monic. $u[k]$ are the input samples of the plant model $G(z) = \frac{B(z)}{A(z)F(z)}$, $y[k]$ are the measured output values and the non-measured contaminating noise $e[k]$ is the input of the noise model $H(z) = \frac{C(z)}{A(z)D(z)}$ and $\delta[k]$ is the unit impulse at time $k$.

**Definition A.11 (State-space models of rational transfer functions).** For a rational transfer function given as

$$P(z) = \frac{B(z)}{A(z)} = \tilde{d} + \frac{\sum_{k=1}^{n} \tilde{b}_k z^{-k}}{1 + \sum_{k=1}^{n} a_k z^{-k}}$$

a state space description can be given as

$$A = \begin{bmatrix} -a_1 & 1 & 0 & \cdots \\ \vdots & 0 & \ddots & 0 \\ \vdots & : & 0 & 1 \\ -a_n & 0 & \cdots & 0 \end{bmatrix}, \quad B = \begin{bmatrix} \tilde{b}_1 \\ \vdots \\ \tilde{b}_n \end{bmatrix}, \quad C = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}^T, \quad D = \tilde{d}$$

and the system output $Y \in \mathbb{R}^N$ corresponding to the input $U \in \mathbb{R}^N$ can be calculated as

$$Y = \Gamma_{B/A} \vec{x}_{B/A}[0] + \vec{H}_{B/A} U$$

where $\vec{x}_{B/A}[0] \in \mathbb{R}^n$ is the initial condition of the system, $\Gamma_{B/A} \in \mathbb{R}^{N \times n}$ is the extended observability matrix of the system

$$\Gamma_{B/A} = [C^T (CA)^T (CA^2)^T \cdots (CA^{N-1})^T]^T$$

and $\vec{H}_{B/A}$ is the lower triangular Toeplitz matrix with first column

$$\vec{H}_{B/A}(; 1) = [D \ CB \ CAB \ \cdots \ CA^{N-2}B]^T$$
By defining the column vectors \( U, Y \) and \( E \) to be constructed from the input, output and noise values, the model can be described as

\[
Y = \Gamma^\frac{1}{A} x_A[0] + H^\frac{1}{A} \left( \Gamma^\frac{g}{F} x^\frac{g}{F}[0] + H^\frac{g}{F} U + \Gamma^\frac{c}{F} x^\frac{c}{F}[0] + H^\frac{c}{F} E \right)
\]

The prediction errors corresponding to a given parameter vector \( \theta \) can be calculated as

\[
E(\theta) = H^\frac{e}{C} \left( H^\frac{A}{F} (Y - \Gamma^\frac{1}{A} x^\frac{1}{A}[0]) - \Gamma^\frac{g}{F} x^\frac{g}{F}[0] - H^\frac{g}{F} U - \Gamma^\frac{c}{F} x^\frac{c}{F}[0] \right)
\]

The unconstrained optimization problem corresponding to the least squares prediction error estimation problem is given as

\[
V_d(\theta) = \frac{1}{2N} E(\theta)^T E(\theta)
\]

where the parameter vector \( \theta \) contains the parameters of the state-space description of the systems involved and their initial condition if they are not assumed to be known.

One important measure of complexity for the identification problem is the degree of polynomials involved in expressing the prediction errors. Let the indicator variables \( 1 \cdot \) be defined as

\[
1_P = \begin{cases} 
1, & \text{the variable } P \text{ contains unknowns} \\
0, & \text{otherwise}
\end{cases}
\]

and let the degree \( \deg (\cdot) \) of a matrix \( M \) with polynomial entries be defined as the maximum of the degree of its elements.

\[
\deg (M) = \max_{i,j} \{ \deg (M(i,j)) \}
\]

The degrees of different terms appearing in the expression of the prediction error can be calculated as

\[
\deg \left( H^\frac{g}{F} \right) = 1_G + 1_A(N - 2) \quad \deg \left( \Gamma^\frac{g}{F} \right) = 1_A(N - 1)
\]
In general the SDP based solution in the $m$-th order relaxation involves at least an LMI of size $\binom{n_\theta + 2m}{2m}$, where $m$ needs to be greater than $\frac{\deg V_d}{2} = \deg \left( \hat{E} \right)$.

**Statement A.12** (Unconstrained general SDP sizes). The degree of the least squares cost function $\deg (V_d)$ can be calculated as

$$\frac{1}{2} \deg (V_d(\theta)) = 1_C + (N - 2)1_D + \max \left\{ 1_A + 1_C \left( 1 + (N - 1)1_A \right) \right\}$$

Depending on the model structure, the size of the SDP corresponding to the $m$ order relaxation of the optimization problem

$$\hat{\theta} = \arg \min V_d(\theta)$$

$$\|\theta\|^2 \leq R^2$$

can be calculated based on the number of unknowns and the degree of the polynomial cost function. There are two, full sized LMI constraints in the SDP formulation. Based on model structure complexity there are three different classes of problem difficulty.

1. Model complexity related difficulty class: it is characterized by $n_f = n_d = 0$ and no initial condition variables. The minimal relaxation order is $m_{\min} = 1$. The model complexity is $n_\theta = n_A + n_B + n_C$. The combinatorial factor of the SDP size growth depends only on $n_\theta$.

2. Model complexity and sample count related difficulty class: Contains problems where the minimal relaxation order depends on $1N$. This class covers models with $D = 1$. The minimal relaxation order increases with the sample count $N$.

3. Model complexity and double sample count related difficulty class: Contains problems where the minimal relaxation order depends on $2N$. Parameter
combinations not fitting into the previous two categories belong here. The minimal relaxation order increases with the double of the sample count $2N$.

I note that for the first class of problems, where the minimal relaxation order is not dependent on the sample count, the first approximation yields the global optimum.

For the other cases higher order relaxations are needed. As the relaxation order depends on the sample count $N$, the size of the corresponding SDPs will depend on it exponentially.

The other approach considers the problem with auxiliary variables corresponding to the noise samples $\varepsilon[k]$ and writing the cost function as

$$V_a(\theta, \varepsilon) = \frac{1}{2N} \varepsilon^T \varepsilon$$

where the parameter vector contains both the system parameters $\theta$ and the noise samples $\varepsilon[k]$ as well. Optimization of the cost function needs to be carried out under the equality constraint

$$A(z)F(z)D(z)y[k] = F(z)D(z)\mathcal{A}(z)\delta[0] +$$
$$+ D(z) (\mathcal{X}_F(z)\delta[0] + \mathcal{B}(z)u[k]) + F(z) (\mathcal{X}_D(z)\delta[0] + \mathcal{C}(z)\varepsilon[k]) \quad \forall k \in [N]$$

defined for every sample.

This results in an optimization problem with quadratic cost function, polynomial constraints, and variables whose number is linearly increasing with the sample count $N$, similarly to the number of constraints.

**Statement A.13** (Constrained general SDP sizes). Depending on the model structure, the size of the SDP corresponding to the $m$ order relaxation of the optimization problem

$$\hat{\theta} = \arg\min V_a(\theta, \varepsilon) = \frac{1}{2N} \sum_{k=1}^{N} \varepsilon^2[k]$$

$$A(z)F(z)D(z)y[k] = F(z)D(z)\mathcal{A}(z)\delta[0] +$$
$$+ D(z) (\mathcal{X}_F(z)\delta[0] + \mathcal{B}(z)u[k]) + F(z) (\mathcal{X}_D(z)\delta[0] + \mathcal{C}(z)\varepsilon[k]) \quad \forall k \in [N]$$

$$\|\theta\|^2 + \|\varepsilon\|^2 \leq R^2$$
can be calculated based on the degree of the equality constraints (both polynomial coefficients, initial condition and error variables should be counted). There are $2N + 2$, full sized LMI constraints in the formulated SDPs.

Based on model structure complexity there are three different classes of problem difficulty.

1. Linear constraints: The minimal relaxation order $m_{\text{min}} = 1$.

2. Quadratic constraints: The minimal relaxation order $m_{\text{min}} = 2$.

3. Cubic constraints: The minimal relaxation order $m_{\text{min}} = 3$.

In every case the number of variables in the optimization problem is $n_\theta + N$. This results in SDP sizes exponentially growing both in the sample count $N$ and the relaxation order $m$.

Apart from the simplest problem (that are in the form of linear regression) both approaches rely on solving SDPs with sizes that exponentially depend on the sample count $N$. This limits the applicability of these SDP relaxations to very simple systems and extremely short data records. It depends on the model structure which approach can be started with smaller SDPs.

**Subthesis II.2: Optimization with sparsity**

I have examined the correlative sparsity structure of the identification optimization problem with equality constraints. I have shown that the CSP graph corresponding to the problem can be extended to a chordal graph and I determined the maximal cliques of this graph. Based on this, I gave formulae to calculate the size of SDPs corresponding to the sparse relaxations. The main conclusion is that due to the structure of the optimization problem the size growth for a given relaxation order can be reduced to linear in the sample count.

The correlative sparsity pattern graph (CSP graph) of an optimization problem is defined as follows. Each variable in the optimization corresponds to a node in the graph. Two nodes are connected if they appear together in a monomial of the cost function or they appear together in a constraint [Waki et al., 2006].
Statement A.14 (Sparsity of the unconstrained optimization problem). If the sample count is greater than the number of identifiable parameters, then the CSP graph of the unconstrained optimization problem is fully connected, thus there is no sparsity in the problem.

Since the unconstrained optimization problem used the minimal number of decision variables, it was not expected that there is any sparsity left in the problem. However, with the introduction of the prediction error variables this changes drastically. The CSP graph of the identification optimization problem with equality constraints is given in Figure A.1.

![Figure A.1: Chordal extension of the CSP graph of the equality constrained identification POP](image_url)

Statement A.15 (Chordal extension of CSP graph). The CSP graph corresponding to the optimization problem with equality constraints has a small chordal extension. Its maximal cliques consist of the model parameters, initial conditions and as many noise variables as the system lag. The number of cliques is proportional to the sample count $N$ and their size is proportional to the model complexity.

Considerable computational efficiency is gained due to the fact that in the chordal extension of the CSP graph the maximal clique size is much smaller than the total number of variables. From the chordal structure of the graph it also follows that its maximal cliques satisfy the so called running intersection property. As proven in [Lasserre, 2006] the sparse SDP relaxations will converge to the global solution if the maximal cliques in the CSP graph satisfy the running intersection property and there is a known upper bound on the norm of the optimal solution.
**Statement A.16** (Constrained sparse SDP sizes). Depending on the model structure, the size of the SDP corresponding to the \( m \) order relaxation of the optimization problem

\[
\hat{\theta} = \arg\min_V \theta_y = \frac{1}{2N} \sum_{k=1}^{N} \epsilon^2[k]
\]

\[
A(z)F(z)D(z)y[k] = F(z)D(z)A(z)\delta[0] +
\]

\[
+D(z)(\mathcal{A}_F(z)\delta[0] + B(z)u[k]) + F(z)(\mathcal{A}_D(z)\delta[0] + C(z)\epsilon[k]) \quad \forall k \in [N]
\]

\[
\|\theta_C\|^2 + \|\epsilon_k\|^2 \leq R_k \quad k \in [p]
\]

can be calculated based on the degree of the equality constraints (both polynomial coefficients, initial condition and error variables should be counted). There are \( 2N + 2p \), reduced size sized LMI constraints in the formulated SDPs, where the size of each LMI is \( (|I|+m) \), where \( p = N - n_C - n_F + 1 \) and \( |I| \) is the maximal clique size of the chordal extension of the CSP graph.

Based on model structure complexity there are three different classes of problem difficulty, this classification is the same as in Statement A.13.

**Theorem A.17** (Convergent sparse relaxations). The polynomial optimization problem in Statement A.16 can be solved using the sparse SDP relaxation defined in [Lasserre, 2006]. The size of the SDPs needed to be solved will grow combinatorially with the model complexity and the relaxation order but it will grow only linearly in the sample count.

It is always better to use the formulation with the equality constraints. This results in LMI constraints involving only a portion of the SDP variables, allowing higher relaxation orders.

Based on the current understanding of polynomial optimization problems and the formulation of the Lasserre hierarchies, further reduction is not likely to be possible. With the currently available computational power the class of tractable problems is limited to model complexity of up to around ten model parameters to be estimated and thirty-fifty samples, using the second relaxation. Numerical results above this complexity become unreliable or cannot even be computed. Experience shows that in situations where the PEM method is trapped in a lo-
cal minimum even the estimates obtained by early relaxations, such as the first or second relaxations, can be used as starting points for PEM to obtain better estimate.

**Thesis III:**

**Effect of finite measurement length on uncertainty estimation**

The following publications are related to this thesis: [3], [13].

Asymptotic confidence regions for parameter estimates are used mostly for two reasons. They are quite reliable if enough data is available when they are constructed, and their structure is simple. In case of linear dynamical systems parameter estimates these are ellipsoids, which can be easily described.

The downside of such confidence regions is that their construction relies only on asymptotic properties of the estimation problem (such as input and noise spectra). The distribution of an unbiased estimate $\hat{\theta}_N$ based on $N$ samples of a nominal parameter vector $\theta_0$ can be described by the asymptotic limiting distribution

$$\sqrt{N}(\hat{\theta}_N - \theta_0) \overset{d}{\to} N(0, \Sigma_\infty) \quad \text{Var}(\hat{\theta}_N - \theta_0) = \frac{1}{N} \Sigma_\infty$$

with an appropriate covariance matrix $\Sigma_\infty$, if the problem satisfies certain regularity conditions. This rule can only account for the effects of the finite record length through the $1/N$ multiplication, but estimating the variance this way results in underestimation.

The goals of this thesis are: *i.* to illustrate the shortcomings of the $1/N$ rule; *ii.* to provide a procedure that is capable of characterizing the error in the variance estimate; *iii.* to suggest a modification of the regular asymptotic estimation procedure.

*I have defined the notion of finite sample variance loss that stands to characterize the difference between the actual variance of the finite sample estimate and its asymptotic estimate based on the $1/N$ rule. I provided an algorithm that can be used to approximate the variance loss for unbiased estimates of linear dynamical system parameters and I propose a heuristic to help decide whether the asymptotic
confidence region can be trusted or not. Based on this, the asymptotic variance estimate can be adjusted using the approximate or exact value of the variance loss to obtain a better approximation. The overall conclusion that I draw is that the variance loss remains significant for much higher sample counts than one might expect.

As in Thesis II, the model structure is assumed to be of the following form

\[
A(z)y[k] = \frac{B(z)}{F(z)} u[k] + \frac{\mathcal{X}_5(z)}{\mathcal{F}(z)} \delta[0] + \frac{\mathcal{C}(z)}{\mathcal{D}(z)} e[k] + \frac{\mathcal{X}_4(z)}{\mathcal{D}(z)} \delta[0] + \mathcal{X}_1(z) \delta[0] \quad k \geq 1
\]

where \(A, B, C, D\) and \(F\) are finite order polynomials of the shift operator \(z^{-1}\) with orders \(n_A, n_B, n_C, n_D\) and \(n_F\) respectively. The polynomials \(\mathcal{X}_i\) are polynomials representing the initial conditions of the system. The polynomials with the exception of \(B(z)\) and the initial conditions are monic. \(u[k]\) are the input samples of the plant model \(G(z) = \frac{B(z)}{A(z)F(z)}\), \(y[k]\) are the measured output values and the non-measured contaminating noise \(e[k]\) is the input of the noise model \(H(z) = \frac{C(z)}{A(z)D(z)}\) and \(\delta[k]\) is the unit impulse at time \(k\). The input \(u[\cdot]\) and noise \(e[\cdot]\) are assumed to be independent.

In order to be able to compare properties of estimates obtained from increasing amounts of data, this data should be acquired in a stationary manner, by continuous time invariant excitation. The input signal can be a deterministic periodic signal, pseudo-random binary signal or the realization of stationary stochastic process. Throughout this thesis the calculations are done assuming white or filtered Gaussian noise input, but the method can be adjusted to other input signals as well. The estimate \(\hat{\theta}_N\) is obtained as the least squared prediction error estimate.

By considering a sample third order system with output error noise model, the parameter vector \(\theta = [b_1 \ b_2 \ b_3 \ f_1 \ f_2 \ f_3]^T\), with nominal parameter vector \(\theta_0 = [0.0880 \ 0.0195 \ -0.0442 \ -2.0536 \ 1.4611 \ -0.3442]^T\). The time constant of this system is around 5 samples. Figure A.2 shows the expected variance of the estimates of these parameters after compensating the \(1/N\) decay (averaged over different input and noise realizations). The example illustrates some of the major characteristics of the variance estimates: \(i.\) the asymptotic estimate underestimates the real variance; \(ii.\) the difference between the empirical and the
asymptotic values can be very significant for small values of the sample count $N$;

iii. the error decays differently for numerator and denominator coefficients.

![Figure A.2: Normalized variances of parameters against sample count.](image)

(a) Normalized variance of $\hat{b}_1, \hat{b}_2, \hat{b}_3$

(b) Normalized variance of $\hat{f}_1, \hat{f}_2, \hat{f}_3$

**Definition A.18** (Finite sample variance loss). For a scalar parameter $\theta$, let $
abla^2_{N|\theta}({\hat{\theta}_N})$ denote the variance of the parameter estimate $\hat{\theta}_N$ obtained using $N$ samples, conditioned on the realization of the input $u$. The asymptotic variance is defined as $\nabla^2_{\infty|u}({\hat{\theta}}) = \lim_{N \to \infty} N\nabla^2_{N|u}({\hat{\theta}_N})$. The finite sample variance loss for $N$ samples is defined as

$$\nabla_{N|u}({\hat{\theta}_N}) = \frac{N\nabla^2_{N|u}({\hat{\theta}_N})}{\nabla^2_{\infty|u}({\hat{\theta}})}$$

The expected finite sample variance loss for $\theta$ is defined as

$$\overline{\nabla}_{N}({\hat{\theta}_N}) = \mathbb{E}_{U} \left( \nabla_{N|u}({\hat{\theta}_N}) \right)$$

In the case when multiple different realizations of the input $U$ correspond to the same asymptotic properties (i.e. multi-sine signal with different initial phase but common spectra, realizations of a stationary stochastic process) the expected variance loss $\overline{\nabla}_{N}$ captures the part of the variance loss that is characteristic to the asymptotic properties of the input, while $\nabla_{N|u}$ depends on the chosen realization of the input $U$. 
Being able to calculate the variance loss exactly is equivalent in difficulty to being able to characterize the variance of the estimates exactly. This is a difficult and analytically intractable problem in most cases, this is one of the reasons why asymptotic estimates are used for finite sample variance estimation. My goal is to present an approximation procedure for the variance loss that can be used in the case when the contaminating noise $E$ is white Gaussian and the estimate $\hat{\theta}_N$ is unbiased. This approximation is computationally more tractable than the exact problem and provides reasonable results.

Introducing the row vectors $\phi[k] \in \mathbb{R}^{1 \times n_\theta} = \frac{\partial \hat{y}[k]}{\partial \theta} \bigg|_{\theta=\theta_0}$, the matrix $\Phi \in \mathbb{R}^{N \times n_\theta} = \{\phi[k]\}_{k=1}^N$ and using the assumption that the estimate $\hat{\theta}_N$ is unbiased, it is known that its covariance is bounded by the Cramér-Rao lower bound, that is

$$
\mathbb{E} \left( (\hat{\theta}_N - \theta_0)(\hat{\theta}_N - \theta_0)^T \bigg| U \right) \geq M_U^{-1}
$$

where

$$M_U = \frac{1}{\sigma_y^2} \mathbb{E} \left( \Phi^T \Phi \bigg| U \right)$$

Algorithm A.2 is a moment matching based approximation method that can be used to evaluate the expected variance loss $\bar{\alpha}_N((\hat{\theta}_N)_i)$.

**Algorithm A.2 Approximating the expected variance loss**

1. Calculate the covariance matrix $\Psi = \mathbb{E}(\phi^T[N]\phi[N])$.
2. Let $X \in \mathbb{R}^{\kappa N \times n_\theta}$ be a matrix whose rows are Gaussian random vectors with covariance $\Psi$.
3. Determine the degrees of freedom $\kappa N$ by means of moment matching between $X^T X$ and $\Phi^T \Phi$.
4. Approximate the expected finite sample variance loss as

$$\bar{\alpha}_N((\hat{\theta}_N)_i) \approx \frac{1}{\sigma_\infty^2((\hat{\theta}_N)_i)} \frac{N}{\kappa N - n_\theta - 1} (\Psi^{-1})_{i,i}$$

The moment matching in the third step can be realized by finding the root of a scalar function $f(X^T X, \Phi^T \Phi)$. For one dimensional problems the trivial choice is the solution of the equation $\mathbb{E}(X^T X - \Phi^T \Phi) = 0$. For higher dimensional problems
matching the expectation of the $i$-th diagonal term is a viable choice, or matching the determinants as $\mathbb{E}(|X^T X|) = \mathbb{E}(|\Phi^T \Phi|)$.

![Graph](image)

(a) Numerator parameters $b_i$.

(b) Denominator parameters $f_i$.

Figure A.3: Empirical and approximate expected variance loss of parameters. The curves for the empirical values have circle markers, while those corresponding to values obtained by the proposed approximation are marked with squares. Lines with the same color correspond to the same parameter.

Figure A.3 shows the behavior of the expected variance loss with respect to sample count for the example system.

Multiple consequences can be drawn from the behaviour of the expected variance loss. The distribution of the variance loss is not concentrated on a single value for small values of $N$, thus it can be above or under the expected value. This shows that, when choosing input signals for short experiments, the emphasis should shift from designing asymptotic properties of a signal towards the design of the actual realization.

The variance loss remains significant for quite long, 100 major time constants in the illustration example.

Before conducting an identification experiment, the length of the data record can be selected such that either the approximate variance of the estimates or their variance losses goes below a prescribed bound. This will result in longer experiments than if the same condition is fulfilled but with the asymptotic variance.

After the data is collected and the estimation is carried out, the corresponding asymptotic variance estimates can be adjusted to account for the small sample variance loss. The variance loss $\alpha_{N\mid U}(\hat{\theta}_N)_i$ of each parameter $\theta_i$ can be approximated...
(based on the presented approximation algorithm, or based on multiple simulations). Assuming that the asymptotic covariance matrix of the estimate is $\Sigma_\infty$, a rescaled finite sample covariance matrix can be obtained by transforming the eigenvectors of $\Sigma_\infty$ as follows. Let $\alpha \in \mathbb{R}^{n_\theta}$ be the vector containing the finite sample variance losses of the individual parameters $\left(\alpha_{N\mid U}(\hat{\Theta}_N)_i\right)^{n_\theta}_{i=1}$ and let the rescaling operator $r(v, \alpha)$ be defined as $r(v, \alpha) : \mathbb{R}^{n_\theta} \times \mathbb{R}^{n_\theta} \rightarrow \mathbb{R}^{n_\theta}; (r(v, \alpha))_i = v_i \sqrt{\alpha_i}$. The finite sample covariance matrix is obtained as

$$\Sigma_N = N \sum_{i=1}^{n_\theta} \lambda_i r(v_i, \alpha) r^T(v_i, \alpha)$$

where $\lambda_i$ and $v_i$ are the eigenvalues and eigenvectors of $\Sigma_\infty$. This method provides a multidimensional rescaling of the asymptotic covariance matrix, thus maintaining a simple description for the uncertainty region while incorporating the finite sample variance loss.
Appendix B

Variance of Filtered White Noise Output

This appendix contains the calculation of the variance of the output of a system whose input is filtered white noise. We are assuming that input of the system $u[k]$ is the output of a filter $F(z)$ driven by white noise samples $e[k]$.

**Statement B.1** (Variance of output). The variance of the output $y[k]$ of the system $G(z)$ driven by filtered white noise created using the filter $F(z)$ driven by zero mean white noise samples $e[k]$ can be given as

$$
\mathbb{E}(y^2[k]) = \sigma^2 \left( \sum_{c=0}^{k-T_y} \left[ \sum_{\ell=0}^{c} gf_{c-\ell} \right]^2 + \sum_{c=k-T_u}^{k-T_y} \left[ \sum_{\ell=0}^{k-T_y} gf_{c-\ell} \right]^2 \right) 
$$  \hspace{1cm} (B.1)

where $\sigma^2$ is the variance of the driving noise signal $e[k]$, $T_u$ is the last time instance when the state of the input filter $F(z)$ is zero and $T_y$ is last time when the state of the plant $G(z)$ is zero. $T_u \leq T_y$ is assumed, i.e. the input signal is ‘on’ before the plant model starts to use it. The impulse responses of the systems $G$ and $F$ are denoted by $g_k$ and $f_k$ respectively.

The parameters $T_u$ and $T_y$ allow setting the initial condition of the systems. If these values are set to 0 then output can appear not earlier than $k = 0$. If $T_u = -\infty$ and $T_y = 0$ then the input signal is stationary at time 0, and the plant $G(z)$ has zero initial condition. If both values are $-\infty$ then the input and the
output can be considered to be already in a stationary state. The proof is straightforward calculation.

**Proof.** With the parameterization, the input signal of the plant $G$ can be written as

\[ u[k] = \sum_{\ell=0}^{k-T_u} f_\ell e[k-\ell] \quad k \geq T_u \tag{B.2} \]

while its output is given by

\[ y[k] = \sum_{\ell=0}^{k-T_y} g_\ell u[k-\ell] \quad k \geq T_y \tag{B.3} \]

The variance of the output sample $y[k]$ can be calculated as

\[ \mathbb{E}(y^2[k]) = \mathbb{E} \left( \sum_{\ell=0}^{k-T_y} g_\ell u[k-\ell] \right)^2 = \mathbb{E} \left( \sum_{\ell=0}^{k-T_y} g_\ell \sum_{h=0}^{k-\ell-T_u} f_h e[k-\ell-h] \right)^2 \]

We need to be able to group the coefficients of the noise values $e[k]$, the groups are characterized by the equality $\ell + h = c$ for all possible values of $c$.

\[ \mathbb{E} \left( \sum_{\ell=0}^{k-T_y} \sum_{h=0}^{k-\ell-T_u} g_\ell f_h e[k-\ell-h] \right)^2 \]

\[ = \mathbb{E} \left( \sum_{c=0}^{k-T_y} e[k-c] \left( \sum_{\ell=0}^{c} g_\ell f_{c-\ell} \right) + \sum_{c=k-T_y+1}^{k-T_u} e[k-c] \left( \sum_{\ell=0}^{k-T_y} g_\ell f_{c-\ell} \right) \right)^2 \]

The statement to be proven follows using the fact that the two sums cover different ranges for $c$ and the fact that the noise values $e[k]$ have zero mean and are independent.
Bibliography

List of publications related to the thesis

Papers published in international journals


Papers published in conference proceedings


**List of publications not related to the thesis**

**Papers published in international journals**


**Papers published in conference proceedings**


**Independent citations**


System identification has gone a long way in the past century but there are still problems to be solved and new directions waiting to be discovered. The mainstream identification literature contains results that rely on practically infinite amount of data (either measurement data or prior knowledge), which is an acceptable compromise for many applications, but not for all.

The focus in this thesis is put on obtaining statistically reliable uncertainty estimates when the identification task is carried out in a non-informative environment, meaning that the number of observations and that of the estimated parameters is barely enough to make the estimation problem meaningful. As this problem is tackled, a series of questions emerge that serve as the subject of the thesis.

First, a recent method (Sign-Perturbed Sums, SPS) for constructing exact confidence regions is described from a new point of view by defining a generic family of such methods (Data Perturbation methods, DP methods). Exact confidence level hypothesis tests can be constructed in the framework of DP methods where some of the elements of a DP method are determined by the assumed characteristics of the noise that contaminates the measurements. The other elements need to be defined to suit the model structure that is estimated. It is shown that the SPS method is a DP method belonging to symmetrically distributed noise sources. Other DP methods are also given and the structure of the corresponding confidence sets is examined.

DP methods for parameter estimates of linear dynamical systems rely on the solution of polynomial optimization problems. The second part of the thesis deals with how the optimum of such problems can be obtained in light of recent developments in the optimization community.

As computing exact confidence regions turns out to be computationally taxing, it needs to be quantified when the asymptotic results can be trusted. The last part of the thesis describes a method that estimates the uncertainty error introduced by asymptotic results. Based on these estimates it can be decided whether the asymptotic confidence regions are to be trusted or some sort of compensation is needed.