Vrije Universiteit Brussel

Budapest University of Technology and Economics
Faculty of Electrical Engineering and Informatics
Department of Automation and Applied Informatics

Vrije Universiteit Brussel
Faculty of Engineering
Department of Fundamental Electricity and Instrumentation

System Identification in Highly Non-Informative Environment

Rendszeridentifikáció Erősen Információhiányos Környezetben

Ph.D. Thesis Booklet

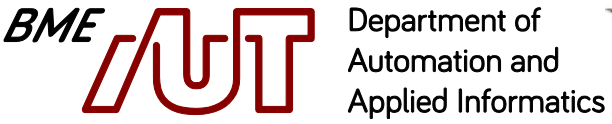
Sándor Kolumbán

11 March 2016

Advisors:

Prof. István Vajk

Prof. Johan Schoukens



Department of Automation and Applied Informatics

Vrije Universiteit Brussel
Departement ELEC

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 [CCGW12]. 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.

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 [CCW12a] 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 [CW05, CCW12a] 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 [CCW12a] 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 [CCW12a] 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 1 (Model structure assumption) *We assume that the measurements come from a fixed model structure described in the following form*

$$Y = f(\boldsymbol{\theta}_0, X, E)$$

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

Assumption 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 $H_0 : \boldsymbol{\theta}_0 = \boldsymbol{\theta}$ and alternative hypothesis $H_1 : \boldsymbol{\theta}_0 \neq \boldsymbol{\theta}$.

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

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

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 $\boldsymbol{\theta}$, if the data can be generated with that model.

Assumption 3 (Transformation invariance) *Let $\mathbb{P} : \mathcal{E} \to [0, 1]$ 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 1 (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 π 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 π.*

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 π.

Definition 2 (Performance measure) *A function can be considered as a performance measure if $Z : \Theta \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.*

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 $\boldsymbol{\theta}$ be the model under test. Let Z be an arbitrary performance measure.

Theorem 1 *Let the dataset $D = (Y, X)$ be generated by a model that satisfies Assumptions 1 and 2 with underlying model parameter $\boldsymbol{\theta}_0$. 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

$$\mathbb{P}(\textsc{TestModel}(\boldsymbol{\theta}_0, X, Y, \mathcal{A}) = \mathbf{Accept}) = \alpha(\mathcal{A}) = \frac{|\mathcal{A}|}{\mathbf{m!}}$$

where TESTMODEL *is defined in Algorithm 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 1 The data perturbation method

procedure TESTMODEL($\boldsymbol{\theta}, X, Y, \mathcal{A}$)
  if $f^*(\boldsymbol{\theta}, X, Y)$ is not defined **then**
    Reject $\boldsymbol{\theta}$.
  else
    Calculate $E(\boldsymbol{\theta}) = f^*(\boldsymbol{\theta}, X, Y)$.
    Select $g_2, \ldots, g_m$ random elements uniformly from G.
    Create m perturbed datasets

$$D^{(1)} = (Y, X) \qquad D^{(i)} = (f(\boldsymbol{\theta}, X, g_i E(\boldsymbol{\theta})), X) \; \forall i = 2, \ldots, m$$

    Define the values Z_i as the performance of the model $\boldsymbol{\theta}$ on the different datasets

$$Z_i = Z(\boldsymbol{\theta}, X, Y^{(i)}) \; \forall i = 1, \ldots, m \qquad Y^{(i)} = f(\boldsymbol{\theta}, X, g_i E(\boldsymbol{\theta}))$$

    Select a random permutation π uniformly from all permutations of $\{1, \ldots, m\}$.
    Calculate the well defined ordering O of the values $Z_1, \ldots, Z_m$ with respect to π.
    if $O \in \mathcal{A}$ **then**
      Accept $\boldsymbol{\theta}$.
    else
      Reject $\boldsymbol{\theta}$.
    end if
  end if
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

$$Y = f(\boldsymbol{\theta}_0, X, E) = X^T \boldsymbol{\theta}_0 + E, \qquad Y, E \in \mathbb{R}^N, \; X \in \mathbb{R}^{n_{\boldsymbol{\theta}} \times N}, \; \boldsymbol{\theta}_0 \in R^{n_{\boldsymbol{\theta}}}$$

This model structure is clearly invertible with respect to noise

$$E = f^*(\boldsymbol{\theta}, X, Y) = Y - X^T \boldsymbol{\theta}$$

Definition 3 (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 2 *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 $\boldsymbol{\theta}_0$ that will result in connected confidence regions.*

The performance measure corresponding to these tests is

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

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 level.

The SPS method is defined with two different performance measures in [CCW12a]. One of these corresponds to the performance measure given in Theorem 2. I have given sufficient conditions on the regressors for both measures given in [CCW12a] that ensure boundedness of the confidence regions. I have shown that the other performance measure reported in [CCW12a] 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

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

where $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, $\mathcal{D}$ and $\mathcal{F}$ are finite order polynomials of the shift operator z^{-1} with orders $n_{\mathcal{A}}$, $n_{\mathcal{B}}$, $n_{\mathcal{C}}$, $n_{\mathcal{D}}$ and $n_{\mathcal{F}}$ respectively. The measured part $\mathcal{X}$ of the dataset D contains the $u[k]$ inputs of the plant model $G(z) = \frac{\mathcal{B}(z)}{\mathcal{A}(z)\mathcal{F}(z)}$, the observables $\mathcal{Y}$ are the measured output values $y[k]$ and the non-measured contaminating noise $\mathcal{E}$ is represented by the input $e[k]$ of the noise model $H(z) = \frac{\mathcal{C}(z)}{\mathcal{A}(z)\mathcal{D}(z)}$.

Statement 1 (Initial conditions and invertibility) *In order for this model to fulfil the invertibility Assumption 2, the initial conditions for both the plant model G and noise model H need to be available.*

The meaning of Statement 1 is that the initial conditions are needed to be known, or the parameter vector $\boldsymbol{\theta}$ 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 $\boldsymbol{\theta}$ that contains the coefficients of the polynomials $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, $\mathcal{D}$ and $\mathcal{F}$ augmented with initial conditions that are not assumed.

Once the parameter vector $\boldsymbol{\theta}$ and our assumptions facilitate the creation of DP methods (Assumptions 2 and 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 [CCW12b].

The performance measure in the linear regression case was

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

which can be interpreted in three different ways. One being the weighted norm of the gradient of the cost function, as used in [CCW12a], the second interpretation being the weighted norm

of the difference between the least-squares estimate and $\boldsymbol{\theta}$. The third possible interpretation is the norm of the difference between the predictions given by the estimated and the tested model

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

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 $\hat{\boldsymbol{\theta}}$ denote a parameter estimate corresponding to this dataset (it can be obtained by any means). Let $\hat{Y}_{\hat{\boldsymbol{\theta}}}$ denote the simulated output of the plant model from $\hat{\boldsymbol{\theta}}$ for the input in D and similarly let $\hat{Y}_{\boldsymbol{\theta}}$ be the output of the plant model from $\boldsymbol{\theta}$ on the same input.

In the linear regression case, when the least squares estimate is used as $\hat{\boldsymbol{\theta}} = [XX^T]^{-1}X^TY$, the performance measure can be written as

$$Z(\boldsymbol{\theta}, X, Y) = \left\| \hat{Y}_{[XX^T]^{-1}X^TY} - \hat{Y}_{\boldsymbol{\theta}} \right\|^2$$

Statement 2 (Performance measure for system models) *Let the data set be composed of the input values $u[k]$ and observations $y[k]$. Let $\hat{\boldsymbol{\theta}}$ denote an estimate of the parameter vector corresponding to this data. Using the performance measure*

$$Z(\boldsymbol{\theta}, U, Y) = \left\| \hat{Y}_{\hat{\boldsymbol{\theta}}} - \hat{Y}_{\boldsymbol{\theta}} \right\|^2$$

in the 6th step of a data perturbation method results in confidence regions that contain $\hat{\boldsymbol{\theta}}$. The connected component of the confidence region containing $\hat{\boldsymbol{\theta}}$ 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 [CCW12b], and the root cause for this is that the problem is not linear in the parameters.

Statement 2 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 $\hat{\boldsymbol{\theta}}$ 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 $\hat{\boldsymbol{\theta}}$ 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 $\hat{\boldsymbol{\theta}}$ 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 $\hat{\boldsymbol{\theta}}$ 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 [Las01].

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 [WKKM06].

In general, the sparsity based problem reduction comes at the cost of loosing 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 [Las06]. 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

$$\mathcal{A}(z)y[k] = \frac{\mathcal{B}(z)}{\mathcal{F}(z)}u[k] + \frac{\mathcal{X}_{\frac{\mathcal{B}}{\mathcal{F}}}(z)}{\mathcal{F}(z)}\delta[0] + \frac{\mathcal{C}(z)}{\mathcal{D}(z)}e[k] + \frac{\mathcal{X}_{\frac{\mathcal{C}}{\mathcal{D}}}(z)}{\mathcal{D}(z)}\delta[0] + \mathcal{X}_{\frac{1}{\mathcal{A}}}(z)\delta[0] \quad k \geq 1$$

where $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, $\mathcal{D}$ and $\mathcal{F}$ are finite order polynomials of the shift operator z^{-1} with orders $n_{\mathcal{A}}$, $n_{\mathcal{B}}$, $n_{\mathcal{C}}$, $n_{\mathcal{D}}$ and $n_{\mathcal{F}}$ respectively. The polynomials $\mathcal{X}_{\cdot}$ are polynomials representing the initial conditions of the system. The polynomials with the exception of $\mathcal{B}(z)$ and the initial conditions are monic. $u[k]$ are the input samples of the plant model $G(z) = \frac{\mathcal{B}(z)}{\mathcal{A}(z)\mathcal{F}(z)}$, $y[k]$ are the measured output values and the non-measured contaminating noise $e[k]$ is the input of the noise model $H(z) = \frac{\mathcal{C}(z)}{\mathcal{A}(z)\mathcal{D}(z)}$ and $\delta[k]$ is the unit impulse at time k.

Definition 4 (State-space models of rational transfer functions) *For a rational transfer function given as*

$$P(z) = \frac{\mathcal{B}(z)}{\mathcal{A}(z)} = \tilde{d} + \frac{\sum_{k=1}^{n}\tilde{b}_k z^{-k}}{1 + \sum_{k=1}^{n} a_k z^{-k}}$$

a state space description can be given as

$$A = \begin{bmatrix} -a_1 & 1 & 0 & \cdots \\ \vdots & 0 & \ddots & 0 \\ \vdots & \vdots & 0 & 1 \\ -a_n & 0 & \cdots & 0 \end{bmatrix} \quad B = \begin{bmatrix} \tilde{b}_1 \\ \vdots \\ \tilde{b}_n \end{bmatrix} \quad C = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}^T \quad D = \tilde{d}$$

and the system output $Y \in \mathbb{R}^N$ corresponding to the input $U \in \mathbb{R}^N$ can be calculated as

$$Y = \boldsymbol{\Gamma}_{\frac{\mathcal{B}}{\mathcal{A}}}\boldsymbol{x}_{\frac{\mathcal{B}}{\mathcal{A}}}[0] + \boldsymbol{H}_{\frac{\mathcal{B}}{\mathcal{A}}}U$$

where $\boldsymbol{x}_{\frac{\mathcal{B}}{\mathcal{A}}}[0] \in \mathbb{R}^n$ is the initial condition of the system, $\boldsymbol{\Gamma}_{\frac{\mathcal{B}}{\mathcal{A}}} \in \mathbb{R}^{N\times n}$ is the extended observability matrix of the system

$$\boldsymbol{\Gamma}_{\frac{\mathcal{B}}{\mathcal{A}}} = \begin{bmatrix} C^T & (CA)^T & (CA^2)^T & \dots & (CA^{N-1})^T \end{bmatrix}^T$$

and $\boldsymbol{H}_{\frac{\mathcal{B}}{\mathcal{A}}}$ is the lower triangular Toeplitz matrix with first column

$$\boldsymbol{H}_{\frac{\mathcal{B}}{\mathcal{A}}}(:,1) = \begin{bmatrix} D & CB & CAB & \cdots & CA^{N-2}B \end{bmatrix}^T$$

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

$$Y = \boldsymbol{\Gamma}_{\frac{1}{\mathcal{A}}}\boldsymbol{x}_{\frac{1}{\mathcal{A}}}[0] + \boldsymbol{H}_{\frac{1}{\mathcal{A}}}\left(\boldsymbol{\Gamma}_{\frac{\mathcal{B}}{\mathcal{F}}}\boldsymbol{x}_{\frac{\mathcal{B}}{\mathcal{F}}}[0] + \boldsymbol{H}_{\frac{\mathcal{B}}{\mathcal{F}}}U + \boldsymbol{\Gamma}_{\frac{\mathcal{C}}{\mathcal{D}}}\boldsymbol{x}_{\frac{\mathcal{C}}{\mathcal{D}}}[0] + \boldsymbol{H}_{\frac{\mathcal{C}}{\mathcal{D}}}E\right)$$

The prediction errors corresponding to a given parameter vector $\boldsymbol{\theta}$ can be calculated as

$$E(\boldsymbol{\theta}) = \boldsymbol{H}_{\frac{\mathcal{D}}{\mathcal{C}}}\left(\boldsymbol{H}_{\frac{\mathcal{A}}{1}}\left(Y - \boldsymbol{\Gamma}_{\frac{1}{\mathcal{A}}}\boldsymbol{x}_{\frac{1}{\mathcal{A}}}[0]\right) - \boldsymbol{\Gamma}_{\frac{\mathcal{B}}{\mathcal{F}}}\boldsymbol{x}_{\frac{\mathcal{B}}{\mathcal{F}}}[0] - \boldsymbol{H}_{\frac{\mathcal{B}}{\mathcal{F}}}U - \boldsymbol{\Gamma}_{\frac{\mathcal{C}}{\mathcal{D}}}\boldsymbol{x}_{\frac{\mathcal{C}}{\mathcal{D}}}[0]\right)$$

The unconstrained optimization problem corresponding to the least squares prediction error estimation problem is given as

$$V_d(\boldsymbol{\theta}) = \frac{1}{2N} E(\boldsymbol{\theta})^T E(\boldsymbol{\theta})$$

where the parameter vector $\boldsymbol{\theta}$ 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 $\mathbb{1}_{\cdot}$ be defined as

$$\mathbb{1}_{\mathcal{P}} = \begin{cases} 1, & \text{the variable } \mathcal{P} \text{ contains unknowns} \\ 0, & \text{otherwise} \end{cases}$$

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

$$\deg\left(\boldsymbol{H}_{\frac{\mathcal{A}}{\mathcal{A}}}\right) = \mathbb{1}_{\mathcal{B}} + \mathbb{1}_{\mathcal{A}}(N-2) \quad \deg\left(\boldsymbol{\Gamma}_{\frac{\mathcal{A}}{\mathcal{A}}}\right) = \mathbb{1}_{\mathcal{A}}(N-1)$$

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

Statement 3 (Unconstrained general SDP sizes) *The degree of the least squares cost function* $\deg(V_d)$ *can be calculated as*

$$\begin{aligned} \frac{1}{2}\deg(V_d(\boldsymbol{\theta})) = \mathbb{1}_{\mathcal{C}} + (N-2)\mathbb{1}_{\mathcal{D}} + \max\Big\{ & \mathbb{1}_{\mathcal{A}} + \mathbb{1}_{\mathcal{X}_{\frac{1}{\mathcal{A}}}}(1 + (N-1)\mathbb{1}_{\mathcal{A}}) \\ & \mathbb{1}_{\mathcal{X}_{\frac{\mathcal{B}}{\mathcal{F}}}}(1 + (N-1)\mathbb{1}_{\mathcal{F}}), \\ & \mathbb{1}_{\mathcal{B}} + (N-2)\mathbb{1}_{\mathcal{F}}, \\ & \mathbb{1}_{\mathcal{X}_{\frac{\mathcal{C}}{\mathcal{D}}}}(1 + (N-1)\mathbb{1}_{\mathcal{D}}) \Big\} \end{aligned}$$

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

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \arg\min V_d(\boldsymbol{\theta}) \\ \|\boldsymbol{\theta}\|^2 &\leq R^2 \end{aligned}$$

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* $n_{\boldsymbol{\theta}} = n_{\mathcal{A}} + n_{\mathcal{B}} + n_{\mathcal{C}}$*. The combinatorial factor of the SDP size growth depends only on* $n_{\boldsymbol{\theta}}$*.*

2. *Model complexity and sample count related difficulty class: Contains problems where the minimal relaxation order depends on $1N$. This class covers models with $\mathcal{D} = 1$. 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

$$V_a(\boldsymbol{\theta}, \boldsymbol{\varepsilon}) = \frac{1}{2N}\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}$$

where the parameter vector contains both the system parameters $\boldsymbol{\theta}$ and the noise samples $\varepsilon[k]$ as well. Optimization of the cost function needs to be carried out under the equality constraint

$$\begin{aligned}\mathcal{A}(z)\mathcal{F}(z)\mathcal{D}(z)y[k] = F(z)\mathcal{D}(z)\mathcal{X}_{\mathcal{A}}(z)\delta[0]+\\ +\mathcal{D}(z)\left(\mathcal{X}_{\mathcal{F}}(z)\delta[0] + \mathcal{B}(z)u[k]\right) + \mathcal{F}(z)\left(\mathcal{X}_{\mathcal{D}}(z)\delta[0] + \mathcal{C}(z)\varepsilon[k]\right) \quad \forall k \in [N]\end{aligned}$$

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 4 (Constrained general SDP sizes) *Depending on the model structure, the size of the SDP corresponding to the m order relaxation of the optimization problem*

$$\begin{aligned}\hat{\boldsymbol{\theta}} = \operatorname{argmin} V_a(\boldsymbol{\theta}, \boldsymbol{\varepsilon}) =& \frac{1}{2N}\sum_{k=1}^{N}\varepsilon^2[k]\\ \mathcal{A}(z)\mathcal{F}(z)\mathcal{D}(z)y[k] =& \mathcal{F}(z)\mathcal{D}(z)\mathcal{X}_{\mathcal{A}}(z)\delta[0]+\\ &+\mathcal{D}(z)\left(\mathcal{X}_{\mathcal{F}}(z)\delta[0] + \mathcal{B}(z)u[k]\right) + \mathcal{F}(z)\left(\mathcal{X}_{\mathcal{D}}(z)\delta[0] + \mathcal{C}(z)\varepsilon[k]\right) \quad \forall k \in [N]\\ \|\boldsymbol{\theta}\|^2 + \|\boldsymbol{\varepsilon}\|^2 \leq& R^2\end{aligned} \tag{1}$$

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 $m_{\min} = 1$.*

2. *Quadratic constraints: The minimal relaxation order $m_{\min} = 2$.*

3. *Cubic constraints: The minimal relaxation order $m_{\min} = 3$.*

In every case the number of variables in the optimization problem is $n_{\boldsymbol{\theta}} + N$. 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 [WKKM06].

Statement 5 (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 1.

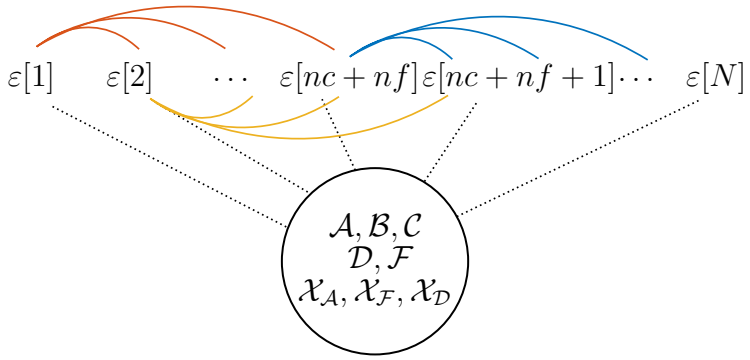


Figure 1: Chordal extension of the CSP graph of the equality constrained identification POP

Statement 6 (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 [Las06] 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 7 (Constrained sparse SDP sizes) *Depending on the model structure, the size of the SDP corresponding to the m order relaxation of the optimization problem*

$$\hat{\boldsymbol{\theta}} = \operatorname{argmin} V_a(\boldsymbol{\theta}, \boldsymbol{\varepsilon}) = \frac{1}{2N} \sum_{k=1}^{N} \varepsilon^2[k]$$
$$\mathcal{A}(z)\mathcal{F}(z)\mathcal{D}(z)y[k] = \mathcal{F}(z)\mathcal{D}(z)\mathcal{X}_{\mathcal{A}}(z)\delta[0] +$$
$$+\mathcal{D}(z)\left(\mathcal{X}_{\mathcal{F}}(z)\delta[0] + \mathcal{B}(z)u[k]\right) + \mathcal{F}(z)\left(\mathcal{X}_{\mathcal{D}}(z)\delta[0] + \mathcal{C}(z)\varepsilon[k]\right) \quad \forall k \in [N]$$
$$\|\boldsymbol{\theta}_C\|^2 + \|\boldsymbol{\varepsilon}_{I_k}\|^2 \leq R_k \quad k \in [p]$$

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 $\binom{|I|+m}{m}$, where $p = N - n_{\mathcal{C}} - n_{\mathcal{F}} + 1$ 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 4.

Theorem 3 (Convergent sparse relaxations) *The polynomial optimization problem in Statement 7 can be solved using the sparse SDP relaxation defined in [Las06]. 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 local 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 $\hat{\boldsymbol{\theta}}_N$ based on N samples of a nominal parameter vector $\boldsymbol{\theta}_0$ can be described by the asymptotic limiting distribution

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(0, \Sigma_\infty) \qquad \mathrm{Var}\left(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0\right) = \frac{1}{N}\Sigma_\infty$$

with an appropriate covariance matrix Σ_∞, 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.*

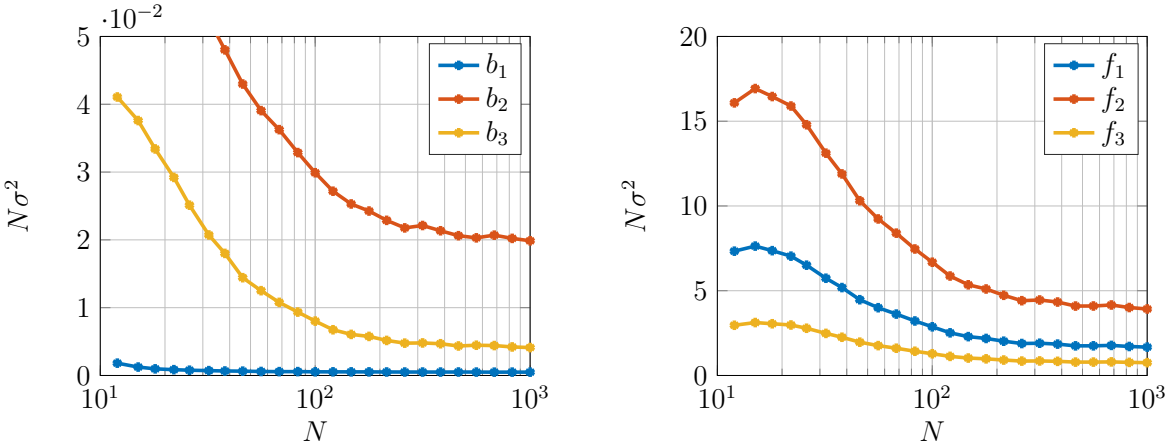
As in Thesis II, the model structure is assumed to be of the following form

$$\mathcal{A}(z)y[k] = \frac{\mathcal{B}(z)}{\mathcal{F}(z)}u[k] + \frac{\mathcal{X}_{\frac{\mathcal{B}}{\mathcal{F}}}(z)}{\mathcal{F}(z)}\delta[0] + \frac{\mathcal{C}(z)}{\mathcal{D}(z)}e[k] + \frac{\mathcal{X}_{\frac{\mathcal{C}}{\mathcal{D}}}(z)}{\mathcal{D}(z)}\delta[0] + \mathcal{X}_{\frac{1}{\mathcal{A}}}(z)\delta[0] \quad k \geq 1$$

where $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, $\mathcal{D}$ and $\mathcal{F}$ are finite order polynomials of the shift operator z^{-1} with orders $n_\mathcal{A}$, $n_\mathcal{B}$, $n_\mathcal{C}$, $n_\mathcal{D}$ and $n_\mathcal{F}$ respectively. The polynomials $\mathcal{X}_\cdot$ are polynomials representing the initial conditions of the system. The polynomials with the exception of $\mathcal{B}(z)$ and the initial conditions are monic. $u[k]$ are the input samples of the plant model $G(z) = \frac{\mathcal{B}(z)}{\mathcal{A}(z)\mathcal{F}(z)}$, $y[k]$ are the measured output values and the non-measured contaminating noise $e[k]$ is the input of the noise model $H(z) = \frac{\mathcal{C}(z)}{\mathcal{A}(z)\mathcal{D}(z)}$ 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 $\hat{\boldsymbol{\theta}}_N$ is obtained as the least squared prediction error estimate.

By considering a sample third order system with output error noise model, the parameter vector $\boldsymbol{\theta} = [b_1\ b_2\ b_3\ f_1\ f_2\ f_3]^T$, with nominal parameter vector $\boldsymbol{\theta}_0 = [0.0880\ 0.0195\ -0.0442\ -2.0536\ 1.4611\ -0.3442]^T$. The time constant of this system is around 5 samples. Figure 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.



(a) Normalized variance of $\hat{b}_1, \hat{b}_2, \hat{b}_3$

(b) Normalized variance of $\hat{f}_1, \hat{f}_2, \hat{f}_3$

Figure 2: Normalized variances of parameters against sample count.

Definition 5 (Finite sample variance loss) *For a scalar parameter $\boldsymbol{\theta}$, let $\sigma^2_{N|U}(\hat{\boldsymbol{\theta}}_N)$ denote the variance of the parameter estimate $\hat{\boldsymbol{\theta}}_N$ obtained using N samples, conditioned on the realization of the input u. The asymptotic variance is defined as $\sigma^2_{\infty|U}(\hat{\boldsymbol{\theta}}) = \lim_{N\to\infty} N\sigma^2_{N|U}(\hat{\boldsymbol{\theta}}_N)$. The* finite sample variance loss *for N samples is defined as*

$$\alpha_{N|U}(\hat{\boldsymbol{\theta}}_N) = \frac{N\sigma^2_{N|U}(\hat{\boldsymbol{\theta}}_N)}{\sigma^2_{\infty|U}(\hat{\boldsymbol{\theta}})}$$

The expected finite sample variance loss *for $\boldsymbol{\theta}$ is defined as*

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

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 $\bar{\alpha}_N$ 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.

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 $\hat{\boldsymbol{\theta}}_N$ is unbiased. This approximation is computationally more tractable than the exact problem and provides reasonable results.

Introducing the row vectors $\phi[k] \in \mathbb{R}^{1\times n_\theta} = \left.\frac{\partial \hat{y}[k]}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$, the matrix $\Phi \in \mathbb{R}^{N\times n_\theta} = (\phi[k])|_{k=1}^{N}$ and using the assumption that the estimate $\hat{\boldsymbol{\theta}}_N$ is unbiased, it is known that its covariance is bounded by the Cramér-Rao lower bound, that is

$$\mathbb{E}\left((\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0)^T \middle| U\right) \geq M_U^{-1}$$

where

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

Algorithm 2 is a moment matching based approximation method that can be used to evaluate the expected variance loss $\bar{\alpha}_N((\hat{\boldsymbol{\theta}}_N)_i)$.

Algorithm 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 Ψ.
3: Determine the degrees of freedom κN by means of moment matching between X^TX and $\Phi^T\Phi$.
4: Approximate the expected finite sample variance loss as

$$\bar{\alpha}_N((\hat{\boldsymbol{\theta}}_N)_i) \approx \frac{1}{\sigma_\infty^2((\hat{\boldsymbol{\theta}}_N)_i)}\frac{N}{\kappa N - n_\theta - 1}(\Psi^{-1})_{i,i}$$

The moment matching in the third step can be realized by finding the root of a scalar function $f(X^TX, \Phi^T\Phi)$. For one dimensional problems the trivial choice is the solution of the equation $\mathbb{E}(X^TX - \Phi^T\Phi) = 0$. For higher dimensional problems matching the expectation of the i-th diagonal term is a viable choice, or matching the determinants as $\mathbb{E}(|X^TX|) = \mathbb{E}(|\Phi^T\Phi|)$.

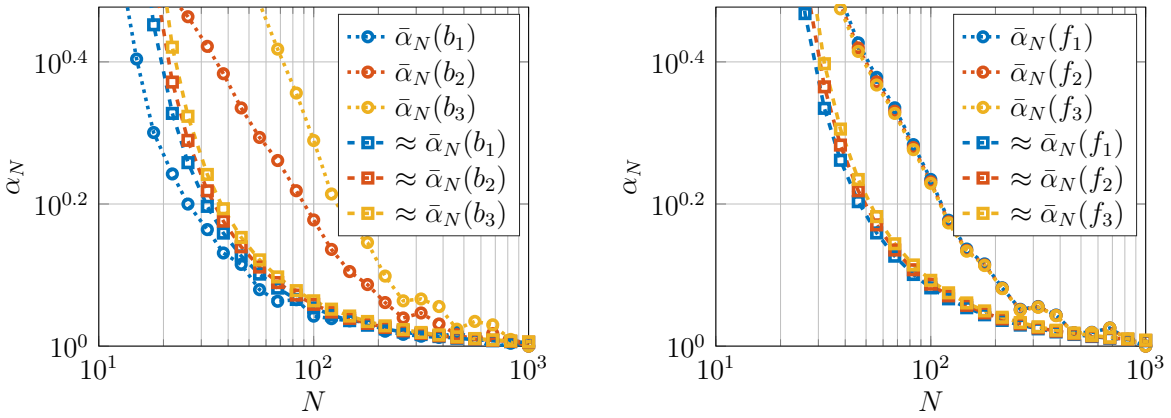
Figure 3 shows the behavior of the expected variance loss with respect to sample count for the example system.

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 bellow 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



(a) Numerator parameters b.

(b) Denominator parameters f.

Figure 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.

variance loss $\alpha_{N|U}((\hat{\boldsymbol{\theta}}_N)_i)$ of each parameter $\boldsymbol{\theta}_i$ can be approximated (based on the presented approximation algorithm, or based on multiple simulations). Assuming that the asymptotic covariance matrix of the estimate is Σ_∞, a rescaled finite sample covariance matrix can be obtained by transforming the eigenvectors of Σ_∞ as follows. Let $\alpha \in \mathbb{R}^{n_\theta}$ be the vector containing the finite sample variance losses of the individual parameters $\left(\alpha_{N|U}((\hat{\boldsymbol{\theta}}_N)_i)\right)_{i=1}^{n_\theta}$ and let the rescaling operator $r(v, \alpha)$ be defined as $r(v, \alpha) : \mathbb{R}^{n_\theta} \times \mathbb{R}^{n_\theta} \to \mathbb{R}^{n_\theta}; (r(v, \alpha))_i = v_i\sqrt{\alpha_i}$. 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 λ_i and v_i are the eigenvalues and eigenvectors of Σ_∞. 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.

3 List of publications related to the thesis

Papers published in international journals

[1] S. Kolumbán, I. Vajk, and J. Schoukens, "Perturbed datasets methods for hypothesis testing and structure of corresponding confidence sets," *Automatica*, vol. 51, no. 0, pp. 326 – 331, 2015.

[2] S. Kolumbán, I. Vajk, and J. Schoukens, "Approximation of confidence sets for output error systems using interval analysis," *Control Engineering and Applied Informatics*, vol. 14, no. 2, pp. 73–79, 2012.

Papers published in conference proceedings

[3] S. Kolumbán, "Effect of reduced sample count on estimated uncertainty," in *Proceedings of the Automation and Applied Computer Science Workshop 2014*, 2014.

[4] S. Kolumbán, "Sps toolbox for matlab," in *Proceedings of the Automation and Applied Computer Science Workshop 2013*, 2013.

[5] S. Kolumbán, "Reliable parameter estimation of pharmacokinetical double peak phenomenon models," in *Proceedings of the Automation and Applied Computer Science Workshop 2012*, 2012.

[6] S. Kolumbán, "On semidefinite representability of constraints," in *Proceeding of Automation and Applied Computer Science Workshop 2011*, 2011.

[7] S. Kolumbán, "Application of semidefinite programming in system identification," in *Proceedings of the Automation and Applied Computer Science Workshop 2010*, 2010.

[8] S. Kolumbán and I. Vajk, "Identification aspects of sdp based polynomial optimization relaxations," in *16th IFAC Symposium on System Identification SYSID 2012*, 2012.

[9] S. Kolumbán and I. Vajk, "Correlative sparsity structure of polynomial optimization problems in system identification," in *Proceedings of the TAMOP PhD Workshop*, 2012.

[10] S. Kolumbán and I. Vajk, "Rendszeridentifikációból származó polinom optimalizálási feladatok ritkasága," in *SzámOkt 2011 XXI. Nemzetközi Számítástechnikai és Oktaás Konferencia kiadványa*, 2011.

[11] S. Kolumbán and I. Vajk, "Using semidefinite programming for l1 identification of the arx model," in *Proc. 11th IEEE International Symposium on Computational Intelligence and Informatics CINTI 2010*, 2010.

[12] S. Kolumbán and I. Vajk, "Constrained identification of arx models using semidefinite programming," in *Proceedings of 1st International Scientific Workshop on DCS*, 2010.

[13] J. Schoukens and S. Kolumbán, "Study of the minimum experiment length to identify linear dynamic systems: a variance based approach," in *Proc. 2015 IEEE International Instrumentation and Measurement Technology Conference*, 2015.

4 List of publications not related to the thesis

Papers published in international journals

[14] M. Balázs, G. Horváth, S. Kolumbán, P. Kovács, and M. Telek, "Fluid level dependent markov fluid models with continuous zero transition," *Performance Evaluation*, vol. 68, no. 11, pp. 1149–1161, 2011.

[15] A. Dudás, S. Juhász, and S. Kolumbán, "Performance analysis of multi-threaded locking in bucket hash tables," *ANNALES Universitatis Scientiarum Budapestinensis de Rolando Eötvös Nominatae Sectio Computatorica*, vol. 36, pp. 63–74, 2012.

Papers published in conference proceedings

[16] I. Albert, S. Kolumbán, H. Charaf, and L. Lengyel, "The solution area-based approach of the content-driven template-based layout system," in *Proc. International Conference on System Science and Engineering*, 2013.

[17] A. Dudás, S. Juhász, and S. Kolumbán, "Recalibrating fine-grained locking in parallel bucket hash tables," in *Facing the Multicore-Challenge III 2012, LNCS 7686*, pp. 60–71, 2013.

[18] A. Dudás and S. Kolumbán, "Performance evaluation of hash tables," in *Proceedings of the Automation and Applied Computer Science Workshop 2008*, pp. 63–74, Budapest University of Technology and Economics, Department of Automation and Applied Informatics, Jun 2008.

[19] A. Dudás, S. Kolumbán, and T. Schrádi, "Effect of cache lines in array-based hashing algorithms," in *Proceedings of the 2011 International Conference on Foundations of Computer Science FCS2011*, pp. 10–15, July 2011.

[20] S. Kolumbán, S. Juhász, and A. Dudás, "Modelling performance impact of caches on hash table performance," in *Proc. of the IADIS European Conference on Informatics 2009, part of the IADIS Multiconference of Computer Science and Information systems 2009*, (Algarve, Portugal), pp. 34–42, June 2009.

5 References

[CCGW12] Marco Campi, Balázs Csanád Csáji, Simone Garatti, and Erik Weyer. Certified system identification - towards distribution-free results. In *Proc. 16th IFAC Symposium on System Identification*, pages 245–255, 2012.

[CCW12a] B.Cs. Csáji, M.C. Campi, and E. Weyer. Non-asymptotic confidence regions for the least-squares estimate. In *Proceedings of the 16th IFAC Symposium on System Identification (SYSID 2012)*, pages 227–232, 2012.

[CCW12b] B.Cs. Csáji, M.C. Campi, and E. Weyer. Sign-perturbed sums (sps): A method for constructing exact finite-sample confidence regions for general linear systems. In *Decision and Control (CDC), 2012 IEEE 51st Annual Conference on*, pages 7321–7326, 2012.

[CW05] M.C. Campi and E. Weyer. Guaranteed non-asymptotic confidence regions in system identification. *Automatica*, 41(10):1751 – 1764, 2005.

[Las01] Jean Bernard Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11(3):796–817, 2001.

[Las06] Jean Bernard Lasserre. Convergent sdp-relaxations for polynomial optimization with sparsity. In Andrés Iglesias and Nobuki Takayama, editors, *Mathematical Software - ICMS 2006*, volume 4151 of *Lecture Notes in Computer Science*, pages 263–272. Springer Berlin / Heidelberg, 2006.

[WKKM06] Hayato Waki, Sunyoung Kim, Masakazu Kojima, and Masakazu Muramatsu. Sums of squares and semidefinite program relaxations for polynomial optimization problems with structured sparsity. *SIAM J. on Optimization*, 17:218–242, January 2006.