European Journal of Control (2009)3-4:275–310
© 2009 EUCA
DOI:10.3166/EJC.15.275–310

European
Journal of
Control

System Identification of Complex and Structured Systems

Håkan Hjalmarsson*

Linnaeus Center, School of Electrical Engineering, KTH – Royal Institute of Technology, 100 44 Stockholm, Sweden

A key issue in system identification is how to cope with high system complexity. In this contribution we stress the importance of taking the application into account in order to cope with this issue. We define the concept "cost of complexity" which is a measure of the minimum required experimental effort (e.g., used input energy) as a function of the system complexity, the noise properties, and the amount, and desired quality, of the system information to be extracted from the data. This measure gives the user a handle on the trade-offs that must be considered when performing identification with a fixed experimental "budget". Our analysis is based on the observation that the identification objective is to guarantee that the estimated model ends up within a pre-specified "level set" of the application objective. This geometric notion leads to a number of useful insights: Experiments should reveal system properties important for the application but may also conceal irrelevant properties. The latter, dual, objective can be explored to simplify model structure selection and model error assessment issues. We also discuss practical issues related to computation and implementation of optimal experiment designs. Finally, we illustrate some fundamental limitations that arise in identification of structured systems. This topic has bearings on identification in networked and decentralized systems.

Keywords: system identification, model accuracy, experiment design.

1. Introduction

Identification of complex systems is a challenging problem from many perspectives. High system order, many inputs and outputs, non-linearities, practical and economical constraints in experimentation, are all issues that add to the complexity of this problem. It was noted in [1] that obtaining the process model is the single most time consuming task in the application of model-based control and [2] reports that three quarters of the total costs associated with advanced control projects can be attributed to modeling. It is therefore important to understand what makes an identification problem difficult. In this contribution we will set aside numerical issues and the problem of finding a suitable model structure, and focus on the modeling accuracy. In particular we will examine its dependence on the application and the experimental cost for obtaining the required accuracy. We will also discuss structural limitations imposed by the model structure.

There exist several results in the literature that point to that the number of estimated parameters is a very limiting factor for the modeling accuracy. Consider the discrete-time causal linear time-invariant (LTI) system[1]

$$y(t) = G_o(q)u(t) + H_o(q)e_o(t),$$

where u is the input, e_o is a white noise disturbance and where y is the output. When a parametric model

$$y(t) = G(q,\theta)u(t) + H(q,\theta)e(t), \tag{1}$$

[1] q is the shift operator.

*E-mail: hakan.hjalmarsson@ ee.kth.se

Received 1 March 2009; Accepted 24 April 2009
Recommend by J. Bokor (IPC Chair)

is identified using prediction error identification (we denote the parameter estimate by $\hat{\theta}_N$, where N is the used sample size), it is shown in [3] that for high model orders the variance of the frequency function estimate at frequency ω is given by

$$\mathrm{V}arG(\mathrm{e}^{\mathrm{j}\omega}, \hat{\theta}_N) \approx n \, \frac{\Phi_v(\mathrm{e}^{\mathrm{j}\omega})}{N\Phi_u(\mathrm{e}^{\mathrm{j}\omega})} \tag{2}$$

for a parametric model of order n. Here $\Phi_v(\mathrm{e}^{\mathrm{j}\omega})/(N\Phi_u(\mathrm{e}^{\mathrm{j}\omega}))$ is the noise power spectrum to signal energy spectrum ratio. This gives a rather pessimistic perspective on the realism of estimating systems of high order. The expression above indicates that estimating a system of order 1000 is 1000 times more expensive (measured in terms of required input energy) than a first order system, e.g., with a given limit on the input power it will take 1000 times longer to estimate the high order system than the first order system to within the same accuracy of the frequency function estimate. A recent result that points in the same direction can be found in [4, 5]. Specializing to the case where $H_o = 1$, when G and H are independently parametrized and the system is operating in open loop, it holds that

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} N\Phi_u(\mathrm{e}^{\mathrm{j}\omega})\mathrm{V}arG(\mathrm{e}^{\mathrm{j}\omega}, \hat{\theta}_N)\mathrm{d}\omega = n_G\,\lambda_e, \tag{3}$$

where n_G is the number of parameters that are used to parametrize G and where λ_e is the noise variance $\mathrm{E}[e_o^2(t)]$. Thus there is a water-bed effect for the variance of $G(\mathrm{e}^{\mathrm{j}\omega}, \hat{\theta}_N)$: if the variance is made small in some frequency regions it must be large in another region to satisfy the equality above (there exists a similar water-bed effect in spectral estimation [6]). The result above also points to that obtaining models with high accuracy over the entire frequency region becomes costly as the model order increases.

Example 1: *Suppose that it is desired that*

$$\mathrm{V}arG(\mathrm{e}^{\mathrm{j}\omega}, \hat{\theta}_N) \le \frac{1}{2\gamma}, \quad \forall\omega \tag{4}$$

(γ will be called the accuracy throughout the paper since higher γ means lower variance and hence higher accuracy). Then (3) implies that the experiment must be such that

$$N\mathrm{E}[u^2(t)] = \frac{1}{2\pi}\int_{-\pi}^{\pi} N\Phi_u(\mathrm{e}^{\mathrm{j}\omega})\mathrm{d}\omega \ge 2\,\gamma\, n_G\,\lambda_e \tag{5}$$

is satisfied. Thus n_G will have a big impact on the required external energy. ■

So how can one deal with these limitations? Well, returning to Example 1, if the input energy budget does not allow (5) to be satisfied, then the bandwidth over which the accuracy constraint (4) is required has to be relaxed sufficiently. We conclude that the demands from the application in regards to both the amount of system information (for example over which bandwidth a model is required) and the accuracy with which this information has to be extracted are very important. The importance of taking the application into account has been stressed in many places in the literature, see, e.g., [7, 8], but this issue cannot be over emphasized.

However, there is one more lesson to be learnt from Example 1. Notice that (3) implies

$$\sup_{\omega} \mathrm{V}arG(\mathrm{e}^{\mathrm{j}\omega}, \hat{\theta}_N) \ge \frac{n_G\,\lambda_e}{N\mathrm{E}[u^2(t)]}.$$

This inequality implies that with a limited input energy budget and many estimated parameters there must be frequencies where the frequency function estimate is of poor quality. If this is acceptable from the point of view of the application, this is not necessarily a bad thing as it means that the model can be lax in some frequency region. We can interpret it as that certain system properties are concealed in the experiment and thus little modeling effort has to be undertaken in regards to these properties. However, it all depends on whether this is acceptable for the application or not. Another important observation is that the factors n and n_G appearing in (2) and (3), respectively, relate to the number of identifiable parameters. Thus these factors can be controlled by performing experiments such that certain parameters are not identifiable (the reader may imagine this situation as that the excitation is such that the system behaves as a simpler system. We illustrate this by an example.

Example 2: *Consider the FIR system*

$$y(t) = \sum_{k=0}^{n-1} \theta_k^o u(t-k) + e_o(t), \tag{6}$$

where $\{e_o(t)\}$ is white noise with variance λ_e. Suppose that the objective is to estimate the static gain $\sum_{k=1}^{n} \theta_k^o$ of the system. When a white input with variance λ_u is used, the variance of the static gain estimate becomes

$$n\frac{\lambda_e}{N\lambda_u},$$

whereas if a constant input $u(t) = u$ is used (with $u = \sqrt{\lambda_u}$ so that the signal has the same power as in the white noise case), the variance is given by

$$\frac{\lambda_e}{N\lambda_u}.$$

Notice that the latter input makes the system equivalent to the static system

$$y(t) = \left(\sum_{k=0}^{n-1} \theta_k^o\right) u + e_o(t)$$

and thus the static gain can also be estimated with a simple static model. ■

The example illustrates that we facilitate the identification problem by performing an experiment that conceals system properties that are not important (in this case the individual impulse response coefficients).

Understanding the fundamental limitations in identification is closely related to optimal experiment design. The concept of least-costly identification was introduced in [9] and developed in a series of conference contributions, appearing later as [10]. The idea is to formulate the experiment design problem such that the objective is to minimize the experimental cost, typically measured in terms of required input and output energy, subject to constraints on the model's accuracy. This concept will be fundamental for our considerations. When the minimum cost of such a design problem is quantified in terms of noise, model structure, model order and required accuracy and amount of extracted system information and when this is coupled to a particular application, we will call this function, the *"cost of complexity*." The cost of complexity thus provides explicit information on how various quantities, such as the system complexity, affect the minimum experimental cost.

In system identification and related communities, optimal experiment design has witnessed a revival during the last decade, see, e.g., [11–30]. Much of the effort has been devoted to reformulate optimal experiment problems to computationally tractable convex optimization problems and to connect optimal experiment design for parametric identification methods, such as prediction error identification, to control applications [10, 31–42].

In this contributions, we use optimal experiment design as a tool for obtaining insights regarding the cost of complexity. In view of this, we have selected a particular problem formulation, to be introduced in Section 2, that facilitates the analysis. In Appendix I, some alternative approaches are discussed briefly.

A research area where complexity issues have been very much in focus is "identification for control." Here the problem of how to identify models that are suitable for control design is studied. The reader is referred to [43–46] for comprehensive treatments of the subject. One of the observations coming out from this research has been that when dealing with models of restricted complexity it is desirable to design the identification set-up such that the identification criterion is proportional to the application objective since then the model best suited for the application is obtained asymptotically (in the number of observations), see, e.g., [47–51] for examples. In this contribution, we will show that this paradigm has a deeper connotation than perhaps recognized before. We will show that optimal experiment design aims at matching the identification criterion to the application criterion. However, an important difference compared to previous work, is that there should be a scaling factor which ensures that the design is optimal for finite observations (as opposed to the bulk of the work in control relevant identification where only the model bias is considered).

For an initial discussion on the concepts discussed above, the reader is referred to Sections 4 and 5 in [45]. The cost of complexity was introduced in [52] where identification of FIR models (6) is considered. The system information to be extracted is the frequency response over a pre-specified bandwidth ξ, i.e., (4) is to be satisfied but only for $\omega \in [-\xi, \quad \xi]$. It is shown that the minimum required input energy is approximately

$$N\mathrm{E}[u^2(t)] \approx 2\gamma\, \lambda_e\, \xi\, n. \tag{7}$$

Notice that the right hand side consists of three factors: 1) The accuracy γ, 2) the noise variance λ_e, and 3) ξn which can be interpreted as the fraction of the total system complexity that has to be extracted in order to meet the quality requirements.

1.1. Outline

The paper is divided into three parts. In the first part (Sections 2–5), a framework for obtaining approximate explicit expressions for the cost of complexity is developed. More specifically, in Section 2 we discuss the problem of identification with a particular objective in mind in broad terms. In Section 3, we become more specific and translate the concepts from Section 2 to a stochastic setting, e.g., maximum likelihood (ML) and prediction error (PE) identification. We proceed by discussing issues related to models of restricted complexity in Section 5. Here we argue that good experimental conditions facilitate the use of such models. We also pursue ideas on how to cope with the limitations imposed by high system/model orders discussed above. In Section 5, we apply the concepts to prediction error identification of single input single output linear time invariant systems. We use model reference control and identification of non-minimum phase zeros as illustrations of concepts. In the second part of the paper we turn to how to numerically solve optimal experiment design problems and how to practically implement the solutions. This is covered in Section 6. Structured systems are discussed in the final part of the paper (Section 7). Here we introduce a geometric interpretation of the variance of an estimate and use this concept to illustrate some limitations that exists in regards to information provided by adding actuators and sensors to a system. Finally, some concluding remarks are given in Section 8.

2. Identification With an Objective

In this section, we will formulate the problem we will study in quite general terms. Subsequently, we will focus on more specific settings where more insight can be obtained.

2.1. The True System, the Model Set, the Data and the Estimate

We will denote the true system by $\mathcal{S}_o$. It can be seen as a map from input and disturbances/noise to the output. We denote the system input by $u(t) \in \mathbb{R}^{n_u}$ and the output by $y(t) \in \mathbb{R}^{n_y}$. We denote one input–output sample by $z(t) = \begin{bmatrix} y^T(t) & u^T(t) \end{bmatrix}^T$.

A model set $\mathcal{M}^*$ is defined as a set of input–output models M.

We assume that N input–output samples $Z^N = \{z(t)\}_{t=1}^N$ are going to be collected from the system and that they are to be used to estimate the true system. We will assume that the estimation method produces a point estimate $M(Z^N) \in \mathcal{M}^*$.

2.2. The Quality of a Model

There are many ways to measure a model's quality. However, since, presumably, the estimated model will be used in some application, involving the true system, e.g., control design, deconvolution filtering in a communication system or failure detection in a vehicle system, it is natural to measure the quality in terms of the performance of the application. We assume that if an exact mathematical model of the true system was available for the design of the application, the desired performance would be obtained. However, when the used model does not correspond to the true system, the performance of the application will degrade. We measure this in terms of a performance degradation "cost" $V_{app}(M)$ which has global minimum $V_{app}(M) = 0$ at $M = \mathcal{S}_o$, c.f. [8, 12]. The performance degradation cost can simply be the achieved performance when M is used for the design of the application compared with the achieved performance when the true system $\mathcal{S}_o$ is used, as in the following example.

Example 3: *Suppose the objective is to design a controller C such that the H_∞-norm of some transfer function matrix $F(\mathcal{S}_o, C)$ is minimized. Let the chosen (model based) control design method be represented by $C = C(M)$. Then one possibility is to take*

$$V_{app}(M) = \|F(\mathcal{S}_o, C(M))\|_\infty - \|F(\mathcal{S}_o, C(\mathcal{S}_o))\|_\infty. \tag{8}$$

■

In Example 3, $F(\mathcal{S}_o, C(M))$ represents the closed loop system property of interest when the model M is used in the design of the controller. One may also measure the performance degradation by comparing this achieved property with the desired one, i.e., $F(\mathcal{S}_o, C(\mathcal{S}_o))$, as in the next example.

Example 4: *(Example 3 continued): Instead of (8) one may choose*

$$V_{app}(M) = \|F(\mathcal{S}_o, C(M)) - F(\mathcal{S}_o, C(\mathcal{S}_o))\|_\infty.$$

■

More generally, with $\mathcal{J}(M)$ denoting the property of interest of the application when the model M is used in the design, the relative performance degradation cost[2]

$$V_{rel}(M) := \frac{1}{2}\left\|\frac{\mathcal{J}(M) - \mathcal{J}(\mathcal{S}_o)}{\mathcal{J}(\mathcal{S}_o)}\right\|^2 \tag{9}$$

where $\|\cdot\|$ is a suitable norm, may be used as V_{app}. In Section 3.5 we will see that the relative performance degradation cost has an interesting interpretation.

Example 5: *Consider the problem of estimating the first impulse response coefficient of a FIR system of order n with parameters $\theta^o = [\theta_1^o, \ \theta_2^o, \ \ldots \ , \theta_n^o]^T$. Then*

$$V_{rel}(M(\theta)) = \frac{1}{2}\left(\frac{\theta_1 - \theta_1^o}{\theta_1^o}\right)^2.$$

■

Example 6: *For a minimum phase linear time-invariant system*

$$y(t) = G_o(q)u(t) + v(t) \tag{10}$$

[2] Here we for simplicity assume that $\mathcal{J}(M)$ is a scalar.

it is desired to design a feedback controller such that the sensitivity function is S. Now given a model G, a model reference controller is given by

$$C(G) = \frac{1}{G}\frac{1-S}{S}, \tag{11}$$

This controller results in the achieved sensitivity function

$$S(G) = \frac{1}{1+G_oC(G)}. \tag{12}$$

The relative performance degradation cost can now be measured by

$$V_{rel}(G) := \frac{1}{2}\left\|\frac{S(G)-S}{S}\right\|^2 \tag{13}$$

using, for example, the $\mathcal{L}_2$-norm. ■

The objective is to design the identification set-up such that[3]

$$V_{app} \le \frac{1}{2\gamma}$$

for some pre-specified accuracy γ. Ideally we would like to determine experimental conditions such that

> *the probability of the event*
>
> $$V_{app}(M(Z^N)) \le \frac{1}{2\gamma}$$
>
> *is at least some high pre-specified value, e.g. the probability that the performance degradation cost is less than $1/(2\gamma)$ is at least 99 % when using the identified model.*

Unfortunately it is in general difficult to compute the probability of the event above. However, it is often possible (at least for large sample sizes), prior to the identification experiment, to establish that the estimate $M(Z^N)$ will belong to a (condensed) model set $U^*(\mathcal{S}_o) \subset \mathcal{M}^*$ with a certain probability:

$$M(Z^N) \in U^*(\mathcal{S}_o) \subset \mathcal{M}^*. \tag{14}$$

We could thus set up the identification such that

$$U^*(\mathcal{S}_o) \subseteq \mathcal{M}^*_{app}, \tag{15}$$

where $\mathcal{M}^*_{app}$ is the "level set"

$$\mathcal{M}^*_{app} := \left\{M \in \mathcal{M}^* : V_{app}(M) \le \frac{1}{2\gamma}\right\} \subseteq \mathcal{M}^*, \tag{16}$$

and such that the probability of the event (14) is sufficiently close to one, see Fig. 1.

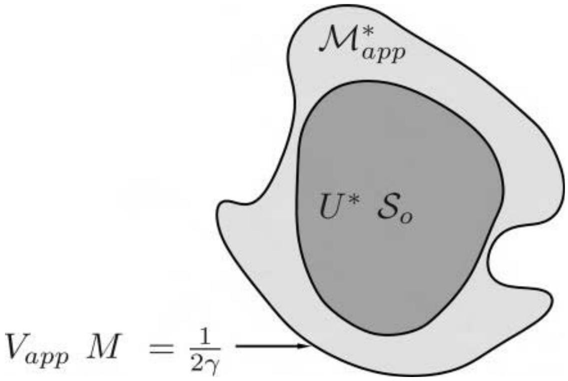

Fig. 1. The figure illustrates that $U^*(\mathcal{S}_o)$ should be contained in the set $\mathcal{M}^*_{app} := \left\{M : V_{app}(M) \le \frac{1}{2\gamma}\right\}$ in order for the performance specifications to be met.

Before we proceed, we make the reader aware that there is a certain degree of decoupling between the application and the identification in (15). The set $\mathcal{M}^*_{app}$ is a function of the application whereas the set $U^*(\mathcal{S}_o)$ depends on the identification setting. The common property is the true system $\mathcal{S}_o$. We will make extensive use of this decoupling in order to reveal the fundamental issues involved in ensuring (15).

We would also like to alert the reader that there are other ways to formulate application-oriented identification; two methods closely related to the approach outlined above are presented in Appendix I. Our path is chosen as it leads to simple expressions in certain cases, providing useful insights.

2.3. Design Variables

The user has several degrees of freedom at her disposal:

- *Experimental conditions*. This includes feedback mechanism, reference signal excitation, experiment length and sampling time.
- *Model structure and model order*. These quantities can be seen as prior information that alleviates the identification.
- *Identification method.*

There may be various constraints imposed on these choices. For example, the experiment length and the input excitation power may be limited or the system may have to be operated in closed loop with a certain controller. Due to computational and/or technological limitations, certain identification methods may have to be used. We denote all design variables by $\mathcal{D}$.

We can see the design of an identification experiment as choosing the design variables $\mathcal{D}$ such that (15) is satisfied. An important observation, which may

[3] The factor 2 is for notational convenience later on.

seem trivial, is that the verification of (15) only involves verifying that all models in $U^*(\mathcal{S}_o)$ satisfy the accuracy specified by γ. This set is specified by user and thus implies that the user does not have to be concerned with the (possibly enormous) set of models outside $\mathcal{M}^*$ that have satisfying accuracy.

2.4. Cost of Complexity

There are many ways to choose the design variables $\mathcal{D}$ such that (15) is satisfied. When the true system is in the model set $\mathcal{M}^*$, one can typically achieve the desired accuracy by just using a persistently exciting input[4] and by increasing the experiment time sufficiently. However, the identification experiment is associated with some cost, e.g., accounting for product waste and man hours. Thus it is desirable to design the experiment such that the identification cost is minimized, subject to that the desired accuracy is reached and experimental constraints are met. This way of viewing the identification problem has been coined "least costly identification" [10]. For simplicity of argument, we will measure the experimental cost for a single input single output system by the input energy $N\,\mathrm{E}[u^2(t)]$. We have already made use of this quantity in Section 1.

Let us denote the minimum cost by Q. Formally

$$\begin{aligned} Q = \ & \arg\min_{\mathcal{D}} N\,\mathrm{E}[u^2(t)] \\ & \text{subject to } U^*(\mathcal{S}_o) \subseteq \mathcal{M}^*_{app}. \end{aligned}$$

Now Q will depend on a number of parameters. First, it will typically depend on the properties of the true system and it may thus be impossible to achieve the cost Q in a practical identification experiment. However, Q provides a lower bound for the practically achievable cost and thus it is an interesting quantity to consider. Second, also the application itself will influence Q. This is related to the desired accuracy γ but also the shape of the level set $\mathcal{M}^*_{app}$ defined in (16).

To understand the role of $\mathcal{M}^*_{app}$, suppose that all models in $\mathcal{M}^*_{app}$ possess a certain property, whereas models outside this set do not possess this property. Then it will be important to choose the design variables $\mathcal{D}$ such that all models in $U^*(\mathcal{S}_o)$ possess this property. This has implications on how the excitation should be chosen. It is easy to imagine that the excitation should be such that this characteristic property is easily detected from the observed data:

> *The experimental data should reveal system properties that are important for the application.*

Now from the least costly paradigm it follows that this is the only experimental excitation that should be used (if the minimum cost Q is to be achieved). This has the implication that, in an optimal experiment with cost Q, system properties that are not particular to $\mathcal{M}^*_{app}$ will not be easily detected from the experimental data, unless this is a side effect of enhancing the "visibility" some important system property in the data. We conclude:

> *The experimental data may hide system properties that are not important for the application.*

This will have the implication that the modeling effort of such properties can be lax, and to a large extent neglected. This is an observation that can be used to the model builder's advantage. We discussed this in connection with the inequality (3) already in Section 1. We will return with more solid mathematical support for the statements above.

The set $\mathcal{M}^*_{app}$ also depends on the performance demands in the application. Imagine a user "knob" $\xi \in [0,1]$ with which the user is able to tweak the performance of the application, where a larger ξ indicates a higher performance demand in the application. Notice that this is different from the accuracy γ which is related to how close the performance of the application based on an identified model is to the performance when knowledge about the true system is used in the design of the application. We clarify this with an example.

Example 7: *(Example 6 continued): Consider again Example 6. Now let the desired sensitivity function S be given by*

$$S_\xi(q) = \frac{1-q^{-1}}{1-\frac{1-\xi}{1+\xi}q^{-1}}, \quad \xi \in [0,1], \tag{17}$$

which is parametrized such that the bandwidth over which S_ξ is small increases when ξ increases, see Fig. 2. ■

Intuitively, one may imagine that as ξ increases towards 1, it will become necessary to model more and more properties of the system accurately in order to obtain a given accuracy γ. A simple example is the following.

Example 8: *For a linear time-invariant system (10) it is desired to estimate the frequency function $G_o(e^{j\omega})$ over a certain frequency region:*

$$V_{app}(G) := \sup_{|\omega| \le \pi\xi} |G_o(e^{j\omega}) - G(e^{j\omega})|^2.$$

[4] We use the definition of persistence of excitation employed in [7]: A stationary signal is persistently exciting if its spectrum is strictly positive for all frequencies. We refer to [53–55] for details on how persistence of excitation relates to the rank of the information matrix and the identification criterion.

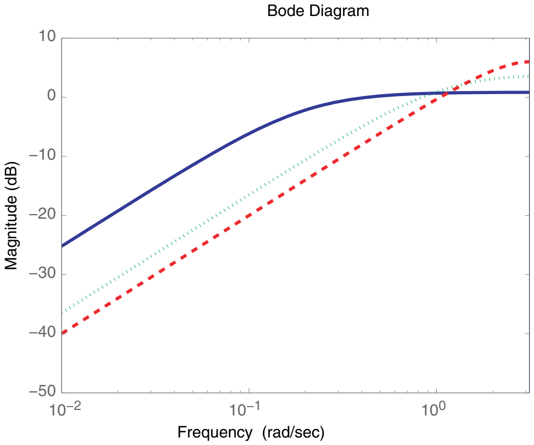

Fig. 2. Magnitude plots of S_ξ in Example 7. Solid line: $\xi = 0.1$. Dotted line: $\xi = 0.5$. Dashed line: $\xi = 1$.

■

Interpreting these increasing demands on the knowledge of the properties of the true system in terms of the set $\mathcal{M}^*_{app}$, we have that this set typically changes when ξ is increased. In particular its "shape" changes so that properties that for small ξ are not characteristic to $\mathcal{M}^*_{app}$ become characteristic as ξ increases. For example, in Example 8 the frequency response at higher frequencies become important for $\mathcal{M}^*_{app}$ when ξ increases. Thus the performance demand ξ in some sense influences the shape of $\mathcal{M}^*_{app}$ whereas the accuracy γ controls the "size". These geometric interpretations will become clearer in the next section.

On the identification side, we can see the model set as a prior. This set in turn has influence on $U^*(\mathcal{S}_o)$ and thus the "size" and "shape" of the model set should also influence the cost Q. Likewise the model order and the used identification method will influence $U^*(\mathcal{S}_o)$ and thus Q.

We will call Q the *cost of complexity* since, as we have discussed above, Q reflects how system complexity, performance demands, accuracy, disturbances etc influence the experimental cost. The main objective of the next few sections is to shed insight into how the quantities discussed above influence Q qualitatively. With such knowledge we will be able to discuss which applications are difficult or easy from an identification perspective. It will supply the user with insights into the trade-offs between application demands/system complexity/prior information/accuracy/experimental cost. We emphasize that our discussion is limited to the case where Q is the minimum required input energy.

3. Stochastic Identification

In order to proceed we will have to become more specific and we will here focus on the case where parametric models are identified under stochastic assumptions.

3.1. Introduction

Suppose that the model set is given by

$$\mathcal{M}^* = \{M(\theta), \quad \theta \in D_{\mathcal{M}} \subset \mathbb{R}^n\}, \tag{18}$$

and let us, for the moment, assume that the true system belongs to $\mathcal{M}^*$, i.e. $\exists\theta = \theta^o$ such that $M(\theta^o)$ describes the true system $\mathcal{S}_o$. We will comment on the case of low complexity modeling in Section 4.1. The parameter estimate $\hat{\theta}_N$ for a wide class of parameter estimation methods developed for such problems possesses asymptotic (in the sample size N) statistical properties of the form

$$\sqrt{N}(\hat{\theta}_N - \theta^o) \sim AsN(0, R^\dagger), \tag{19}$$

$$N(\hat{\theta}_N - \theta^o)^T R(\hat{\theta}_N - \theta^o) \sim As\chi^2(\tilde{n}), \tag{20}$$

where the "information matrix" R depends on the experimental conditions, identification method, etc. In (20), $\tilde{n}$ is the rank of R. As we already have seen in Example 2, it may be beneficial to perform experiments where the input signal is not sufficiently rich and thus results in a singular R. In order to cover such situations, we are using the Moore-Penrose pseudo-inverse $R^\dagger$ in (19) instead of the inverse R^{-1}. The notation in (19) is to be interpreted as that the estimate of any identifiable quantity $\mathcal{J}(\theta)$ has asymptotic distribution

$$\sqrt{N}(\mathcal{J}(\hat{\theta}_N) - \mathcal{J}(\theta^o)) \\ \sim AsN\Big(0, [\mathcal{J}'(\theta^o)]^T R^\dagger \mathcal{J}'(\theta^o)\Big).$$

We refer to Appendix II for details, see also [56]. For a characterization of the relationship between the information matrix and identifiability we refer to the interesting work [53, 54].

Now we will discuss how to form the set $U^*(\mathcal{S}_o)$ used in (15). Recall that $U^*(\mathcal{S}_o)$ is a set for which we a priori know that $\hat{\theta}_N$ will end up in with a certain probability. In fact, it would be desirable to take $U^*(\mathcal{S}_o) = \mathcal{M}^*_{app}$ and then design the experiment such that the probability for $\hat{\theta}_N \in \mathcal{M}^*_{app}$ is a desired value (e.g. 99 %). However, it is in general very difficult to compute this probability and we shall have to contend with other sets that we discuss next.

Let $\chi^2_\alpha(m)$ be the α-percentile of a χ^2 distribution with m degrees of freedoms. In view of (20) we have that for sufficiently large sample sizes, the ellipsoid

$$\mathcal{E}_{id} = \left\{ \theta : (\theta - \theta^o)^T R (\theta - \theta^o) \leq \frac{\chi^2_\alpha(\tilde{n})}{N} \right\} \tag{21}$$

will contain $\hat{\theta}_N$ with probability α. This means that the set of models corresponding to the ellipsoid above can be used as the set $U^*(\mathcal{S}_o)$ in (15), i.e.

$$U^*(\mathcal{S}_o) = \{ M(\theta) : \theta \in \mathcal{E}_{id} \}.$$

We alert the reader of the existence of results concerning non-ellipsoidal confidence regions, see [57] and [58, 59].

Alternatively, condition (15) can be expressed in the parameter space as

$$\mathcal{E}_{id} \subseteq \Theta_{app}, \tag{22}$$

where

$$\Theta_{app} := \left\{ \theta : V_{app}(M(\theta)) \leq \frac{1}{2\gamma}, \; \theta \in D_{\mathcal{M}} \right\}.$$

With regards to the cost of complexity, a concern with (21) is the factor $\chi^2_\alpha(\tilde{n})$ which controls the volume of the ellipsoid. This factor can be approximated by

$$\chi^2_\alpha(\tilde{n}) \approx (\beta + \sqrt{\tilde{n}})^2 = \mathcal{O}(\tilde{n}) \tag{23}$$

for some constant β [60], and thus grows linearly with $\tilde{n}$. We illustrate the effect this has with an example.

Example 9: *Consider the problem of estimating the first impulse response coefficient of a FIR system of order n. Assuming for simplicity that the desired accuracy is $\gamma = 1$, and that the true impulse response coefficient in question is 1, gives that the desired accuracy is given by*

$$(\theta_1 - \theta^o_1)^2 \leq 1 \tag{24}$$

when using the relative error as in Example 5.

Assuming the noise variance to be $\lambda_e = 1$ and a white input (it can be shown that this input is optimal) with variance λ_u so that $R = \lambda_u I$. Then

$$\mathcal{E}_{id} = \left\{ \theta : (\theta - \theta^o)^T (\theta - \theta^o) \leq \frac{\chi^2_\alpha(n)}{N \lambda_u} \right\}. \tag{25}$$

Now for all θ in $\mathcal{E}_{id}$ to satisfy (24) we must choose $N \cdot \lambda_u \geq \chi^2_\alpha(n)$. In view of (23) this means that the input energy grows linearly with the model/system order. ■

From Example 9, we see that it may be desirable to tune the ellipsoidal uncertainty set better to the level set $\mathcal{M}^*_{app}$ for the performance degradation cost. Notice that (24) imposes a constraint in only one direction in the parameter space, in fact it is a degenerate ellipsoid, whereas the ellipsoid (25) is a unit ball. For the purpose of obtaining better approximations of the level set of the performance degradation cost, which is expressed in the parameter space as Θ_{app}, let $\Gamma \in \mathbb{R}^{n \times m}$ where $m \leq n$ and suppose that γ is full rank. Now, from (19) we have that

$$\sqrt{N} \Gamma^T (\hat{\theta}_N - \theta^o) \sim AsN(0, \Gamma^T R^\dagger \Gamma),$$

and hence

$$N(\hat{\theta}_N - \theta^o)^T \Gamma \left(\Gamma^T R^\dagger \Gamma \right)^{-1} \Gamma^T (\hat{\theta}_N - \theta^o) \sim As\chi^2(m).$$

Thus

$$\mathcal{E}_{id}(\Gamma) := \left\{ \theta : (\theta - \theta^o)^T \Gamma \left(\Gamma^T R^\dagger \Gamma \right)^\dagger \Gamma^T (\theta - \theta^o) \leq \frac{\chi^2_\alpha(m)}{N} \right\} \tag{26}$$

will contain $\hat{\theta}_N$ with probability α. Notice that $\mathcal{E}_{id}$ defined in (21) can be written as $\mathcal{E}_{id}(I)$.

Example 10: *(Example 9 continued): Take $\Gamma = [1, 0, \ldots, 0]^T$.*

Then

$$\Gamma \left(\Gamma^T R^\dagger \Gamma \right)^\dagger \Gamma^T = \begin{bmatrix} 1 & 0_{1 \times (n-1)} \\ 0_{(n-1) \times 1} & 0_{(n-1) \times (n-1)}, \end{bmatrix}$$

and thus

$$\mathcal{E}_{id}(\Gamma) = \left\{ \theta : (\theta_1 - \theta^o_1)^2 \leq \frac{\chi^2_\alpha(1)}{N \lambda_u} \right\},$$

and we see that it is sufficient that $N\lambda_u \geq \chi^2_\alpha(1)$ in order for the constraint (24) to be satisfied. This constraint does not depend on the model order n. ■

We conclude that

$$\mathcal{E}_{id}(\Gamma) \subseteq \Theta_{app} \tag{27}$$

with a suitably chosen Γ rather than (22) should be considered. This is one of the key observations in the paper. We will postpone the discussion on how to choose Γ to Section 3.4.

3.2. Maximum Likelihood Estimation

In maximum likelihood estimation, R in (19) is typically given by the average Fisher information matrix

$$\begin{aligned} I_{id}(\theta^o) &:= \lim_{N \to \infty} -\frac{1}{N} \mathrm{E} \left[\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log f(\theta; Z^N) \Big|_{\theta = \theta^o} \right] \\ &= \lim_{N \to \infty} \left\langle \frac{\mathrm{d}}{\mathrm{d}\theta} \log f(\theta; Z^N) \Big|_{\theta = \theta^o}, \frac{\mathrm{d}}{\mathrm{d}\theta} \log f(\theta; Z^N) \Big|_{\theta = \theta^o} \right\rangle, \end{aligned} \tag{28}$$

where f is the probability density function for the observations Z^N, and where

$$\langle g(Z^N), h(Z^N) \rangle = \frac{1}{N} \mathrm{E}\big[g(Z^N) h^T(Z^N)\big].$$

According to the asymptotic Cramér-Rao bound, $I_{id}(\theta^o)$ majorizes all possible inverse covariance matrices R in (19), and thus corresponds to the smallest possible confidence ellipsoid $\mathcal{E}_{id}(\Gamma)$ (that was defined in (26)). When our interest is to study the cost of complexity it is thus relevant to use $R = I_{id}(\theta_o)$.

It is also possible to express the condition (22) in terms of the log-likelihood function. To see this notice that it holds [61]

$$\frac{1}{N}\mathrm{E}\big[-\log f(\theta; Z^N)\big] \approx \\ \frac{1}{N}\mathrm{E}\big[-\log f(\theta^o; Z^N)\big] + \frac{1}{2}\,(\theta - \theta^o)^T\, I_{id}(\theta^o)\,(\theta - \theta^o)$$

since $\frac{\mathrm{d}}{\mathrm{d}}\mathrm{E}[-\log\ f(\theta; Z^N)] = 0$. Thus,

$$\begin{aligned}\mathcal{E}_{id} := &\left\{\theta : (\theta - \theta^o)^T I_{id}(\theta^o)(\theta - \theta^o) \le \frac{\chi^2_\alpha(\tilde{n})}{N}\right\} \\ \approx &\left\{\theta : \frac{2}{N}\mathrm{E}\big[-\log\ f(\theta; Z^N)\big] - . \right. \\ &\left. \frac{2}{N}\mathrm{E}\big[-\log\ f(\theta^o; Z^N)\big] \le \frac{\chi^2_\alpha(\tilde{n})}{N}\right\}\end{aligned}$$

for large N. In other words, $\mathcal{E}_{id}$ corresponds to the level set

$$\left\{\theta : V_{id}(\theta) \le \frac{\chi^2_\alpha(\tilde{n})}{2N}\right\} \tag{29}$$

for the average Kullback–Leibler information distance

$$V_{id}(\theta) := \lim_{N\to\infty} \\ \frac{1}{N}\mathrm{E}\big[-\log f(\theta; Z^N)\big] - \frac{1}{N}\mathrm{E}\big[-\log f(\theta^o; Z^N)\big]. \tag{30}$$

The link between confidence sets and the cost function has been pointed out in [62]. Thus, in view of (29), we see that in order to ensure (22) the identification experiment should be designed such that the level set of the Kullback–Leibler information distance V_{id}, corresponding to level $\chi^2_\alpha(\tilde{n})/(2N)$, is contained in the level set of the performance degradation cost V_{app}, corresponding to level $1/(2\gamma)$:

$$V_{id}(\theta) \le \frac{\chi^2_\alpha(\tilde{n})}{2N} \quad \Rightarrow \quad V_{app}(M(\theta)) \le \frac{1}{2\gamma}. \tag{31}$$

This observation is remarkable in its simplicity and has some far reaching implications as we will see later.

3.3. Prediction Error Identification

We will in this section briefly derive the condition corresponding to (31) that ensures (22) when PE identification is used rather than ML identification. We consider a quadratic prediction error criterion and single-input/single-output systems. Let the one-step ahead predictor of the system output corresponding to the model $M(\theta)$ be given by the linear time-invariant predictor

$$\hat{y}(t|t-1; \theta) = W(q, \theta)\, z(t)$$

Assuming the true system to be in the model set and the predictor to be differentiable with respect to θ, then R is given by

$$R = I_{id}(\theta^o),$$

where I_{id} is the average information matrix

$$I_{id}(\theta^o) := \frac{1}{\lambda_e}\ \mathrm{E}\Big[W'(q, \theta^o)\, z(t)(W'(q, \theta^o)\, z(t))^T\Big], \tag{32}$$

where λ_e is the variance of the innovations of the true system and where $W'(q, \theta) = \partial W(q, \theta)/\partial\theta$. Now, a second order Taylor approximation gives (ε is the prediction error)

$$\begin{aligned}\mathrm{E}[\varepsilon^2(t, \theta)] &\approx \mathrm{E}[\varepsilon^2(t, \theta^o)] + (\theta - \theta^o)^T \lambda_e I_{id}(\theta^o)(\theta - \theta^o) \\ &= \lambda_e + (\theta - \theta^o)^T \lambda_e\, I_{id}(\theta^o)(\theta - \theta^o).\end{aligned}$$

Using the same arguments as in the ML-case, this expansion suggests that if we instead of the Kullback–Leibler information distance (30), that we used in the ML-case, take

$$V_{id}(\theta) := \frac{\mathrm{E}\big[\varepsilon^2(t, \theta)\big] - \lambda_e}{2\lambda_e} \tag{33}$$

then the condition (31) implies that the performance degradation constraint (22) will be met for PE identification when the demanded accuracy is high.

3.4. An Information Matrix Condition

With some abuse of notation, $V_{app}(\theta)$ will denote $V_{app}(M(\theta))$. Suppose now that

$$V_{app}(\theta) \approx \frac{1}{2}(\theta - \theta^o)^T V''_{app}(\theta^o)(\theta - \theta^o) \tag{34}$$

holds in the set of admissible model parameters Θ_{app}. This is true if V_{app} is three times continuously differentiable, the true system belongs to the model set, and γ is sufficiently large (recall that V_{app} is constructed to have global minimum 0 at $M = \mathcal{S}_o$ so its derivative at $M = \mathcal{S}_o$ will be zero). We have the following example.

Example 11: *Consider again Example 6. Let the model G be parametrized by θ. Using the L_2 norm in (13) and $S = S_\xi$ from Example 7 give that close to θ^o it holds that*

$$V_{rel}(\theta) \approx \frac{1}{2}(\theta - \theta^o)^T V''_{rel}(\theta^o)(\theta - \theta^o),$$

where

$$V''_{rel}(\theta^o) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \frac{1 - S_\xi(e^{j\omega})}{G_o(e^{j\omega})} \right|^2 G'(\theta^o, e^{j\omega})\left(G'(\theta^o, e^{j\omega})\right)^* d\omega. \tag{35}$$

■

Now, (34) implies that the level set Θ_{app} for V_{app} in the parameter space (see (22)) can be approximated by

$$\Theta_{app} \approx \mathcal{E}_{app} := \left\{ \theta : (\theta - \theta^o)^T V''_{app}(\theta^o)(\theta - \theta^o) \leq \frac{1}{\gamma} \right\}. \tag{36}$$

Similar to $\mathcal{E}_{id}(\Gamma)$, this is an ellipsoid centered at θ^o. Thus the condition (15), with the ellipsoid (26) corresponding to $U^*(\mathcal{S}_o)$, can be approximated with that the ellipsoid (26) is contained in the ellipsoid (36):

$$U^*(\mathcal{S}_o) \subseteq \mathcal{M}^*_{app} \quad \Leftrightarrow \quad \mathcal{E}_{id}(\Gamma) \subseteq \mathcal{E}_{app}.$$

Thus the performance degradation cost is acceptable if the ellipsoid $\mathcal{E}_{id}(\Gamma)$, related to the identification, is contained in the ellipsoid $\mathcal{E}_{app}$, related to the application.

We are now in position to discuss how to select Γ in (26).

Theorem 3.1: *Consider the ellipsoids $\mathcal{E}_{id}(\Gamma)$ and $\mathcal{E}_{app}$ defined in (26) and (36), respectively. Suppose that $V''_{app}(\theta^o)$ has rank m and take $\Gamma \in \mathbb{R}^{n\times m}$ such that*

$$V''_{app}(\theta^o) = \Gamma M \Gamma^T$$

for some $M > 0$. Then $\mathcal{E}_{id}(\Gamma) \subseteq \mathcal{E}_{app}$ if and only if

$$N \cdot R \geq \gamma\chi^2_\alpha(m) V''_{app}(\theta^o). \tag{37}$$

Proof: See Appendix III ■

Theorem 3.1 has several implications. Let $\lambda_i(X)$ denote the eigenvalues of the matrix X ordered in descending order. Then (37) implies

$$N \cdot \lambda_i(R) \geq \gamma\chi^2_\alpha(m)\lambda_i(V''_{app}(\theta^o)), \quad i = 1, \ldots, n,$$

see, e.g., [63], and in particular

$$\operatorname{Tr} N \cdot R \geq \gamma\chi^2_\alpha(m) \operatorname{Tr} V''_{app}(\theta^o).$$

Notice also that the degrees of freedom in the factor $\chi^2_\alpha(m)$ in (37) is given by the rank of $V''_{app}(\theta^o)$. This indicates that the shape of the performance degradation cost is very important for the cost of complexity.

3.5. The Application Demand Matrix

As in Section 2.2, take $\mathcal{J}(M)$ to denote the system property of interest that is obtained in the application when the model M is used in the design and consider the relative performance degradation cost (9). Suppose that[5] $\mathcal{J}(\theta)$ is two times differentiable. Then the Hessian of the relative performance degradation cost can be expressed as

$$V''_{rel}(\theta) = I_{rel}(\theta) := \left\langle \frac{\mathcal{J}'(\theta)}{\mathcal{J}(\theta)}, \frac{\mathcal{J}'(\theta)}{\mathcal{J}(\theta)} \right\rangle = \left\langle \frac{d}{d\theta}\log \mathcal{J}(\theta), \frac{d}{d\theta}\log \mathcal{J}(\theta) \right\rangle. \tag{38}$$

We see that (38) has the same structure as the average information matrix (28). In view of (37) this matrix enforces a demand on the average information matrix, and we will therefore call this matrix the application demand matrix.

3.6. Optimal Input Design

When we increase the experimental effort (experiment time, input power, etc.), $\mathcal{E}_{id}(\Gamma)$ will shrink. In view of the experimental cost, we would like to use precisely the amount of experimentation effort to fit $\mathcal{E}_{id}(\Gamma)$ inside Θ_{app}. One may imagine that it would be ideal to have $\mathcal{E}_{id}(\Gamma) = \Theta_{app}$, i.e., the two level sets in question of V_{id} and V_{app} should coincide. In any case, considering this situation gives an upper bound on the cost of complexity.

Notice now first that in this situation, $\mathcal{E}_{id}(\Gamma) = \mathcal{E}_{id}$ since

$$\Gamma(\Gamma^T R^\dagger \Gamma)^{-1}\Gamma^T = R$$

when equality holds in (37). Thus $\mathcal{E}_{id}(\Gamma)$ corresponds to a level set for V_{id}, according to the discussion in Sections 3.2 and 3.3. Furthermore, in the region where

[5] Here we again abuse notation; what we really mean is $\mathcal{J}(M(\theta))$.

the quadratic approximation (34) is valid, all level curves of V_{id} and V_{app} coincide (when appropriately scaled), i.e.,

$$V_{id}(\theta) = \frac{\gamma\chi_\alpha^2(n)}{N} V_{app}(\theta). \tag{39}$$

We can phrase this as:

> *the identification experiment should be set up such that the identification cost V_{id} resembles the performance degradation cost V_{app}, scaled to take into account the used number of data, the number of parameters and the desired accuracy.*

In maximum likelihood estimation and using a relative performance degradation cost, this corresponds to that

> *the experimental conditions should be set so that the average Kullback–Leibler information distance is a scaled version of relative performance degradation cost.*

In view of the above, it should come as no surprise that sometimes the optimal experimental conditions coincide with the desired operating conditions for the application. This happens for example for minimum variance control [8, 64, 12, 37]. We have further comments on this in Section 4.1 and in Section 5.3 we show that this happens in model reference control.

Before we conclude this section, we remark that it may not be possible to find experimental conditions such that (39) holds. The problem in Example 9 is one such example.

3.7. The Impact of the Model Structure

The problem of ensuring (15) is very much tied to the used model set. In (37), this manifests itself in that both R and V''_{app} depend on the model parametrization. In the interest of making the impact of the model parametrization visible, one may introduce a generic model set which includes a range of other model sets. The generic model set is in turn parametrized, with possibly infinitely many parameters. For simplicity, we will assume that the generic "parameter vector" belongs to ℓ_2. For a specific model set parametrized by $\theta \in \mathbb{R}^n$, there is a map to the generic parameters: $\tau : \mathbb{R}^n \to \ell_2$. This map represents the model structure.

Example 12: *Consider the class of ℓ_2 stable linear time invariant models. Then the impulse response coefficients can be used as generic parameters. For the model (1) the map τ is given by*

$$\tau(\theta) = \{g_0(\theta), h_0(\theta), g_1(\theta), h_1(\theta), \ldots\},$$

where

$$g_k(\theta) = \frac{1}{2\pi}\int_{-\pi}^{\pi} G(\mathrm{e}^{\mathrm{j}\omega}, \theta)\mathrm{e}^{\mathrm{j}\omega k}\mathrm{d}\omega,$$
$$h_k(\theta) = \frac{1}{2\pi}\int_{-\pi}^{\pi} H(\mathrm{e}^{\mathrm{j}\omega}, \theta)\mathrm{e}^{\mathrm{j}\omega k}\mathrm{d}\omega.$$

■

For non-linear systems the reader may think of the generic model set as a flexible basis function expansion.

We assume that for the generic model parametrization, the performance degradation cost is measured by a function V_{gen}. Now, for each parametric model set which is embedded in the generic model parametrization (as described above) the performance degradation cost can be expressed as $V_{app}(\theta) = V_{gen}(\tau(\theta))$. Assuming the involved functions to be differentiable of the required orders, this function has Hessian

$$V''_{app}(\theta^o) = [\tau'(\theta^o)]^T V''_{gen}(\tilde{\tau}^o)\,\tau'(\theta^o).$$

where $\tilde{\tau}^o = \tau(\theta^o)$. Similarly, it follows from (28) that the information matrix corresponding to a model set parametrized by θ is related to the information matrix corresponding to the generic parameterization according to (we use I_{id} to denote both information matrices, the argument reveals to which parametrization it corresponds to)

$$I_{id}(\theta^o) = [\tau'(\theta^o)]^T I_{id}(\tilde{\tau}^o)\,\tau'(\theta^o),$$

Thus, in the case of maximum likelihood estimation, we may write the condition (37) as

$$[\tau'(\theta^o)]^T \left(N \cdot I_{id}(\tilde{\tau}^o) - \gamma\chi_\alpha^2(m)\, V''_{gen}(\tilde{\tau}^o)\right) \tau'(\theta^o) \geq 0, \tag{40}$$

where $m = \operatorname{Rank} V''_{gen}(\tilde{\tau}^o)\tau'(\theta^o) = \operatorname{Rank} V''_{app}(\theta^o)$.

Thus we see that the model structure is helpful in ensuring (40) if the kernel of $[\tau'(\theta^o)]^T$ includes the eigenvectors of

$$N \cdot I_{id}(\tilde{\tau}^o) - \gamma\chi_\alpha^2(m) V''_{gen}(\tilde{\tau}^o) \tag{41}$$

which correspond to negative eigenvalues. We also notice that if (41) is positive semi-definite then the identification objective (40) is met regardless of the model set (as long as it is a subset of the generic model set and includes the true system). This in turn implies that the model structure selection problem is simplified considerably as the accuracy will not degrade with over parametrization.

4. Dealing With System and Model Complexity

Up until now we have studied how to set up an identification experiment such that a given performance objective is met. In the estimation step, we have assumed that the model order is known. To further our understanding of optimal experiments, we will now examine the case where the true system may not be in the model set or when the system order is unknown, i.e., the model order selection problem. In the last subsection we will consider how to handle high system complexity when the application information is not singular but has a large spread of the eigenvalues.

4.1. Restricted Complexity Models

Assume now that the true system $\mathcal{S}_o$ belongs to a model set (18), parametrized by $\theta \in \mathbb{R}^n$. Following the reasoning in Section 3.6, we assume that the identification experiment is set up such that the identification criterion matches the performance degradation cost so that (39) holds.

Now we are interested in what happens when we use a model set which is a subset of $\mathcal{M}^*$ but which does not contain $\mathcal{S}_o$. In the parameter space, our model set is parametrized by $\eta \in \mathbb{R}^{\tilde{n}}$ where $\tilde{n} < n$. A given model parameter η for the model set of restricted complexity corresponds to a model parameter $\theta = \theta(\eta)$ of $\mathcal{M}^*$. Thus the performance degradation cost of a model with parameter η is given by $V_{app}(\theta(\eta))$.

One very reassuring observation can be made when (39) holds, or even when the identification set-up is such that it only holds that

$$V_{id}(\theta) = \alpha\, V_{app}(\theta) \tag{42}$$

for some fixed constant $\alpha > 0$. The implication is that

> *the identified model, within the model set of restricted complexity, will approach the best possible model as the sample size grows, in the sense that of all models in the model set (of restricted complexity), this model minimizes the performance degradation cost V_{app}.*

This is simply due to that $V_{id}(\theta(\eta))$ becomes the identification criterion for the model set of restricted complexity as the sample size grows, and thus with V_{app} proportional to V_{id}, $V_{app}(\eta(\theta))$ will also be minimized[6]

One direct implication of the above is that when (39) holds, then as long as the model has sufficient degrees of freedoms to make V_{id} zero, the property of interest will be consistently estimated. The problem in Example 2 of estimating the static gain is one such example. In fact this holds sometimes even if it is not possible to match the identification and the performance degradation costs exactly. Exact conditions are established in [20, 65]. We will see an example of this in Section 5.4.

As pointed out in Section 1, the benefits of matching the identification criterion with the application objective have been noted in the context of identification for control, see, e.g., [47–51]. Notice however that we here make an important additional observation. The scaling in (39) is very important in order to ensure that the performance degradation cost constraint is not violated for finite sample size N. The correct scaling can in general only be obtained with proper design of the excitation, the feedback and the experiment length and not with other methods such as prefiltering. We will illustrate this in Section 5.3.

4.2. Model Order Selection

In prediction error identification V_{id} is itself the performance degradation cost when the application is one-step ahead prediction. Now, there is a significant body of literature on how to determine a model that is good for prediction. For example AIC [7] provides an estimate of the prediction error variance when the estimated model is applied to one-step ahead prediction. The prediction error variance is indeed a measure of the performance degradation cost when a model is used for prediction.

Now we make the simple observation that $V_{app} = V_{id}$ (this is what we have assumed at the outset of this section). But this implies that

> *all available theory on identification when the intended model use is prediction is applicable for the application in question.*

For example, suppose that we are considering a sequence of increasing model sets with increasing number of parameters, and that one of the model sets is known to contain the true system. Denote the least-squares estimate of the model structure with $\tilde{n}$ parameters by $\hat{\eta}$ and a corresponding parameter vector in the model structure of the full order model by $\theta(\hat{\eta})$.

[6] The attentive reader may be concerned with our argument above since there may be several values of θ that correspond to a given η due to non-identifiability, either due to the model parameterization (e.g. pole-zero cancellations), or due to lack of persistence of excitation, so that the map $\theta(\eta)$ is not well defined. Notice, however, that models corresponding to such θ's are indistinguishable from the data (see the discussion in Appendix II) and hence V_{id} is exactly the same for these different models and thus, in view of (42), also V_{app} is identical for these models. Thus it suffices for our arguments above to hold that we define a unique map $\theta(\eta)$.

Then AIC can be used to provide an estimate of $V_{app}(\theta(\hat{\eta}))$. In our context this estimate is given by

$$\hat{V}_{app}(\theta(\hat{\eta})) = \frac{\frac{1}{N}\sum_{t=1}^{N} \varepsilon(t,\theta(\hat{\eta}))^2 + 2\frac{\lambda_e \tilde{n}}{N} - \lambda_e}{\gamma \chi^2_\alpha(n)\lambda_e/N}$$

when the noise variance λ_e is known. In Appendix IV it is shown that this is an unbiased estimate for deterministic linear regression problems subject to white Gaussian noise. In this appendix the statistical properties of the estimate are also derived.

4.3. Further Measures to Manage System Complexity

One of the more important observations in Section 3 was that the experimental effort can be reduced when the Hessian of the performance degradation cost is singular. This manifests itself in (37) by the factor $\chi^2_\alpha(m) = \mathcal{O}(m)$ where m is the rank of $V''_{app}(\theta^o)$. Now it may be that $V''_{app}(\theta^o)$ is full rank but that there is a large spread of the eigenvalues. In this situation it is tempting to replace $V''_{app}(\theta^o)$ by a low rank approximation since this will reduce the scaling factor $\chi^2_\alpha(m) = o(m)$ which in turn will reduce the experimental cost since (37) becomes a milder constraint. Now there are several issues to consider:

(1) How much should the rank be reduced with, and which directions should be removed?
(2) How should the estimation be performed?

Suppose that the accuracy γ is given and that it is desired to minimize some measure of the experimental cost. Then it turns out that, ideally, the following procedure should be followed:

First make an eigenvalue decomposition

$$V''_{app}(\theta^o) = EDE^T,$$

where D is a diagonal matrix consisting of the eigenvalues $\{\lambda_k\}$ of V''_{app}, ordered in descending order of $\alpha_k^2\lambda_k$ where $\alpha = [\alpha_1 \ \ \ldots \ \ \alpha_n]^T = E^T\theta$. Now for each m_a such that $\sum_{k=m_a+1}^{n} \alpha_k^2\lambda_k \leq 1/\gamma$ holds, perform the following steps:

(i) Partition $E = [E_a \quad E_\Delta]$ with $E_a \in \mathbb{R}^{n\times m_a}$ and take D_a to be the upper $m_a \times m_a$ block of D.
(ii) Take

$$\gamma_a = \frac{1}{\frac{1}{\gamma} - \sum_{k=m_a+1}^{n}\alpha_k^2\lambda_k}.$$

(iii) Take

$$\tilde{V}''_{app}(\theta^o) := E_a D_a E_a^T$$

as approximation of $V''_{app}(\theta^o)$ and design an experiment such that (37) holds with $V''_{app}(\theta^o)$ replaced by $\tilde{V}''_{app}(\theta^o)$, γ replaced by γ_a and m replaced by m_a.

Pick the rank m_a for which the experimental cost computed above is minimized. Use the experimental design which corresponds to this rank of $\tilde{V}''_{app}(\theta^o)$. Estimate $\hat{\theta}_N$ and form the projection

$$\hat{\theta}_N^{proj} = E_a E_a^T \hat{\theta}_N.$$

It then holds that

$$(\hat{\theta}_N^{proj} - \theta^o)^T V''_{app}(\hat{\theta}_N^{proj} - \theta^o) \leq \frac{1}{\gamma} \tag{43}$$

with probability α, and this is achieved at the lowest possible cost. The above procedure gives the following observation:

The number of eigenvalues that dominate the eigenvalue distribution of $V''_{app}(\theta^o)$ represents the amount of system information that has to be extracted from the system.

We will denote the number of dominating eigenvalues by ξn, where n is the dimension of θ. It may seem tempting to also reduce the model complexity, i.e. the number of parameters, so that it corresponds to the reduction of rank in $V''_{app}(\theta^o)$. However, this could cause an increase in the variability of the estimate so that (43) does not hold. This is due to the invariance principle [66] which is a separation principle that says that one should first model as well as possible and then use the resulting model for all further computations. This is extensively discussed in Section 4 in [45]. See also Appendix VI for further ramifications of this principle. An exception is when the experiment design is such that the inverse covariance matrix R is singular. Then it is sufficient to use a model with the number of parameters equal to the rank of R, see Appendix II, and Example 2 for an example.

5. Application to Prediction Error Identification

In this section we will illustrate how our insights from the previous sections can be used. In the interest of simplicity, we will consider prediction error identification of causal linear time invariant systems.

5.1. Preliminaries

Consider the output-error model

$$y(t) = G(q)u(t) + e(t),$$

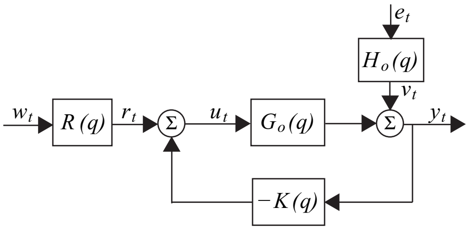

Fig. 3. Block diagram of a single input single output LTI system with output feedback.

where G is parametrized such that the true system is in the model set. Let n denote the number of parameters that parametrize G and let λ_e denote the true noise variance. The experimental set-up is shown in Fig. 3 which we assume to be stable. In the figure w is white noise with unit variance and thus R can be seen as the minimum phase stable spectral factor of the reference signal r.

Consider now that it is of interest to achieve a certain transfer function L_o, that involves the true system G_o. Based on the model, $L = L(G)$ is designed, where L is such that $L(G_o) = L_o$. Model reference control, discussed in Examples 6, 7 and 11, is one application which fits into this framework. Other such applications include filter design and simulation. We measure the performance degradation cost by the relative cost (9) using the $\mathcal{L}_2$ norm with the quantity of interest taken as $\mathcal{J}(\theta) = L(G(\theta))$.

We will use (39) as a way to assess the cost of complexity. Thus we need to establish expressions for V_{id} and V_{app}. Starting with V_{id}, the prediction error is given by

$$\varepsilon(t, G) = (y - Gu(t)) = (G_o - G)u(t) + e(t).$$

Using this, straightforward calculations give that V_{id} defined in (33) can be expressed as

$$V_{id}(\theta) = \frac{1}{2\pi}\int_{-\pi}^{\pi} |G(\mathrm{e}^{\mathrm{j}\omega}) - G_o(\mathrm{e}^{\mathrm{j}\omega})|^2 \frac{\Phi_u(\mathrm{e}^{\mathrm{j}\omega})}{\lambda_e}\,\mathrm{d}\omega. \tag{44}$$

By a first order Taylor approximation, we have $(L(G) - L(G_o))/L(G_o) \approx \frac{L'(G_o)}{L(G_o)}(G - G_o) = (G - G_o)$ $\frac{\mathrm{d}}{\mathrm{d}G}\log L(G)|_{G=G_o}$ and thus the relative performance degradation cost V_{rel} can be expressed approximately as

$$V_{rel}(G) \approx \frac{1}{2\pi}\int_{-\pi}^{\pi} |G(\mathrm{e}^{\mathrm{j}\omega}) - G_o(\mathrm{e}^{\mathrm{j}\omega})|^2 \left|\frac{\mathrm{d}}{\mathrm{d}G}\log L(G)|_{G=G_o(\mathrm{e}^{\mathrm{j}\omega})}\right|^2 \mathrm{d}\omega. \tag{45}$$

5.2. Cost of Complexity

Considering experimental conditions such that (39) holds will provide us with an upper bound on the cost of complexity. Clearly, (39) holds for (44) and (45) if

$$N\Phi_u(\mathrm{e}^{\mathrm{j}\omega}) = \gamma\chi^2_\alpha(n)\lambda_e \left|\frac{\mathrm{d}}{\mathrm{d}G}\log L(G)\Big|_{G=G_o(\mathrm{e}^{\mathrm{j}\omega})}\right|^2 \tag{46}$$

holds for all frequencies. The reader should observe that (46) is an overbound of the cost of complexity for two reasons: (1) The equality (39) may not be the optimal choice of experimental conditions. (2) There may exist experimentally cheaper ways to match (44) and (45) than matching the integrands frequency by frequency. In regards to (2), it can be shown, using the asymptotic theory in [67], that matching the integrands implies no loss in the cost when the model complexity increases. Taking into account the ease with which the result (46) was obtained, we will stick to this condition. Thus we have the upper bound

$$\begin{aligned} Q &:= \min N \cdot \mathrm{E}[u^2(t)] \\ &\leq \gamma\chi^2_\alpha(n)\lambda_e \frac{1}{2\pi}\int_{-\pi}^{\pi} \left|\frac{\mathrm{d}}{\mathrm{d}G}\log L(G)\Big|_{G=G_o(\mathrm{e}^{\mathrm{j}\omega})}\right|^2 \mathrm{d}\omega \\ &=: \gamma\chi^2_\alpha(n)\lambda_e \left\|\frac{\mathrm{d}}{\mathrm{d}G}\log L(G)\Big|_{G=G_o}\right\|_2^2. \end{aligned} \tag{47}$$

It is instructive to compare this bound with the minimum energy required when the input is white noise, which is a commonly used type of input. In order to ensure

$$N \cdot \Phi_u(\mathrm{e}^{\mathrm{j}\omega}) \geq \gamma\chi^2_\alpha(n)\lambda_e \left|\frac{\mathrm{d}}{\mathrm{d}G}\log L(G)\Big|_{G=G_o(\mathrm{e}^{\mathrm{j}\omega})}\right|^2$$

over all frequencies with a white input (which has spectrum $\Phi_u(\mathrm{e}^{\mathrm{j}\omega}) = \lambda_u$, where λ_u denotes the input power), the minimum input energy is given by

$$\begin{aligned} N\lambda_u^{\min,white} &:= \gamma\chi^2_\alpha(n)\lambda_e \max_\omega \left|\frac{\mathrm{d}}{\mathrm{d}G}\log L(G)\Big|_{G=G_o}\right|^2 \\ &=: \gamma\chi^2_\alpha(n)\lambda_e \left\|\frac{\mathrm{d}}{\mathrm{d}G}\log L(G)\Big|_{G=G_o}\right\|_\infty^2. \end{aligned} \tag{48}$$

Comparing (48) with (47) , we see that a white input requires significantly more input energy than necessary when $\frac{\mathrm{d}}{\mathrm{d}G}\log L(G)|_{G=G_o}$ has narrow spikes in the frequency domain.

5.3. Application 1: Model Reference Control

We will now return to the model reference control problem discussed in Examples 6, 7 and 11. The objective is thus to obtain the closed loop sensitivity $S = S_\xi$ (defined in (17)) and thus $L(G) = S(G)$ (see (12)). Using (11)–(12),

$$\frac{\mathrm{d}}{\mathrm{d}G}\log L(G)\Big|_{G=G_o} = L'(G_o)/L_o = (1-S_\xi)/G_o. \tag{49}$$

We will begin with a specific example.

Example 13: *Consider the FIR system*

$$y(t) = \theta_1^o u(t) + \theta_2^o u(t-1) + e_o(t),$$

where $\theta_1^o = 2, \theta_2^o = 1$ *and where* e_o *is zero mean white noise with variance 1.*

Starting with a low required bandwidth by choosing $\xi = 0.025$ *gives a cost function* $V_{rel}(\theta)$ *with contour lines as in* Fig. 4. *From the figure we see that the performance degradation is not very sensitive in one direction in the parameter space. Optimal input design (see Section 6) gives contour lines of the identification criterion according to* Fig. 5. *We see that the identification cost is oriented in the same direction as the relative performance degradation cost.*

Now we increase the bandwidth significantly by taking $\xi = 1$, *see* Fig. 2. *Then the performance degradation cost changes to the one in* Fig. 6. *We see that the contour lines around the optimum are now much more concentric. Thus both system parameters are important for the performance, meaning that more system information has to be extracted in order for the performance specifications to be met. Optimal input design gives an identification cost with contour lines according to* Fig. 7. *We see that again the optimal input ensures that the cost functions are matched to each other.* ■

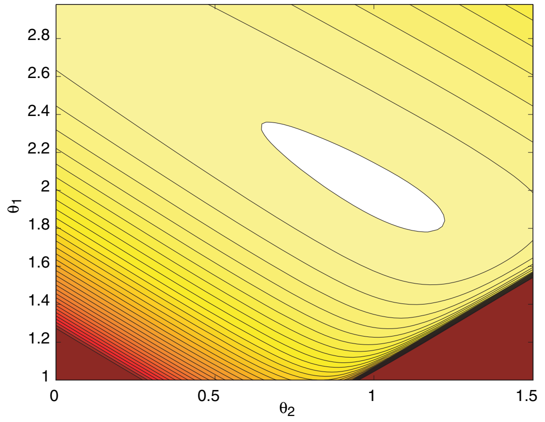

Fig. 4. Contour plot of V_{rel} when $\xi = 0.025$.

Let $\Phi_u^{\text{desired}} = \lambda_e \left|\frac{1-S_\xi(\mathrm{e}^{\mathrm{j}\omega})}{G_o(\mathrm{e}^{\mathrm{j}\omega})}\right|^2$. This is the input spectrum when the closed loop operates according to

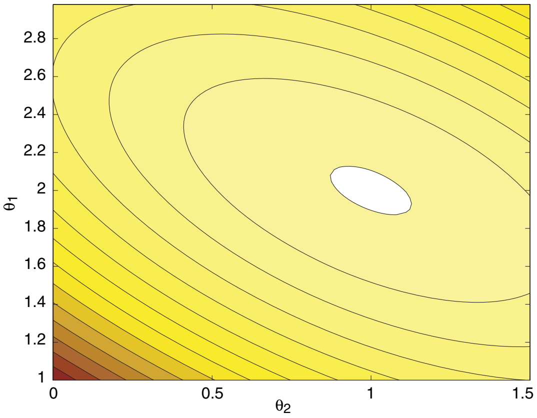

Fig. 5. Contour plots of V_{id} when $\xi = 0.025$.

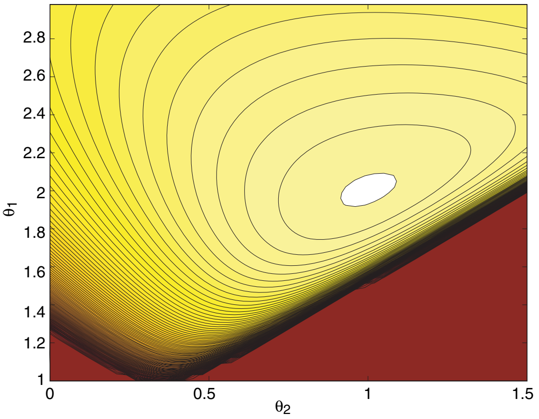

Fig. 6. Contour plot of V_{rel} when $\xi = 1$.

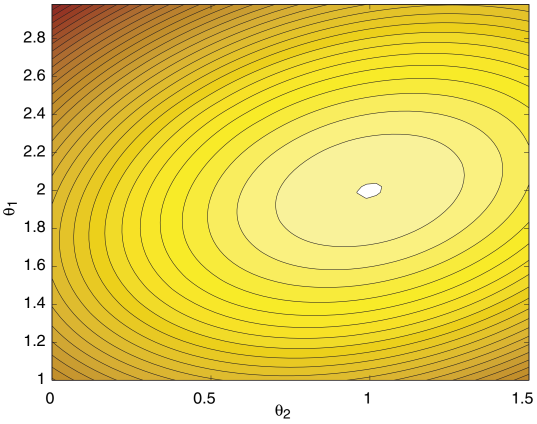

Fig. 7. Contour plots of V_{id} when $\xi = 1$.

the desired specifications, i.e. when the sensitivity function is given by S_ξ and only noise excites the system. From Section 5.2 we have that the desired performance degradation objective will be met in an experiment of length N if the input energy spectrum during the identification experiment is

$$N\widetilde{\Phi}_u(e^{j\omega}) := \gamma\chi^2_\alpha(n)\,\Phi_u^{\text{desired}}(e^{j\omega}), \qquad (50)$$

Thus one way of achieving the identification objective is to perform an experiment under the desired closed operating conditions with an experiment length $N = \gamma\chi^2_\alpha(n)$. This is an illustration of when the experimental conditions coincide with the desired operating conditions, cf. the discussion at the end of Section 3.6.

The reader should also observe that the same cost and result can be obtained in open loop, e.g. simply choose the input spectrum as Φ_u^{desired} and $N = \gamma\chi^2_\alpha(n)$. This results in a different output spectrum compared to the closed loop solution.

When the data has been collected using a different input spectrum, let us call it Φ_u^{id}, it may be tempting to use prefiltering to shape the identification objective into the desired V_{rel}. Replacing the prediction error $\varepsilon(t, G)$ by the filtered prediction error $\varepsilon_F(t, G) = F(q)\varepsilon(t, G)$ gives

$$V_{id}(\theta) = \frac{1}{2\pi}\int_{-\pi}^{\pi} |G(e^{j\omega}) - G_o(e^{j\omega})|^2 \frac{|F(e^{j\omega})|^2\,\Phi_u^{\text{id}}(e^{j\omega})}{\lambda_e}\,d\omega,$$

which shows that by taking F such that $|F|^2 = \Phi_u^{\text{desired}}/\Phi_u^{\text{id}}$ will give the desired identification objective. However, this will not work since our derivations in Section 3 were based on that the true system is in the model set and prefiltering can be seen as changing the noise model from 1 (our output-error model) to F^{-1} which is not consistent with our assumption that the true system is of output-error type. This illustrates the comment made at the end of Section 4.1 that matching identification and application criteria can in general only be achieved through the experiment design.

Now we turn our attention to how the performance specifications of the application influence the cost of complexity. For our model reference control problem we see from (50) that the required energy increases when the bandwidth of $1 - S_\xi$ is increased (this happens when ξ increases, see Fig. 2). We illustrate this with an example.

Example 14: *Suppose that the noise is white with variance λ_e and that G_o has constant magnitude. Suppose also that $S = S_\xi$, with S_ξ defined in (17). In this scenario, (49) is given by*

$$L'(G_o)/L_o = (1 - S_\xi)/G_o = \frac{2\xi}{(1+\xi)G_o}\frac{z^{-1}}{1 - az^{-1}},$$

where $a = (1-\xi)/(1+\xi)$. The input spectrum in (50) can be written as

$$N\Phi_u(z) = \frac{\lambda_e\gamma\chi^2_\alpha(n)\xi}{|G_o|^2}\sum_{i=-\infty}^{\infty} a^{-|i|}z^{-i}.$$

It can be shown that this is the input with minimum energy that achieves the identification objective, also for finite model order. We see that the required input energy is given by

$$\begin{aligned} N\lambda_u &= N\frac{1}{2\pi}\int_{-\pi}^{\pi}\Phi_u(e^{j\omega})\,d\omega \\ &= \frac{\lambda_e\gamma\chi^2_\alpha(n)\,\xi}{|G_o|^2} \approx \frac{\lambda_e\gamma\,\xi n}{|G_o|^2}. \end{aligned} \qquad (51)$$

This expression clearly indicates the trade-offs the user has to make if the energy budget is limited. For given system complexity n and noise variance λ_e, either the accuracy γ or the performance specification ξ has to be sacrificed. The factor ξ can be seen as a measure of the fraction of the total system complexity (measured as the total number of parameters) that has to be extracted from the system. In Fig. 8 *the distribution of the eigenvalues of the application information matrix is shown when $\xi = 0.2$ and the system order is $n = 20$. Notice that the fraction of dominating eigenvalues is around $4/20 = 0.2 = \xi$. This is consistent with the discussion below (43) in Section 4.3. The attentive reader will notice that in this example we have not used the procedure suggested in Section 4.3. The reason is that for the considered problem it is possible to match the*

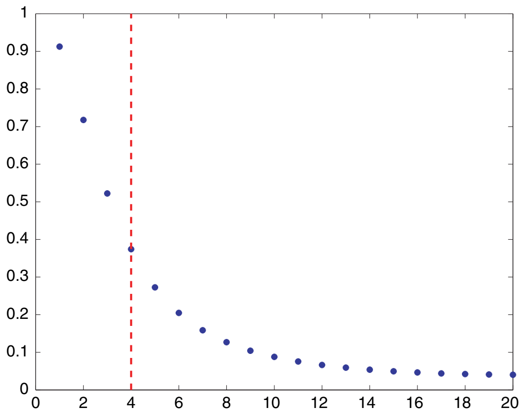

Fig. 8. Example 14: Solid line: The eigenvalues for the application demand matrix plotted in descending order when $\xi = 0.2$ and the system order $n = 20$. Dashed line: Marks the fraction ξn of dominating eigenvalues.

information matrix in the identification to the application demand matrix. When this is not possible it is beneficial to follow the procedure in Section 4.3.

In Fig. 9.a *the obtained relative performance degradation cost (13) is plotted for a sample size increasing from* $N = 100$ *to* $N = 200$*, when the optimal input design is used. The figure shows 100 Monte Carlo runs when the input design is for* $\lambda_e = 0.1, N = 200, \xi = 0.1, \gamma = 1000,$ *and* $\alpha = 95\%$. *The used model order is 2 and the true system is* $G_o(q) = q^{-1}$. *For comparison the cost when white noise, with the same variance as the optimal input, is shown in* Fig. 9b. *From* Fig. 9a *we see that about 3% of the realizations do not satisfy the specifications when* $N = 200$. *This is consistent with the choice* $\alpha = 95\%$. *For the white noise input 7% of the realizations do not satisfy the specifications.*

In Section 5.2 it was noted that (47) and (48) implied that if L'/L_o *is spiky, then the required input energy can be significantly lower for the optimal input than for white noise, when the model order is high. In this example, the minimum required energy for white noise is*

$$N\lambda_u^{min} \approx \frac{\lambda_e \gamma n}{|G_o|^2}.$$

Comparing with (51), we see that when ξ *is small (low bandwidth) and the model order is high (which is the situation where our expressions are accurate), a white noise input will require many times the minimum necessary energy. Exact calculations show that with a model order of 10 and* $\xi = 0.1$*, 5 times as much energy as necessary is required for a white noise input.* ■

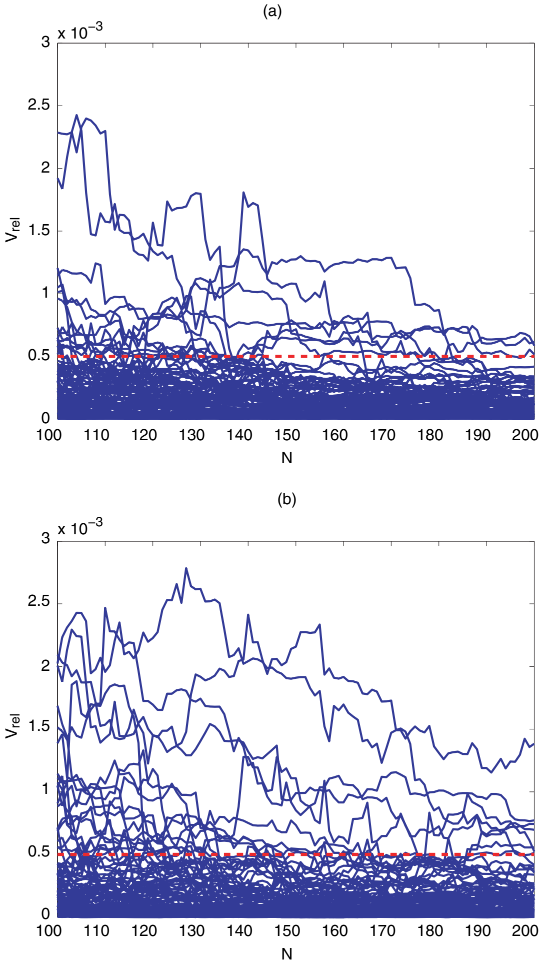

Fig. 9. Example 14: Relative performance degradation cost over 100 Monte Carlo runs. (a) Optimal input. (b) White input. Dashed line: Desired accuracy after $N = 200$ observations.

5.4. Application 2: Identification of Non-minimum Phase zeros

As another application, we will now turn to estimation of non-minimum phase (NMP) zeros. Assume that the stable system $G_o = \sum_{k=1}^{\infty} g_k q^{-k}$ has a real-valued NMP zero at z_o of multiplicity 1 and that this zero is our quantity of interest. We thus take $\mathcal{J}(\theta^o) = z_o$. Now we will not consider a specific model parametrization but instead we will follow Example 12 and use the impulse response coefficients of the system model and noise model dynamics as generic parameters. It can be shown [68] that

$$\frac{\mathrm{d}z_o}{\mathrm{d}g_k} = -\frac{z_o}{\widetilde{G}_o(z_o)} z_o^{-k}, \tag{52}$$

where

$$\widetilde{G}_o(z) = G_o(z)/(1 - z^o z^{-1}). \tag{53}$$

Now as the sensitivity of $\mathcal{J}$ with respect to the impulse response of the noise model is zero, we can ignore these coefficients in the following. Using (52), the part of the application demand matrix that depends on the system impulse response coefficients is thus given by the rank-1 Hankel matrix

$$I_{rel}(\tilde{\tau}^o) = \frac{1}{|z_o\,\widetilde{G}_o(z_o)|^2}\begin{bmatrix} 1 & z_o^{-1} & z_o^{-2} & \cdots \\ z_o^{-1} & z_o^{-2} & z_o^{-3} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \tag{54}$$

where $\tilde{\tau}^o = \tau(\theta^o)$. The corresponding part of the information matrix is given by the Toeplitz matrix

$$I_{id}(\tilde{\tau}^o) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\Phi_u(\mathrm{e}^{\mathrm{j}\omega})}{\Phi_v(\mathrm{e}^{\mathrm{j}\omega})} \begin{bmatrix} 1 & \mathrm{e}^{\mathrm{j}\omega} & \mathrm{e}^{\mathrm{j}2\omega} & \cdots \\ \mathrm{e}^{-\mathrm{j}\omega} & 1 & \mathrm{e}^{\mathrm{j}\omega} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \mathrm{d}\omega.$$

Now if we take

$$N\Phi_u(z) = \Phi_v(z)\frac{\gamma\chi^2_\alpha(1)}{|z_o\,\widetilde{G}_o(z^o)|^2}\sum_{i=-\infty}^{\infty}(z^o)^{-|i|}z^{-i}$$
$$= \Phi_v(z)\frac{\beta}{|\widetilde{G}_o(z^o)|^2}\frac{1-(z^o)^{-2}}{(z^o-z^{-1})(z^o-z)}, \tag{55}$$

we obtain

$$N I_{id}(\tilde{\tau}^o) = \frac{\gamma\chi^2_\alpha(1)}{|z_o\,\widetilde{G}_o(z_o)|^2}\begin{bmatrix} 1 & z_o^{-1} & z_o^{-2} & \cdots & \\ z_o^{-1} & 1 & z_o^{-1} & & \cdots \\ \vdots & \vdots & \vdots & \ddots & \end{bmatrix}. \tag{56}$$

For these experimental conditions, it is now easy to show [33] that

$$N I_{id}(\tilde{\tau}^o) \geq \gamma\chi^2_\alpha(1) I_{rel}(\tilde{\tau}^o),$$

which means that the condition (37) (with $R = I_{id}(\theta^o)$) holds for the generic parametrization. But then according to the discussion in Section 3.7, (37) is satisfied regardless of the model structure. We have thus shown that choosing the input spectrum according to (55) will ensure that the accuracy of the zero estimate is the desired value γ regardless of the model structure.

The input spectrum (55) (or any scaled version of it) has another interesting property. It allows the zero at z_o to be consistently estimated for restricted complexity models. Recall from Section 4.1 that if the identification criterion is matched to the performance degradation cost and the model structure is flexible enough to make this criterion zero (which effectively means that it is flexible enough to model the system property of interest), then the property of interest will be estimated consistently as the sample size grows. In the zero estimation problem above, the input spectrum (55) gives the average information matrix (56) which in turn does not match the application demand matrix (54) perfectly. Thus in this case the two cost functions in question do not match perfectly. However, the fit is sufficiently good and it can be shown [20, 65] that when (55) is used as input spectrum and the system is operated in open loop, the zero z_o will be consistently estimated regardless of the system order and noise dynamics for ARX, ARMAX, Box-Jenkins, FIR and output-error systems when $G(q,\theta)$ has more zeros than poles. For example an FIR model with two parameters can be used.

6. Computation and Implementation of Optimal Input Designs

So far we have mainly been concerned with obtaining explicit expressions for approximations of the minimum experimental cost (measured in terms of the required input energy), i.e., the cost of complexity. Now, we will turn to the practical side of designing optimal experiments.

In order to compute the optimal experimental conditions for a given V_{app} there are two crucial issues that must be addressed:

(1) The level set $\mathcal{M}^*_{app}$ of the performance degradation cost, defined in (15), must be characterized.
(2) Experimental conditions must be determined such that $U^*(\mathcal{S}_o)$ satisfies (15).

In order to obtain a computationally tractable problem, these two issues must be closely linked. In the literature most attention has been given to the case when $U^*(\mathcal{S}_o)$ is an ellipsoid. This is the case, e.g., when the asymptotic theory for maximum likelihood and prediction error estimation is used, cf. Section 3. In this case (14) corresponds to (27).

6.1. Characterization of the Level Set for $\mathcal{M}^*_{app}$

By far the simplest way to define the level set (16) is to use the quadratic approximation (34). As we have seen in Section 3.4 this leads to the simple matrix inequality (37) when $U^*(\mathcal{S}_o)$ is an ellipsoid. When the design can be explicitly expressed as a function of the model (or its parameters) it is often easy to compute the Hessian of V_{app}. However, there are applications where it is non-trivial to compute the required sensitivities, e.g., model predictive control (MPC).

To verify that a given ellipsoid belongs to a convex set is a convex problem. However, despite this, depending on the function in question this may not be a computationally tractable problem. There is a growing body of results in the literature where, for various classes of functions, computationally tractable conditions are formulated that guarantee that a given ellipsoid belongs to the level set of a function in the considered function class. The derived conditions are typically in the form of a feasibility test of a semidefinite program. In [32] this is done for the squared chordal distance [70]

$$\frac{|G(e^{j\omega},\theta) - G_o(e^{j\omega})|^2}{(1+|G(e^{j\omega},\theta)|^2)(1+|G_o(e^{j\omega})|^2)}. \tag{57}$$

This is extended in [34] to a family of functions in the frequency domain which includes, e.g., (57) as well as the weighted relative model error

$$T\frac{G_o - G(\theta)}{G(\theta)}.$$

In [40], the feasibility of a linear matrix inequality (LMI) is shown to guarantee the existence of a state-feedback controller such that a certain H_∞ bound is satisfied for all possible models in an ellipsoidal parameter set. Feasibility conditions in terms of LMIs are provided in [10, 71] which are equivalent to that a given controller satisfies a weighted 4-block H_∞ performance bound under ellipsoidal parametric system uncertainty.

The S-procedure, see, e.g., [72], is the key tool for obtaining the results above. For multi-input multi-output (MIMO) systems, relaxations based on a linear fractional transformations/multiplier/separation of graphs framework has lead to advances so that bounds on worst case performance of H_∞-norms can be computed under ellipsoidal uncertainty [27]. The same techniques have been applied to robust H_2 deconvolution filter design [28].

It is natural to also use a design method in the application that is robust to ellipsoidal uncertainty. Two such results are the H_∞ state feedback in [40] and the deconvolution filter design method in [28].

6.2. Solving Optimal Experiment Design Problems

There exists an extensive literature on optimal input design, see, e.g., [73–77] and the surveys [26, 78]. Let us for simplicity focus on open loop prediction error identification of linear time-invariant systems. For the model structure (1) where the true system corresponds to $G_o(q) := G(q, \theta^o)$ and $H_o(q) := H(q, \theta^o)$, the average information matrix (32) is given by [7]

$$I_{id}(\theta^o) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \Psi(e^{j\omega})\Psi^*(e^{j\omega})d\omega, \tag{58}$$

where

$$\Psi(z) = \frac{1}{\sqrt{\lambda_e}H_o(z)}\left[\, G'(z,\theta^o)R(z) \quad \sqrt{\lambda_e}H'(z,\theta^o)\,\right],$$

where λ_e is the variance of the innovations and where R is the stable minimum phase spectral factor of the input.

From (58) we see that the only design variable with which the user may influence the average information matrix is the input spectrum. For closed-loop systems it can be shown that also the cross-spectrum between the input and the noise will influence I_{id} [7]. This means that the experiment design can be divided into two steps: (1) First the input spectrum and the aforementioned cross-spectrum are determined by solving an optimization problem. (2) A feedback controller and an input signal are determined that are consistent with these spectra. The signal generation in the second step can be achieved by spectral factorization of the desired input spectrum and then filtering white noise through the stable minimum-phase spectral factor. A key issue in the first step is to obtain a tractable optimization problem. In recent years the focus has been on transforming the problems to semi-definite programs.

Notice from (58) that I_{id} is affine in the input spectrum. Thus, if the input spectrum is linearly parametrized, the matrix inequality (37) becomes an LMI in the decision variables. The use of LMIs in optimal input design problems was first proposed in [13] and has subsequently become standard. We refer to [34] for a comprehensive treatment of the subject.

Although alternatives exists, a common approach to optimal open-loop input design consists of the following steps:

(1) Parametrization of the input spectrum
(2) Choice of objective function, signal constraints, quality constraints and computation of their parametrizations
(3) Conversion of objective function and constraints to convex counterparts.

Notice that with the design variable being a spectrum, it is in general not possible to include time-domain constraints.

(1) Parametrization of the input spectrum: There are several possibilities when it comes to the parametrization of the input spectrum. The starting point is to expand the spectrum in some basis functions $\{\mathcal{B}_k\}$:

$$\Phi_u(e^{j\omega}) = \sum_{k=-\infty}^{\infty} \tilde{c}_k\, \mathcal{B}_k(e^{j\omega}), \tag{59}$$

where now the coefficients $\tilde{c}_k$ are seen as decision variables. One possibility is to use sinusoidal basis functions. In fact a finite number of sinusoids parametrize all possible information matrices [74]. However the frequencies of these sinusoids may depend on the system. There are also other basis functions which allow all possible information matrices to be parametrized using a finite number of terms. But as for sinusoids, these basis functions generally depend on the true system. An exception is FIR models: For an

nth order FIR model the first n autocorrelation coefficients of the input parametrize all information matrices. Relaxing the objective of being able to parametrize all information matrices one may use known basis functions and truncate (59) to get a finite number of decision variables. A common parametrization is

$$\Phi_u(\mathrm{e}^{\mathrm{j}\omega}) = \sum_{k=-m}^{m} \tilde{c}_{|k|} \; {}^{-\jmath\omega k}. \tag{60}$$

Below we will, for simplicity, use this parametrization and $\{\tilde{c}_k\}_{k=0}^{m}$ will be our decision variables. The parametrization must be accompanied by a positivity constraint on (60). This infinite dimensional constraint can be transformed to an LMI by way of the Kalman–Yakubovich–Popov (KYP) lemma [79].

(2) Objective function and constraints: From a least-costly perspective, the objective function should relate to the identification cost. A standard cost is the input energy. However, also the output power, or a weighted version thereof, may be included in the objective function. Constraints may include quality qualifiers such as (37) but also signal constraints. There may also be frequency domain constraints. Constraints and objectives that can be written as weighted L_2 functions of the input spectrum become LMIs in the decision variables. For example, for (60) the constraint

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \Phi_u X(\mathrm{e}^{\mathrm{j}\omega})\mathrm{d}\omega + \tilde{X} \geq 0$$

becomes

$$\sum_{k=-m}^{m} \tilde{c}_{|k|} \; X_k + \tilde{X} \geq 0, \tag{61}$$

where

$$X_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(\mathrm{e}^{\mathrm{j}\omega})\mathrm{e}^{-\jmath\omega k}\mathrm{d}\omega.$$

Also a variety of frequency domain constraints can be converted to LMIs through the use of the KYP-lemma or the generalized KYP-lemma [80].

Alternatively one may use a quality measure as objective function and use energy related functions as constraints only. Recently, it has been shown that the solutions to such problems are equivalent [23].

(3) Quality constraints: Notice that the average information matrix (58) will have a parametrization following the left-hand side of (61). Thus quality constraints such as (37) can be written as LMIs.

When the quadratic approximation (34) is not used, the handling of the quality constraint (27) becomes more intricate. It is desirable to find a characterization of (27) such that it is convex in the decision variables $\{\tilde{c}_k\}$. Returning to the characterizations discussed in Section 6.1, for the family of functions considered in [34], the constraint for one frequency becomes an LMI. The characterization in [40] is also an LMI in the decision variables and the same holds for the quality constraint in [10, 81]. The MIMO constraint developed in [27] becomes a bilinear matrix inequality in the decision variables and the same holds for the robust H_2 deconvolution filter design in [28]. The first contribution that to our knowledge treated a problem related to (27) for non-quadratic functions is [32]. It is shown that the worst-case chordal distance over the set $\mathcal{E}_{id}$ (the supremum of (57) over $\mathcal{E}_{id}$) is quasi-convex in the decision variables.

(4) Optimal closed loop experiment design: It is also possible to include the controller as a decision variable. However, it turns out that the appropriate, and equivalent, decision variable is the Youla-parameter which, when linearly parametrized, allows a wide class of experiment design problems to be formulated as semi-definite programs. We refer to [41] for details.

6.3. Implementation aspects

The Achilles' heel of optimal input design is that the optimization problem depends on the true system. In fact both the the level set $\mathcal{M}^*$ and the confidence ellipsoid $\mathcal{E}_{id}(\Gamma)$ depend typically on the true system (parameters). There are basically two main routes around this:

(1) Robust experiment design
(2) Adaptive (or sequential) experiment design

(1) Robust experiment design: In robust experiment design one tries to design an experiment that is satisfying for all systems in the a priori model set $\mathcal{M}^*$. This is typically done in a worst-case sense, i.e., in a min–max formulation. Indicating the dependence of V_{app} on the true system $\mathcal{S}_o$ by adding a second argument $V_{app} = V_{app}(M, \mathcal{S}_o)$, a robust version of the quality constraint (15) is

$$U^*(\mathcal{S}_o) \subseteq \mathcal{M}^*_{app}(\mathcal{S}_o), \quad \forall \mathcal{S} \in \mathcal{M}^*,$$

where

$$\mathcal{M}^*_{app}(\mathcal{S}_o) = \left\{ M : \; V_{app}(M, \mathcal{S}_o) \leq \frac{1}{2\gamma} \right\}.$$

There are limited results available on robust experiment design. In [82], the expected value of the determinant of the Fisher information matrix is minimized. A max-min approach is taken in [83]. The authors in [10] suggest a scenario approach. Through the use of game-theory methods, [5, 21] establish formal results on min–max designs, which show that $1/f$-noise possesses very interesting robustness properties.

(2) Adaptive experiment design: In adaptive, or sequential, design, one iteratively improves the experiment design as more and more measurements become available. There is a substantial literature on sequential design in the statistics literature, see, e.g., [11] and [26] for references. A framework for adaptive input design for dynamical systems using prediction error identification is presented in [84]. The basic idea is very simple. The input is generated by a filter

$$u(t) = F(q, \beta)w(t),$$

where w is zero mean white noise. Using the certainty equivalence principle, at each point N in time the optimal input filter is computed as a function of the most recent system parameter estimate $\hat{\theta}_N$, i.e., $\beta = \beta(\hat{\theta}_N)$ (the solution is obtained from a semi-definite program as outlined above) and the updated filter $F(q, \beta(\hat{\theta}_N))$ is used to generate one new input sample. Subsequently a new output sample is collected, the model is updated as well as the input filter. This is then repeated. It is argued that under quite general conditions, the same asymptotic statistical properties as for the optimal input design are achieved. In [42], this result is formally established for ARX-models. An early contribution based on high model order approximation of the variance error is [64].

Example 15 ***(Example 14 continued):*** *Using the same settings as in Example 14 the optimal input was replaced with an adaptive algorithm where for the first 30 samples (not shown) white noise with the same variance as the optimal input is used as input in order to obtain an initial parameter estimate. The algorithm then was shifted to adaptive mode with an update of the input filter for every newly acquired data sample. In* Fig. 10, *the average performance degradation cost (13) is shown over 100 Monte Carlo runs when adaptive input design is used. Comparing with* Fig. 9a *we see that the performance of the adaptive algorithm is quite similar to when the optimal input is used.* ■

In practice it is often not necessary to adapt the input design for every sample. It may be sufficient with a few iterations as the following example illustrates.

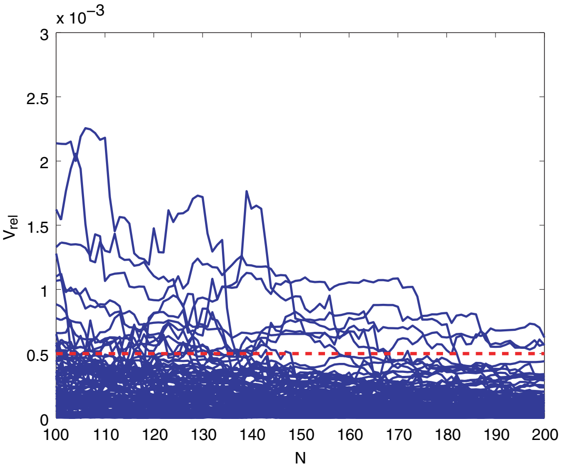

Fig. 10. Example 15: Relative performance degradation cost over 100 Monte Carlo runs when adaptive input design is used. Dashed

Example 16 ***([36]):*** *The process plant is an ARX structure*

$$A(q)y(t) = B(q)u(t) + e(t)$$

with $A(q) = 1 - 1.511q^{-1} + 0.5488q^{-2}$ *and* $B(q) = 0.02059\ q^{-1} + 0.01686\ q^{-2}$. *The sampling time is 10 seconds and* $e(t)$ *has variance 0.01. This is a slight modification of a typical process control application considered in [85]. The process has a rise time of 227 seconds and consequently, as the process response is slow, collecting data samples for the identification takes a long time. Therefore the objective of using optimal input design for this plant is to keep the experiment time below some preset value.*

The optimal design is based on a data length of $N_{opt} = 500$ *and the objective is to design a controller such that the closed loop complementary sensitivity function is given by the dashed line in* Fig. 11. *Also the optimal input spectrum is given in* Fig. 11, *and as a comparison the minimum required power spectrum for a white noise input is shown. Notice that, consistent with the discussion in Section 5.2, the white noise input needs to have a spectrum which is the maximum of the optimal spectrum. Since in this case the required bandwidth is low, the required energy for the white noise input is about 10 times higher than for the optimal input, cf. Example 14.*

To handle the more realistic situation where the true system is unknown, we will replace the optimal design strategy by a two-step procedure. In the first step an initial model is estimated based on a PRBS input[7]. *This model estimate is used as a replacement for the true system in the input design problem. The obtained sub-optimal solution is then applied to the process in the second step.*

[7] PRBS is a periodic, deterministic signal with white-noise-like properties [7].

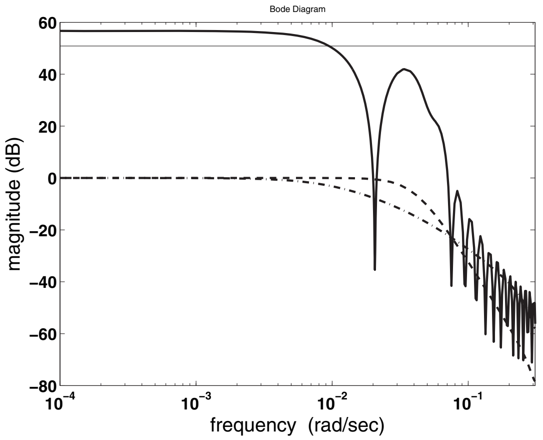

Fig. 11. Example 16. The process plant. Thick solid line: optimal input spectrum. Dashed line: transfer function *T*. Dash-dotted line: open loop system. Thin solid line: white input spectrum.

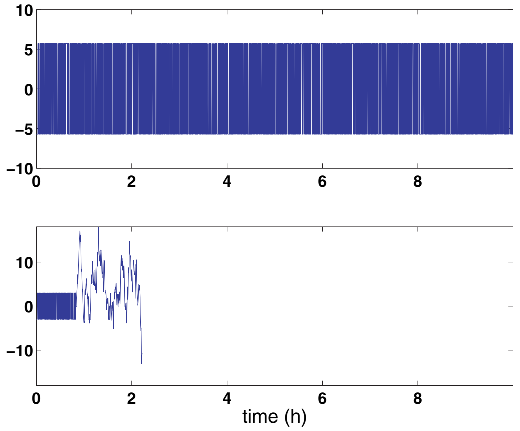

Fig. 12. The process plant. Above: the input sequence not involving optimal input design. Below: the input sequence when involving optimal input design. The first part of the signal is used to identify an initial model estimate. Both signals give the same accuracy.

First consider the two-step adaptive input design approach. We use a PRBS with length $N_{init} = 300$ and amplitude 3 to estimate an initial model estimate G_m of the true system. This model is used for input design with no upper bound on the input spectrum and experiment length $N_{opt} = 500$. This strategy is compared to the approach where a single set of PRBS is used in each Monte-Carlo run. For the comparison's sake the amplitude of the PRBS is tuned so that the signal has the same input power as the average power of the input in the two-step approach. Furthermore, the experiment length is adjusted so that the same confidence level is achieved as with the two-step approach. One realization of the input sequences for both strategies is plotted versus time in hours in Fig. 12. *We clearly see that the experiment time when input design is involved is less than 2 hours and 15 minutes, but more than 10 hours for the PRBS input. We conclude that for the considered quality constraint, the experiment time can be shortened substantially even when a sub-optimal design is used.* ■

As we noted in Section 4.1, it is sometimes possible to estimate a certain property consistently even if a model of restricted complexity is used if the identification cost is matched to the performance degradation cost. When the optimal input depends only on this property, it is sometimes possible to device an adaptive algorithm that estimates this property consistently when the model is of restricted complexity. One such case is the estimation of real valued NMP zeros. Recall from Section 5.4 that, except for a scaling factor, the input (55) depends only on the zero of interest. In [24] it is shown that if an FIR model with only two parameters is estimated with the same adaptive algorithm as outlined above, complemented with a projection mechanism, then the model zero will converge to the desired zero regardless of the complexity of the true system, provided the system has only one NMP zero. We illustrate this with an example from [24].

Example 17: *Consider the system described by*

$$y(t) = \frac{(q-3)(q-0.1)(q-0.2)(q+0.3)}{q^4(q-0.5)}u(t) + \frac{q}{q-0.8}e_o(t),$$

where $\{e_o(t)\}$ is Gaussian white noise of variance 0.01. Notice that the system has exactly one NMP zero at $z_o = 3$.

In order to initialize the algorithm, the first 20 data samples are used to estimate a second order FIR model via the least squares method, with a white noise input signal of variance 1. One realization of the algorithm is shown in Fig. 13. ■

7. Structured Systems

7.1. Introduction

Many systems are highly structured, consisting of interconnected subsystems. Examples range from process industry to sensor networks. It is often of interest to keep the structure of the model consistent with the structure of the system as this will facilitate the interpretation of the model. It is also typically beneficial in terms of model accuracy (recall that the model structure can be seen as a prior), e.g., if it

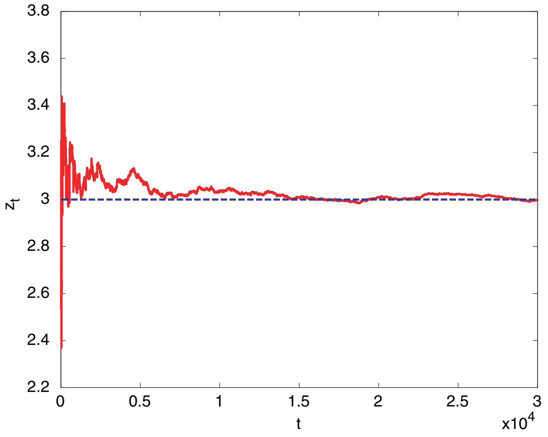

Fig. 13. Example 17. Solid line: One realization of the zero estimate for the adaptive algorithm based on a two parameter FIR model. Dashed line: True location of the zero.

known that the same (physical) parameter appears at several places in the system, the model should reflect this. However, imposing structural information generally compounds the identification problem as the identification criterion typically becomes non-convex. It may also be hard to find good initial parameter values for parameter estimation. There are also identification methods where it is difficult to impose a certain structure, e.g., subspace identification. Thus it is of interest to understand how much accuracy is gained by imposing structural information in the model. Another issue is that there may be a range of actuators and sensors available and it is then of interest to understand how much the use of each one of these can contribute to a model's accuracy.

Example 18: *([86]): Consider the cascade system depicted in* Fig. 14 *given by*

$$y_1(t) = G_1(q,\theta_1)u(t) + e_1(t),$$
$$y_2(t) = G_2(q,\theta_2)G_1(q,\theta_1)u(t) + e_2(t),$$

with two first order FIR transfer functions

$$G_1(t) = 1 + b_1 q^{-1}, \quad \theta_1 = b_1,$$
$$G_2(t) = 1 + b_2 q^{-1}, \quad \theta_2 = b_2.$$

The input signal is white noise with variance λ_u. The true value of the first FIR parameter is $b_1^o = 1$. The noise processes $\{e_1(t)\}$ and $\{e_2(t)\}$ are independent Gaussian white noise stochastic processes, with known variances $\lambda_1 = 1$ and $\lambda_2 = 0.01$, respectively. The second sensor thus provides considerably more accurate measurements than the first sensor. Suppose now that the objective is to estimate b_1^o. The reader may think of b_2^o as a nuisance parameter associated with the second sensor. Since only b_1^o is of interest it is obvious that the second sensor is not necessary for estimating this parameter. However, due to the much better accuracy of the second sensor, it seems intuitive that it would be beneficial to also use this sensor, even if the objective is to estimate b_1^o only.

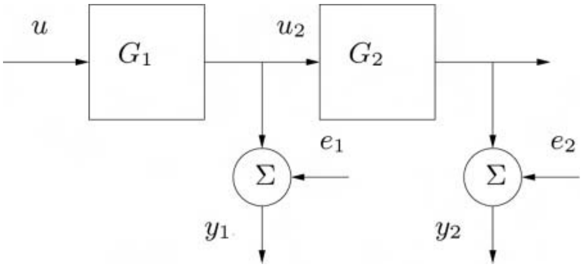

Fig. 14. Cascade system in Example 18.

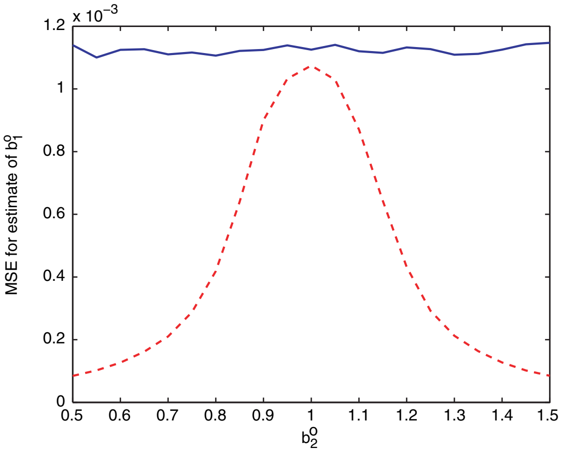

Fig. 15. Example 18. Mean square error of the estimate of b_1^o as a function of the location of b_2^o. Dashed line: y_1 and y_2 used as sensors. Solid line: only y_1 used as sensor.

Now Fig. 15 *shows the mean square error (MSE, computed over 20,000 Monte Carlo runs and with a sample size of $N = 1000$) of the maximum likelihood estimate of b_1^o for two cases:* (1) *Both y_1 and y_2 are used as sensors, and* (2) *only y_1 is used. The MSE is shown as a function of the true value of the second FIR parameter b_2^o. As predicted, the accuracy using both sensors is higher. However, around $b_2^o \approx b_1^o$ the improvement is very small. Thus for systems where $b_2^o \approx b_1^o$, it does not really pay off to use the second sensor even if it is of very high quality.* ■

All the issues discussed above are captured in the covariance matrix $R^\dagger$, see (19), of the parameter estimate. However, unfortunately it is not easy to decode the information "hidden" in $R^\dagger$. If we take prediction error identification as example, the information is captured in the inverse of the average information matrix (32) which in the case of open loop identification of linear time-invariant models is given by (58). Due to the inverse, it is not easy to see how properties such as model structure, input spectrum, model order, etc influence $I_{id}^{-1}(\theta^o)$.

Example 19 ***(Example 18 continued):*** *The average information matrix in Example 18 is given by (58) with*

$$\Psi(z) = \Psi_1(z) := [f(z) \quad 0], \tag{62}$$

where $f(z) = \sqrt{\frac{\lambda_u}{\lambda_1}} z^{-1}$*, when only* y_1 *is used as sensor, and by*

$$\Psi(z) = \Psi_2(z) := \begin{bmatrix} f(z) & \sqrt{\frac{\lambda_u}{\lambda_2}}(1 + b_2^o z^{-1}) z^{-1} \\ 0 & \sqrt{\frac{\lambda_u}{\lambda_2}}(1 + b_1^o z^{-1}) z^{-1} \end{bmatrix} \tag{63}$$

when both y_1 *and* y_2 *are used. It is not obvious that there is something special happening here when* $b_2^o \approx b_1^o$ *as indicated in Fig. 15 without further analysis.*

7.2. A Geometric Framework

A geometric interpretation of the asymptotic covariance matrix has been developed in [65, 87–90] which is helpful for analyzing the impact various quantities have. In order to introduce the concepts that are used we first consider the projection of a vector $\gamma \in \mathbb{R}^{1\times n}$ on the row space of a matrix $\Psi \in \mathbb{R}^{m\times n}$. The projection is given by the standard expression

$$\mathrm{Proj}_{\mathrm{rowspan}\Psi}\{\gamma\} = \gamma\Psi^T\left(\Psi\Psi^T\right)^{-1}\Psi$$

and the squared norm of the projection is given by

$$\|\mathrm{Proj}_{\mathrm{rowspan}\Psi}\{\gamma\}\|^2 = \gamma\Psi^T\left(\Psi\Psi^T\right)^{-1}\Psi\gamma^T.$$

By introducing the inner product[8] $\langle X, Y\rangle = XY^T$, we can write this last expression as

$$\|\mathrm{Proj}_{\mathrm{rowspan}\Psi}\{\gamma\}\|^2 = \langle\gamma, \Psi\rangle\langle\Psi, \Psi\rangle^{-1}\langle\Psi, \gamma\rangle. \tag{64}$$

Now, the starting point for the connection between the asymptotic covariance matrix and this projection result is the observation that the average information matrix can be written as

$$I_{id} = \langle\Psi, \Psi\rangle \tag{65}$$

for some function Ψ with range $\mathbb{C}^n$ (recall that n is the number of parameters) and where $\langle\cdot,\cdot\rangle$ denotes a suitable inner product (for prediction error identification of linear systems the inner product is given by (58)). The next observation is that the asymptotic covariance matrix (which we now denote by P) is the inverse of the information matrix

$$P = \langle\Psi, \Psi\rangle^{-1}.$$

Next, as in Section 2.2, take $\mathcal{J}(M) \in \mathbb{C}$ to denote the system property of interest in the application when the model M is used in the design. Suppose that the model set is parametrized by θ and suppose that $\mathcal{J}(\theta)$ (which is short-hand notation for $\mathcal{J}(M(\theta))$) is a smooth function. Then $\sqrt{N}(\mathcal{J}(\hat{\theta}_N) - \mathcal{J}(\theta^o))$ is asymptotically normally distributed with asymptotic variance given by

$$\begin{aligned}\mathrm{AsVar}\mathcal{J}(\hat{\theta}_N) &= (\mathcal{J}'(\theta^o))^T P \mathcal{J}'(\theta^o) \\ &= (\mathcal{J}'(\theta^o))^T\langle\Psi, \Psi\rangle^{-1}\mathcal{J}'(\theta^o).\end{aligned} \tag{66}$$

Now the key observation is that there exists functions γ[9] such that $\langle\Psi, \gamma\rangle = \mathcal{J}'(\theta^o)$ whereby we can rewrite (66) as

$$\mathrm{AsCov}\mathcal{J}(\hat{\theta}_N) = \langle\gamma, \Psi\rangle\,\langle\Psi, \Psi\rangle^{-1}\,\langle\Psi, \gamma\rangle. \tag{67}$$

But this expression has exactly the form (64) and we arrive at

$$\mathrm{AsCov}\mathcal{J}(\hat{\theta}_N) := \|\mathrm{Proj}_{\mathrm{rowspan}\Psi}\{\gamma\}\|^2.$$

So what are the benefits of this expression? We first return to the cascade example.

Example 20: *(Example 19 continued): Recall the expressions (62) and (63) for* Ψ *that generates the information matrix. With* $\mathcal{J}(\theta^o) = b_1^o$ *being the quantity of interest, it follows that the asymptotic variance of the corresponding estimate is given by*

$$\langle\Psi_1, \Psi_1\rangle^{-1} \quad \text{and} \quad [1 \quad 0]\langle\Psi_2, \Psi_2\rangle^{-1}\begin{bmatrix}1\\0\end{bmatrix} \tag{68}$$

for the cases of one and two sensors, respectively.

Notice now that if $\gamma = [h(z) \quad 0]$ *is such that* $\langle\gamma, \Psi_1\rangle = 1$ *(the precise expression of* γ *is immaterial for our discussion), then* $\langle\gamma, \Psi_2\rangle = [1 \quad 0]$ *and thus the same* γ *can be used to re-write the expressions in (68) as*

$$\|\mathrm{Proj}_{\mathrm{rowspan}\Psi_1}\{\gamma\}\|^2 \quad \text{and} \quad \|\mathrm{Proj}_{\mathrm{rowspan}\Psi_2}\{\gamma\}\|^2. \tag{69}$$

[8] Again we abuse notation. Entry i,j of $\langle X, Y\rangle$ is the inner product between row i of X and row j of Y.

[9] $\gamma = (\mathcal{J}'(\theta^o))^T\langle\Psi, \Psi\rangle^{-1}\Psi$ is one such function, but the exact expression will be of no concern for our discussion.

Next observe that projecting γ on rowspan Ψ_1 or on the joint row-span of Ψ_1 and Ψ_2, the latter given by

$$\begin{aligned}\text{rowspan}\begin{bmatrix} f(z) & 0 \\ f(z) & \sqrt{\frac{\lambda_u}{\lambda_2}}(1+b_2^o z^{-1})z^{-1} \\ 0 & \sqrt{\frac{\lambda_u}{\lambda_2}}(1+b_1^o z^{-1})z^{-1}\end{bmatrix} &= \\ \text{rowspan}\begin{bmatrix} f(z) & 0 \\ 0 & \sqrt{\frac{\lambda_u}{\lambda_2}}(1+b_2^o z^{-1})z^{-1} \\ 0 & \sqrt{\frac{\lambda_u}{\lambda_2}}(1+b_1^o z^{-1})z^{-1}\end{bmatrix}&,\end{aligned} \tag{70}$$

gives the same result. This is due to the structure of γ, i.e. the zero second element. However, the row-span of Ψ_2 is a subspace in the joint row-span of Ψ_1 and Ψ_2. But in view of that the same function γ is projected in the two expressions in (69)*, we conclude that*

$$\|\text{Proj}_{\text{rowspan}\Psi_1}\{\gamma\}\|^2 \geq \|\text{Proj}_{\text{rowspan}\Psi_2}\{\gamma\}\|^2,$$

i.e. we have shown that using two sensors will be no worse than using one sensor. It may seem like we have spent a lot of work establishing this rather evident result. However, it is now also immediate that the asymptotic variances will be equal when $b_1^o = b_2^o$ since then (70) *equals the row-span of Ψ_2:*

$$\begin{aligned}\text{rowspan}\begin{bmatrix} f(z) & 0 \\ 0 & \sqrt{\frac{\lambda_u}{\lambda_2}}(1+b_2^o z^{-1})z^{-1} \\ 0 & \sqrt{\frac{\lambda_u}{\lambda_2}}(1+b_2^o z^{-1})z^{-1}\end{bmatrix} &= \\ \text{rowspan}\begin{bmatrix} f(z) & \sqrt{\frac{\lambda_u}{\lambda_2}}(1+b_2^o z^{-1})z^{-1} \\ 0 & \sqrt{\frac{\lambda_u}{\lambda_2}}(1+b_2^o z^{-1})z^{-1}\end{bmatrix} &= \\ \text{rowspan}\Psi_2.&\end{aligned}$$

Referring to Fig. 15, this explains why there is only minor improvement when using two sensors around $b_2^o = b_1^o$. Notice also that this is a structural property which holds regardless of how small the noise variance λ_2 of the second sensor is (as long as it is positive). ■

The particular behavior around $b_2^o = b_1^o$ in the example is due to that there will be a double root of the polynomial that determines the zeros of the joint system G_1G_2. For a more detailed discussion around this we refer to [86, 91].

7.3. Structural Analysis Using Geometric Tools

Example 20 illustrates that the geometric formulation (67) of the asymptotic variance can be used to deduce structural properties of an identification. Notice that in the example we did not have to determine the function γ that was projected. We only made use of its structure in our derivations. This type of geometric analysis has many applications. In [91] the reader can find results on how adding model parameters and inputs impact a model's accuracy, as well as the difference between open and closed loop identification. The methodology is also used to derive variance expressions for identification of non-linear systems.

(1) *Adding sensors:* The result in Example 20 can be generalized [65, 88]. With G_1 and G_2 being independently parametrized transfer functions and with the input having arbitrary spectrum, it follows that the second sensor is useless for estimating any parameters in G_1 if the row-span of $G_1'G_2$ are in the row-span of G_1G_2'.

(2) *Adding inputs:* A problem dual to the above is to determine when adding inputs can help improve the model quality. This problem is analyzed in [92] using algebraic tools. This problem can also be analyzed using the geometric approach [65, 88]. Consider the model

$$y(t) = G_1(q,\theta,\beta)u^1(t) + G_2(q,\beta)u^2(t) + e(t)$$

where θ and β are unknown parameter vectors, where u^1 and u^2 are two separate inputs to the system, and where $\{e(t)\}$ is white noise. The property of interest is the parameter θ. The user has planned to perform an experiment where the first input u^1 is persistently exciting. If we assume that $G(q,\theta,\beta)$ is globally identifiable it will then be possible to estimate both θ and β using only this input excitation. However, the user also has the possibility of performing an experiment where, in addition to u^1, also u^2 excites the system. The question now is whether this will help improve the estimate of θ. It is immediate that u^2 will be helpful for estimating β since the dynamics G_2 that this input excites depends on β. However, one may suspect that an improved estimate of β in turn will result in an improved estimate of θ. To see this take the extreme case that the power of u^2 is taken very large and that this input is persistently exciting, then $G_2(q,\beta)$ will be estimated exactly which in turn means that all energy of u^1 can be used for estimating θ. Let us now analyze whether this is correct using the geometric tools above. Denoting the stable minimum phase spectral factors of the two inputs by R^1 and R^2, then the two cases when u^1 is used and when both u^1 and u^2 are used correspond to information matrices (65) where

$$\Psi = \Psi_1 := \begin{bmatrix} G_\theta^1 R^1 & 0 \\ G_\beta^1 R^1 & 0, \end{bmatrix}$$

and

$$\Psi = \Psi_2 := \left[\begin{matrix} G^1_\theta R^1 & 0 \\ G^1_\beta R^1 & G^2_\beta R^2 \end{matrix}\right. = \Psi_1 + \begin{bmatrix} 0 & 0 \\ 0 & G^2_\beta R^2 \end{bmatrix},$$

respectively. Consider now a scalar property $\mathcal{J}$ which depends on θ only. Then the function γ should be chosen such that

$$\begin{bmatrix} \mathcal{J}_\theta \\ 0 \end{bmatrix} = \langle \Psi_1, \gamma \rangle = \left\langle \begin{bmatrix} G^1_\theta R^1 & 0 \\ G^1_\beta R^1 & 0 \end{bmatrix}, \gamma \right\rangle, \tag{71}$$

where $\mathcal{J}_\theta = \partial \mathcal{J} / \partial \theta$ evaluated at the true parameter. Now with u^1 persistently exciting it is always possible to find a γ which belongs to the row-span of Ψ_1 such that (71) is satisfied. For such γ we have

$$\langle \Psi_2, \gamma \rangle = \langle \Psi_1, \gamma \rangle + \left\langle \begin{bmatrix} 0 & 0 \\ 0 & G^2_\beta R^2 \end{bmatrix}, \gamma \right\rangle = \langle \Psi_1, \gamma \rangle = \begin{bmatrix} \mathcal{J}_\theta \\ 0 \end{bmatrix}$$

since $\begin{bmatrix} 0 & G^2_\beta R^2 \end{bmatrix}$ is orthogonal to the row-span of Ψ_1 (to which γ belongs). Thus we can use this γ when computing the asymptotic variance of the estimate of $\mathcal{J}$ both when using one input and two inputs. However since we now have constructed γ to belong to the row-span of Ψ_1 the asymptotic variance when only one input is used is given by the squared norm of γ whereas when two inputs are used, γ should first be projected on the row-space of Ψ_2. This projection will obviously not increase the norm and thus we conclude that the asymptotic variance of $\mathcal{J}(\hat{\theta}_N)$ when two inputs are used cannot be larger than when only one input is used. The question now is if there can be equality, i.e. can there be cases when using the second input does not help in improving the accuracy? For this to hold, γ has to belong to the row-space of Ψ_2. Consider the case when $G^1_\theta R^1$ is orthogonal to $G^1_\beta R^1$. Then we see from (71) that γ belongs to the row-space of $\begin{bmatrix} G^1_\theta R^1 & 0 \end{bmatrix}$, but this is a subspace of the row-space of Ψ_2 and thus γ also belongs to this row-space and the asymptotic variances will be equal. The analysis can be extended to show that this orthogonality condition is necessary and sufficient for there to be no function of θ for which using u^2 will improve the accuracy.

Example 21: *Consider the model*

$$y(t) = G^1(q, \theta, \beta) u^1(t) + G^2(q, \beta) u^2(t),$$

where $G^1(q, \theta, \beta) = \sum_{k=1}^n \theta_k q^{-k} + \sum_{k=n+1}^m \beta_k q^{-k}$.

In this case

$$G^1_\theta(z) R^1(z) = \begin{bmatrix} z^{-1} & \ldots & z^{-n} \end{bmatrix}^T R^1(z),$$
$$G^1_\beta(z) R^1(z) = \begin{bmatrix} z^{-(n+1)} & \ldots & z^{-m} \end{bmatrix}^T R^1(z).$$

When u^1 is white noise, these two vectors become orthogonal and thus the estimate of θ will not be improved asymptotically when u^2 is used. This holds regardless of the power of u^2 (as long as it is finite) and how G_2 is parametrized (β may for example be a scalar). ■

(3) *LTI systems:* For prediction error identification of LTI systems, the geometric expression (67) can be reformulated. Consider prediction error identification of the system in Fig. 3 using the model structure (1) and input-output measurements $u(t), y(t)$.

Let $\mathcal{J}(\theta)$ be a two times continuously differentiable real valued scalar function of θ which does not depend on the parameters in the noise model $H(q, \theta)$ and express $\mathcal{J}$ in terms of the system impulse response coefficients $g = \{g_1, g_2, \ldots\}$ according to $\mathcal{J}(\theta) = J_g(g(\theta))$ for some function J_g. The asymptotic variance expression then takes the form

$$\text{As Var } \mathcal{J}(\hat{\theta}_N) = \left\| \text{Proj}_{\text{rowspan}} \Psi \left\{ \nabla J_g \left[\tfrac{\sqrt{\lambda_0} H^*_o}{S^*_o R^*} \quad 0 \right] \right\} \right\|^2, \tag{72}$$

where S_o is the closed loop sensitivity function and where ∇J_g is the z-transform of the sensitivities of J_g:

$$\nabla J_g(z) = \sum_{k=1}^{\infty} \left(\frac{\partial J_g(g^o)}{\partial g_k} \right) z^{-k}$$

(g^o denotes the true impulse response). In (72) we notice that the only quantity that depends on the property of interest is $\nabla J_g(z)$. We also recognize $\sqrt{\lambda_e} H^*_o / (S^*_o R^*)$ as the non-minimum phase unstable spectral factor of the noise to signal ratio Φ_v / Φ^r_u, where Φ^r_u denotes the part of the input spectrum that is due to the external reference signal r. Thus the asymptotic variance depends on the signal to noise ratio. Finally, Ψ is related to the model structure and the experimental conditions.

From (72) we directly obtain bounds on the variance if we remove the projection:

$$\text{AsVar} \mathcal{J}(\hat{\theta}_N) \leq \left\| \nabla J_g \frac{\sqrt{\lambda_0} H^*_o}{S^*_o R^*} \right\|^2 = \| \nabla J_g \|^2_{\frac{\Phi_v}{\Phi^r_u}}.$$

For example the bounds in Table 1 can be derived [65, 87].

The expression has also been used to derive experimental conditions such that the asymptotic variance becomes insensitive to the model and system complexity [20, 65]. This is closely linked to the discussion in Section 3.6 concerning when it is possible to guarantee the model quality constraint regardless of

Table 1. Bounds on the asymptotic variance for some quantities. $\widetilde{G}_o$ is defined in (53).

Property	Variance bound
Real NMP-zero at $z = z_o$	$\frac{1}{\lvert\widetilde{G}_o(z_o)\rvert^2}\left\lVert\frac{1}{1-z_o^{-1}z^{-1}}\right\rVert^2_{\frac{\Phi_v}{\Phi_u^r}}$
$\mathcal{L}_2$-norm	$\lvert G_o\rVert^2_{\frac{\Phi_v}{\Phi_u^r}}/\lVert G_o\rVert^2$
Impulse response coefficient	$\lVert 1\rVert^2_{\frac{\Phi_v}{\Phi_u^r}}$

model structure. The results in [20, 65] generalize the results in Section 5.4 on NMP-zero estimation to other system properties.

8. Conclusions

So what is the essence of our discourse? The main message is that it is important to understand the shape of the level-sets of the performance degradation cost. In hindsight this is not surprising; after all the application should govern the identification set-up. But once the user has a grasp of this issue, the problem is captured by the simple constraint (37) and it is straightforward to compute the cost of complexity and suitable experiment designs, at least when the user has some prior understanding of what the process is (i. e., in the language of the paper, a reasonable guess of θ^o is available.). When the application is designed explicitly based on the model, cf. model reference control, it is straightforward to compute the Hessian $V''_{app}(\theta^o)$. However, there are many applications where this is not trivial at all, examples include optimal control designs such as H_∞, LQG and MPC, and robust filter design. Notice that this is an issue that has to be coped with no matter what experiment design approach that is chosen.

We have also discussed how to cope with a large number of estimated parameters. Here we have pointed out the importance of adapting the confidence set to the level curve of the performance degradation cost. To this end, we have suggested to use confidence ellipsoids of the type (26) with Γ adapted to the performance degradation cost (see Theorem 3.1) and to project the performance degradation cost (see Section 4.3).

The amount of system information that has to be extracted from the system corresponds to the fraction of eigenvalues that dominate the Hessian of the performance degradation cost. This is the factor ξ that we have encountered in the quoted result (7) on the cost of complexity for frequency function estimation and in the model reference application (see Example 14). In fact, (37) suggests the simple rule of thumb that the cost of complexity (measured in terms of required input energy) is proportional to

$$\gamma\,\xi n\,\lambda_e$$

where γ is the desired accuracy, n the total number of parameters and λ_e the noise variance. In the lucky situation that V''_{app} is rank deficient and where the rank does not depend on the system or model complexity, the cost of complexity becomes independent of the system complexity.

We have also been able to connect this framework with earlier studies in identification for control where the benefits of matching the identification criterion to the performance degradation cost of the application were stressed from a bias error perspective. Here we have shown that this concept is inherent in optimal designs, and that this simplifies model structure and model order selection. A key observation is that the scaling of the identification criterion, relative to the performance degradation cost, is important in order to ensure that the desired objective is met. Achieving this requires often proper experiment design.

We have also discussed existing work on how to compute and implement optimal experiment designs and we have argued that adaptive (sequential) designs is a promising avenue for future research. With the new insights that have appeared over the last years on experiment design it is perhaps time to revisit adaptive and dual control?

We have also discussed some aspects regarding identification of structured systems. We have illustrated that there may arise certain structural limitations. In order to uncover these we employed a geometric variance analysis framework recently developed and we have given some glimpses of how this framework can be used.

Acknowledgments

There are many external impulses that have lead to the developments presented in this paper. Lennart Ljung suggested modeling accuracy as my PhD topic and 20 years later I may conclude that the topic turned out to be a tad richer than could fit into one thesis. Maybe now it is time to move on. I have been fortunate to have Laszlo Gerencsér, Michel Gevers, Brett Ninness and Bo Wahlberg as collaborators. Lennart's and their influence cannot be overstated. A special thanks goes to Michel for providing valuable and detailed comments on a quite raw draft. Xavier Bombois, Roland Hildebrand and Cristian Rojas are more recent acquaintances that also deserve credits. The contributions of my former students Kristian

Lindqvist, Henrik Jansson, Jonas Mårtensson and Märta Barenthin have been instrumental. A very special thanks goes to Kristian whose work set everything in motion.

References

1. Ogunnaike BA. A contemporary industrial perspective on process control theory and practice. *Annu Rev Control* 1996; 20: 1–8
2. Hussain MA. Review of the applications of neural networks in chemical process control – simulation and on-line implementation. *Artif Intell Eng* 1999; 13(1): 55–68
3. Ljung L. Asymptotic variance expressions for identified black-box transfer function models. *IEEE Trans Autom Control* 1985; 30(9): 834–844
4. Rojas CR, Welsh JS, Agüero JC. Fundamental limitations on the variance of parametric models. *IEEE Trans Autom Control* 2008, submitted
5. Rojas CR. Robust experiment design. PhD dissertation, The University of Newcastle, June 2008
6. Ninness B. The asymptotic CRLB for the spectrum of ARMA processes. *IEEE Trans Signal Process* 2003; 51(6): 1520–1531
7. Ljung L. *System Identification: Theory for the User*, 2nd edn. Prentice-Hall, Englewood Cliffs, 1999
8. Gevers M, Ljung L. Optimal experiment designs with respect to the intended model application. *Automatica* 1986; 22(5): 543–554
9. Bombois X, Scorletti G, Van den Hof PMJ, Gevers M, Hildebrand R. Least costly identification experiment for control: A solution based on a high-model order approximation. In: American Control Conference, Boston, MA, 2004
10. Bombois X, Scorletti G, Gevers M, Van den Hof PMJ, Hildebrand R. Least costly identification experiment for control. *Automatica* 2006; 42(10): 1651–1662
11. Pronzato L. Adaptive optimization and D-optimum experimental design. *Ann Stat* 2000; 28(6): 1743–1761
12. Forssell U, Ljung L. Some results on optimal experiment design. *Automatica* 2000; 36(5): 749–756
13. Lindqvist K, Hjalmarsson H. Optimal input design using linear matrix inequalities. In: Proceedings of the 12th IFAC Symposium on System Identification, 2000
14. Lindqvist K, Hjalmarsson H. Identification for control: adaptive input design using convex optimization. In: *Conference on Decision and Control*. IEEE, Orlando, December 2001, pp. 4326–4331
15. Belforte G, Gay P. Optimal experiment design for regression polynomial models identification. *Int J Control* 2002; 75(15): 1178–1189
16. Bernaerts K, Servaes RD, Kooyman S, Versyck KJ, Impe JFV. Optimal temperature input design for estimation of the square root model parameters: parameter accuracy and model validity restrictions. *Int J Food Microbiol* 2002; 73(2–3): 145–157
17. Belforte G, Gay P. Optimal input design for set-membership identification of Hammerstein models. *Int J Control* 2003; 76(3): 217–225
18. Jauberthie C, Denis-Vidal L, Coton P, Joly-Blanchard G. An optimal input design procedure. *Automatica* 2006; 42(5): 881–884
19. Stigter JD, Vries D, Keesman KJ. On adaptive optimal input design: a bioreactor case study. *AIChE J* 2006; 52(9): 3290–3296
20. Mårtensson J, Hjalmarsson H. Robustness issues in experiment design for system identification. *IEEE Trans Autom Control* 2009, provisionally accepted
21. Rojas CR, Welsh JS, Goodwin GC, Feuer A. Robust optimal experiment design for system identification. *Automatica* 2007; 43(6): 993–1008
22. Swevers J, Schutter WVJD. Dynamic model identification for industrial robots: integrated experiment design and parameter estimation. *IEEE Control Syst Mag* 2007; 27(5): 58–71
23. Rojas CR, Agüero JC, Welsh JS, Goodwin GC. On the equivalence of least costly and traditional experiment design for control. *Automatica* 2008; 44(11): 2706–2715
24. Rojas C, Hjalmarsson H, Gerencsér L, Mårtensson J. Consistent estimation of real NMP zeros in stable LTI systems of arbitrary complexity. In: 15th IFAC Symposium on System Identification, Saint-Malo, France, July 2009, to appear
25. Franceschini G, Macchietto S. Model-based design of experiments for parameter precision: state of the art. *Chem Eng Sci* 2008; 63(19): 4846–4872
26. Pronzato L. Optimal experimental design and some related control problems. *Automatica* 2008; 44(2): 303–325
27. Barenthin M, Bombois X, Hjalmarsson H, Scorletti G. Identification for control of multivariable systems: controller validation and experiment design via LMIs. *Automatica* 2008; 44(12): 3070–3078
28. Bombois X, Hjalmarsson H. Optimal input design for robust H2 deconvolution filtering. In: 15th IFAC Symposium on System Identification, Saint-Malo, France, July 2009, to appear
29. Barenthin M. Complexity issues, validation and input design for control in system identification. Doctoral thesis, KTH, Stockholm, Sweden, 2008
30. Hildebrand R. Optimal inputs for FIR system identification. In: 47th IEEE Conference on Decision and Control, Cancun, Mexico, 2008, pp. 5525–5530
31. Lindqvist K. On experiment design in identification of smooth systems. Licentiate thesis, Department of Signals, Sensors and Systems, KTH, Stockholm, Sweden, June 2001
32. Hildebrand R, Gevers M. Identification for control: optimal input design with respect to a worst-case ν-gap cost function. *SIAM J Control Optim* 2003; 41(5): 1586–1608
33. Jansson H. Experiment design with applications in identification for control. Doctoral thesis, KTH, Stockholm, Sweden, 2004
34. Jansson H, Hjalmarsson H. Input design via LMIs admitting frequency-wise model specifications in confidence regions. *IEEE Trans Autom Control* 2005; 50(10): 1534–1549
35. Mårtensson J, Jansson H, Hjalmarsson H. Input design for identification of zeros. In: 16th World Congress on Automatic Control, IFAC, Prague, Czech Republik, 2005

36. Barenthin M, Jansson H, Hjalmarsson H. Applications of mixed H∞ and H2 input design in identification. In: 16th World Congress on Automatic Control, IFAC, Prague, Czech Republik, 2005, paper Tu-A13-TO/1
37. Hildebrand R, Solari G. Identification for control: optimal input intended to identify a minimum variance controller. *Automatica* 2007; 43(5): 758–767
38. Barenthin M, Jansson H, Hjalmarsson H, Mårtensson J, Wahlberg B. A control perspective on optimal input design in system identification. In: Forever Ljung in System Identification. Studentlitteratur, Sept. 2006, ch. 10
39. Hjalmarsson H, Mårtensson J. Optimal input design for identification of non-linear systems: learning from the linear case. In: American Control Conference, New York City, USA, July 11–13, 2007
40. Barenthin M, Hjalmarsson H. Identication and control: joint input design and H∞ state feedback with ellipsoidal parametric uncertainty via LMIs. *Automatica* 2008; 44(2): 543–551
41. Hjalmarsson H, Jansson H. Closed loop experiment design for linear time invariant dynamical systems via LMIs. *Automatica* 2008; 44(3): 623–636
42. Gerencsér L, Hjalmarsson H. Identification of ARX systems with non-stationary inputs – asymptotic analysis with application to adaptive input design. *Automatica* 2009; 45(3): 623–633
43. Gevers M. Identification for control: from the early achievements to the revival of experiment design. *Eur J Control* 2005; 4–5: 335–352, 2005, Semi-plenary lecture at IEEE Conference on Decision and Control – European Control Conference
44. Gevers M. A decade of progress in iterative process control design: from theory to practice. *J Process Control* 2002; 12: 519–531
45. Hjalmarsson H. From experiment design to closed loop control. *Automatica* 2005; 41(3): 393–438
46. Zhu Y. System identification for process control: recent progress and outlook. In: 14th IFAC Symposium on System Identification, Newcastle, Australia, March 27–29, 2006, pp. 20–32, plenary address
47. Rivera DE, Morari M. Control-relevant model reduction problems for SISO H2, H∞ and μ-controller synthesis. *Int J Control* 1987; 46(2): 505–527
48. Schrama RJP. Accurate models for control design: the necessity of an iterative scheme. *IEEE Trans Autom Control* 1992; 37(7): 991–994
49. Zang Z, Bitmead RR, Gevers M. Iterative weighted leastsquares identification and weighted LQG control design. *Automatica* 1995; 31: 1577–1594
50. Campi MC, Lecchini A, Savaresi SM. Virtual reference feedback tuning: a direct method for the design of feedback controllers. *Automatica* 2002; 38(8): 1337–1346
51. Hjalmarsson H, Lindqvist K. Identification for control: L2 and L∞ methods. In: IEEE Conference on Decision and Control, Orlando, Florida, USA, December 2001
52. Rojas C, Barenthin M, Hjalmarsson H. The cost of complexity in identification of FIR systems. In: 17th IFAC World Congress, Seoul, South Korea, 2008, pp. 11451–11456
53. Gevers M, Bazanella AS, Miskovic L. Informative data: how to get just sufficiently rich? In: 47th IEEE Conference on Decision and Control, Cancun, Mexico, 2008, pp. 1962–1967
54. Gevers M, Bazanella AS, Bombois X, Miskovic L. Identification and the information matrix: how to get just sufficiently rich? *IEEE Trans Autom Control*, 2009, to appear
55. Gevers M, Bazanella AS, Bombois X. Connecting informative experimetns, the information matrix and the minima of a prediction error identification criterion. In: 15th IFAC Symposium on System Identification, Saint Malo, France, July 6–8, 2009
56. Stoica P, Marzetta TL. Parameter estimation problems with singular information matrices. *IEEE Trans Signal Process* 2001; 49(1), 87–90
57. Campi MC, Ooi SK, Weyer E. Guaranteed non-asymptotic confidence regions in system identification. *Automatica* 2005; 41(10): 1751–1764
58. Van den Hof PMJ, Douma S, den Dekker AJ, Bombois X. Probabilistic model uncertainty bounding in prediction error identification based on alternative test statistics. *Automatica*, 2009, submitted
59. den Dekker AJ, Bombois X, Van den Hof PMJ. Finite sample confidence regions for parameters in prediction error identification using output error models. In: Proceedings of the 17th IFAC World Congress, Seoul, Korea, 2008, pp. 5024–5029
60. Lin J-T. Approximating the cumulative chi-square distribution and its inverse. *The Statistician* 1988; 37(1): 3–5
61. Kullback S. *Information Theory and Statistics*. Wiley, New York, 1959
62. Ljung L, Hjalmarsson H. System identification through the eyes of model validation. In: Proceedings of European Control Conference, Rome, Italy, 1995, pp. 949–954
63. Gantmacher F. *Matrix Theory*. Chelsea Publishing Inc., New York, 1959
64. Hjalmarsson H, Gevers M, De Bruyne F. For model based control design criteria, closed loop identification gives better performance. *Automatica* 1996; 32: 1659–1673
65. Mårtensson J. Geometric analysis of stochastic model errors in system identification, Doctoral thesis, KTH, Stockholm, Sweden, 2007
66. Zehna PW. Invariance of maximum likelihood estimation. *Ann Math Statist* 1996; 37: 755
67. Grenander U, Szegö G. *Toeplitz Forms and Their Applications*. University of California Press, Berkeley, 1958
68. Oppenheim A, Schafer R. *Discrete-Time Signal Processing*. Prentice-Hall, Englewood Cliffs, 1989
69. Bombois X, Anderson BDO, Gevers M. Mapping parametric confidence ellipsoids to Nyquist plane for linearly parametrized transfer functions. In: Goodwin GC (ed), Model Identification and Adaptative Control, Springer Verlag, 2000, pp. 53–71
70. Vinnicombe G. Frequency domain uncertainty and the graph topology. *IEEE Trans Autom Control* 1993; 38: 1371–1382
71. Bombois X, Gevers M, Scorletti G, Anderson BDO. Robustness analysis tools for an uncertainty set obtained by prediction error identification. *Automatica* 2001; 37: 1629–1636

72. Boyd S, El Ghaoui L, Feron E, Balakrishnan V. *Linear Matrix Inequalities in System and Control Theory*. SIAM Studies in Applied Mathematics, Philadelphia, 1994
73. Fedorov VV. *Theory of Optimal Experiments*, ser. Probability and Mathematical Statistics. Academic Press, 1972, vol. 12
74. Goodwin GC, Payne RL. *Dynamic System Identification: Experiment Design and Data Analysis*. Academic Press, New York, 1977
75. Zarrop M. *Optimal Experiment Design for Dynamic System Identification*. Lecture Notes in Control and Information Sciences. Sci. 21, Springer Verlag, Berlin, 1979
76. Pukelsheim F. *Optimal design of experiments*. John Wiley: New York, 1993
77. Fedorov VV, Hackl P. *Model-Oriented Design of Experiments*. Lecture Notes in Statistics; 125. Springer Verlag, Berlin, 1996
78. Walter E, Pronzato L. Qualitative and quantitative experiment design for phenomenological models – a survey. *Automatica* 1990; 26(2): 195–213
79. Yakubovich V. The solution of certain matrix inequalities in automatic control theory. *Soviet Math Dokl* 1962; 3: 620–623
80. Iwasaki T, Hara S. Generalized KYP lemma: unified frequency domain inequalities with design applications. *IEEE Trans Autom Control* 2005; 50(1): 41–59
81. Bombois X, Scorletti G, Anderson BDO, Gevers M, Van den Hof PMJ. A new robust control design procedure based on a PE identification uncertainty set. In: IFAC World Congress, Barcelona, Spain, 2002
82. Pronzato L, Walter E. Robust experiment design via stochastic approximation. *Math Biosci* 1985; 75(1): 103–120
83. Pronzato L, Walter E. Robust experiment design via maximin optimization. *Math Biosci* 1988; 89(2): 161–176
84. Gerencsér L, Hjalmarsson H. Adaptive input design in system identification. In: Proceedings of the 44th IEEE Conference on Decision and Control and European Control Conference, Seville, Spain, December 12–15, 2005, pp. 4988–4993
85. Skogestad S. Simple analytic rules for model reduction and PID controller tuning. *J Process Control* 2003; 13: 291–309
86. Wahlberg B, Hjalmarsson H, Mårtensson J. Variance results for identification of cascade systems. *Automatica* 2009, to appear
87. Mårtensson J, Hjalmarsson H. Variance analysis in SISO linear systems identification. *IEEE Trans Autom Control* 2009, submitted
88. Hjalmarsson H, Mårtensson J. A geometric approach to variance analysis in system identification. *IEEE Trans Autom Control* 2009, submitted
89. Mårtensson J, Hjalmarsson H. A geometric approach to variance analysis in system identification: linear time-invariant systems. In: 46th IEEE Conference on Decision and Control, New Orleans, USA, December 2007, pp. 4269–4274
90. Hjalmarsson H, Mårtensson J. A geometric approach to variance analysis in system identification: theory and nonlinear systems. In: 46th IEEE Conference on Decision and Control, New Orleans, USA, December 2007, pp. 5092–5097
91. Mårtensson J, Hjalmarsson H. Variance error quantifications for identified poles and zeros. *Automatica* 2009, to appear
92. Gevers M, Miskovic L, Bonvin D, Karimi A. Identification of multi-input systems: variance analysis and input design issues. *Automatica* 2006; 42(4): 559–572
93. Xavier J, Barroso V. The Riemannian geometry of certain parameter estimation problems with singular Fisher information matrices. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2004
94. Zacks S. *The Theory of Statistical Inference*. Wiley, 1971

Appendix I

Alternative Approaches to Identification With an Objective

An approach to identification with an objective, closely related to the approach used in this paper, is to try to ensure that the average performance degradation cost of a property of interest is below a given level. As in Section 2.2, take $\mathcal{J}(M)$ to denote the system property of interest in the application when the model M is used in the design and consider the relative performance degradation cost (9). Suppose that the model set is parametrized by θ and suppose that $\mathcal{J}(\theta)$ (which is short-hand notation for $\mathcal{J}(M(\theta))$) is a smooth function. The average relative performance degradation cost is then given by

$$\mathrm{E}\left[(\hat{\theta}_N-\theta^o)^T\left\langle\frac{\mathcal{J}'(\theta^o)}{\mathcal{J}(\theta^o)},\frac{\mathcal{J}'(\theta^o)}{\mathcal{J}(\theta^o)}\right\rangle(\hat{\theta}_N-\theta^o)\right] \\ =\frac{1}{N}Tr\,R^{\dagger}\left\langle\frac{\mathcal{J}'(\theta^o)}{\mathcal{J}(\theta^o)},\frac{\mathcal{J}'(\theta^o)}{\mathcal{J}(\theta^o)}\right\rangle=\frac{1}{N}Tr\,R^{\dagger}V''_{app}(\theta^o)$$

when (19) holds. We can then impose the constraint

$$\mathrm{Tr}R^{\dagger}V''_{app}(\theta^o)\le\frac{N}{\gamma}. \tag{73}$$

When the Euclidean norm is used, the left hand side of (73) can also be interpreted as the variance of $\mathcal{J}(\hat{\theta}_N)$ normalized by the square of the desired quantity $\mathcal{J}(\theta^o)$.

To see the relation between this approach and the approach in this paper we will compare (73) with (37). Factorize

$$V''_{app}(\theta^o)=\Gamma\Gamma^T.$$

Then (73) can be expressed as

$$\mathrm{Tr}\Gamma^T R^{\dagger}\Gamma \leq \frac{N}{\gamma}. \tag{74}$$

With

$$R \geq \frac{\gamma\,\mathrm{Rank}\Gamma}{N}\,\Gamma\Gamma^T = \frac{\gamma\,\mathrm{Rank}\,V''_{app}(\theta^o)}{N}\,V''_{app}(\theta^o) \tag{75}$$

(74) will be satisfied. We see that the condition (75) is very similar to (37). The main difference is that $\chi^2_\alpha(\tilde{n})$ (recall that $\tilde{n}$ is the rank of $V''_{app}(\theta^o)$) is replaced by $\mathrm{Rank}\,V''_{app}(\theta^o)$. However, from (23) we see that these quantities are of the same order. Of course, (75) is only a sufficient condition for (74), but the above clearly shows the connection between the two approaches to robust identification.

Another, but closely related, way of considering the identification problem is to see the model estimation step as a way of condensing the data into a set $U^*(Z^N) \subset \mathcal{M}$ of candidate models *which contains the true system* (with a certain probability). Notice that, as opposed to $U^*(\mathcal{S}_o)$ used in this paper, the set $U^*(Z^N) \subset \mathcal{M}$ depends on the data and hence is stochastic, e.g. it is typically centered around the parameter estimate. Then the application uses some robust design method to account for the uncertainty represented by $U^*(Z^N)$. The objective is now to choose design variables such that a guaranteed performance is achieved for all systems in the set $U^*(Z^N)$. For example, [10] uses this approach. This approach is generally to be preferred since the application explicitly takes the model uncertainty into account. However, for the same reason, and also due to the stochastics involved, it seems much harder to analyze and get insights from than the approach used in the paper.

Appendix II

Non-informative Data Sets and Non-identifiable Model Structures

The minimum amount of input excitation required to obtain a certain accuracy of a certain system property may correspond to an excitation such that the true parameter vector is not identifiable from the data, even if the used model structure is globally identifiable. In this section we shall therefore briefly discuss the statistical properties of a prediction error estimate when the information matrix is singular. ML identification with a singular Fisher information matrix may be treated in a similar way. Formally, Riemannian geometry may be used [91] but here we will take a more pedestrian approach.

Suppose that the model set is defined by some, possibly nonlinear, predictors parametrized by $\theta \in D_{\mathcal{M}} \subset \mathbb{R}^n : W = W(\theta)$. The output predictor is given by $\hat{y}(t|t-1;\theta) = W(Z^t, \theta)$ (recall that $z(t)$ is the joint input-output sample and that $Z^t = \{z(s)\}_{s=1}^t$). Assuming the true system to be in the model set, when the model structure is globally identifiable at the parameter θ^o corresponding to the true system and when

$$I_{id}(\theta^o) = \frac{1}{\lambda_e}\,\mathrm{E}\Big[W'(Z^\infty, \theta^o)(W'(Z^\infty, \theta^o))^T\Big] > 0, \tag{76}$$

where λ_e is the variance of the innovations of the true system and where $W'(Z^t, \theta^o) := \partial W(Z^t, \theta)/\partial\theta$, it holds under mild conditions that

$$\hat{\theta}_N \to \theta^o, \quad \text{w.p.1},$$

$$\sqrt{N}(\hat{\theta}_N - \theta^o) \sim AsN(0, I_{id}^{-1}(\theta^o)), \tag{77}$$

$$N(\hat{\theta}_N - \theta^o)^T\, I_{id}(\theta^o))\,(\hat{\theta}_N - \theta^o) \sim As\chi^2(n). \tag{78}$$

Now, let $Z^\infty = \{z(1), z(2), \ldots\}$ be a given data set and let $W(Z^t)$ and $\bar{W}(Z^t)$ denote two predictors corresponding to two different models. In prediction error identification with a quadratic cost function, it is only possible to discriminate these two models if

$$\mathrm{E}\Big[(W(Z^t) - \bar{W}(Z^t))^2\Big] \neq 0. \tag{79}$$

Example 22: *Consider the model*

$$y(t) = \sum_{k=1}^{n} \theta_k u(t-k) + e(t).$$

If $u(t) \equiv u$ (constant), then all models with the same $\sum_{k=1}^n \theta_k$ give rise to the same predictor and are indistinguishable. ■

There are two reasons for why two parameters $\theta \neq \bar{\theta}$ may give rise to the same prediction error. First, it may be due to that the model structure is not identifiable. Second, as in Example 22, it may be due to that the data set Z^∞ is not sufficiently informative with respect to the model structure. In this case it is only guaranteed that $\hat{\theta}_N$ converges to a set corresponding to the best predictor for the data set Z^∞. When the true system belongs to the model set, but the data set is not informative, i.e., (79) does not hold for two different predictors, this set will not only include the predictor that is optimal for all types of excitation but also other predictors optimal only for the data set Z^{∞}. We will now provide an interpretation to (77)–(78) in this case.

Now let us lump together equivalent predictors for Z^{∞}, the data set under consideration, by introducing $\eta = \eta(\theta) \in \mathbb{R}^{nz}, n_Z \leq n$, and corresponding predictors $\widetilde{W}(\eta)$ such that

$$\begin{aligned}&\forall\theta \in D_{\mathcal{M}},\ \exists\eta = \eta(\theta):\\ &\mathrm{E}\Big[\big(W(Z^t,\theta) - \widetilde{W}(Z^t,\eta)\big)^2\Big] = 0.\end{aligned} \tag{80}$$

We will make the assumptions that it is possible to construct $\eta(\theta)$ such that

(1) $\eta(\theta)$ is differentiable.
(2) $\widetilde{W}(\eta)$ is differentiable.
(3) There is a unique $\eta := \eta^o$ corresponding to the optimal predictor for the true system for the data set Z^{∞}.
(4) It holds

$$I_{id,\eta} := \frac{1}{\lambda_e}\,\mathrm{E}\Big[\widetilde{W}'(Z^{\infty},\eta(\theta^o))(\widetilde{W}'(Z^{\infty},\eta(\theta^o)))^T\Big] > 0.$$

where $\widetilde{W}'(Z^t,\eta) := \partial\widetilde{W}(Z^t,\eta)/\partial\eta$.

Notice that $\eta^o = \eta(\theta^o)$ and that this quantity depends on the properties of the data set Z^{∞} under consideration. The last two assumptions imply that the model structure defined by $\widetilde{W}(\eta)$ is locally identifiable at η^o. For models that are linear in the parameters it is easy to construct η: Let I_{id} have rank n_Z with eigendecomposition $I_{id} = EDE^T$ with $D \in \mathbb{R}^{n_Z \times n_Z}$ non- singular. Then take $\eta = E^T\theta$ and $\widetilde{W}(\eta) := W(E\eta)$.

Now let $\hat{\eta}_n$ be the prediction error estimate corresponding to the model structure defined by $\widetilde{W}(\eta)$. Under mild conditions it holds that

$$\hat{\eta}_N \to \eta^o, \quad \text{w.p.1}, \tag{81}$$

$$\sqrt{N}(\hat{\eta}_N - \eta^o) \sim AsN\big(0, I_{id}, \eta^{-1}(\eta^o)\big), \tag{82}$$

where I_{id}, η is defined similar to (76), with $\widetilde{W}_{\eta}$ replacing W_{θ}.

Under the assumptions above, we have also that $\hat{\theta}_N$ will converge to the set $\Theta^* = \{\theta : \eta(\theta) = \eta^o\}$. Below θ^* denotes an arbitrary parameter from this set.

Now we observe that since $W(Z^t,\theta) = \widetilde{W}(Z^t,\eta(\theta))$ (in a mean-square sense according to (80)), it holds that

$$\frac{\partial W(Z^t,\theta^*)}{\partial\theta} = \Lambda^T\frac{\partial\widetilde{W}(Z^t,\eta(\theta^*))}{\partial\eta},$$

where $\Lambda = \frac{d\eta(\theta)}{d\theta}|_{\theta=\theta^*}$, and hence

$$I_{id}(\theta^*) = \Lambda^T I_{id,\eta}(\eta^o)\Lambda. \tag{83}$$

Notice that $I_{id}(\theta^*)$ is singular when $n_Z < n$ (by assumption). From (83)

$$I^{-1}_{id,\eta}(\eta^o) = \Lambda\, I^{\dagger}_{id}(\theta^*)\,\Lambda^T. \tag{84}$$

Now the only functions $\mathcal{J}(\theta)$ of the system parameters θ that are identifiable, are those that are functions of η, i.e. when $\mathcal{J}(\theta) = \widetilde{\mathcal{J}}(\eta(\theta))$. For such a function it holds that

$$\sqrt{N}(\mathcal{J}(\hat{\theta}_N) - \mathcal{J}(\theta^*)) \sim AsN(0, \mathrm{AsCov}\mathcal{J}(\hat{\theta}_N)), \tag{85}$$

where the asymptotic covariance matrix is given by Gauss approximation formula:

$$\begin{aligned}\mathrm{AsCov}\mathcal{J}(\hat{\theta}_N) &= \mathrm{AsCov}\widetilde{\mathcal{J}}(\hat{\eta}_N)\\ &= \Big[\widetilde{\mathcal{J}}'(\eta^o)\Big]^T I_{id},\eta^{-1}(\eta^o)\widetilde{\mathcal{J}}'(\eta^o).\end{aligned}$$

But inserting (84) gives

$$\begin{aligned}\mathrm{AsCov}\mathcal{J}(\hat{\theta}_N) &= \Big[\Lambda^T\widetilde{\mathcal{J}}'(\eta^o)\Big]^T I^{\dagger}_{id}(\theta^*)\Lambda^T\widetilde{\mathcal{J}}'(\eta^o)\\ &= [\mathcal{J}'(\theta^*)]^T I^{\dagger}_{id}(\theta^*)\mathcal{J}'(\theta^*).\end{aligned} \tag{86}$$

From (85) and (86) we see that even if θ^o is not identifiable, we can still use the use the formula (77) if it only is applied to functions of θ which are identifiable and if $I^{-1}_{id}(\theta^o)$ is replaced by the pseudo-inverse $I^{\dagger}_{id}(\theta^o)$ (we can take $\theta^* = \theta^o$).

In particular notice that

$$\begin{aligned}&(\hat{\theta}_N - \theta^o)^T I_{id}(\theta^o)(\hat{\theta}_N - \theta^o)\\ &= \Big[\Lambda(\hat{\theta}_N - \theta^o)\Big]^T I_{id,\eta}(\eta^o)\Lambda(\hat{\theta}_N - \theta^o)\\ &\approx (\hat{\eta}_N - \eta^o)^T I_{id,\eta}(\eta^o)(\hat{\eta}_N - \eta^o).\end{aligned}$$

Thus, even though $\hat{\theta}_N$ is not well defined, it holds that $N(\hat{\theta}_N - \theta^o)^T I_{id}(\theta^o)(\hat{\theta}_N - \theta^o)$ has the same asymptotic distribution as $(\hat{\eta}_N - \eta^o)^T I_{id,\eta}(\eta^o)(\hat{\eta}_N - \eta^o)$. In view of (82) we thus have

$$N(\hat{\theta}_N - \theta^o)^T I_{id}(\theta^o)(\hat{\theta}_N - \theta^o) \sim \chi^2(n_Z)$$

rather than (78) when we do not have identifiability in the original parametrization.

Next notice that n_Z, the number of identifiable parameters from Z^∞, can be determined from the rank of $I_{id}(\theta^o)$. This follows from (83) where $I_{id}, \eta(\eta^o)$ has full rank.

Summarizing, the notation

$$\sqrt{N}(\hat{\theta}_N - \theta^o) \sim AsN\Big(0, I_{id}^{\dagger}(\theta^o)\Big)$$

implies that the estimate of any identifiable quantity $\mathcal{J}(\theta)$ has asymptotic distribution

$$\begin{aligned}&\sqrt{N}(\mathcal{J}(\hat{\theta}_N) - \mathcal{J}(\theta^o))\\ &\quad \sim AsN\Big(0, [\mathcal{J}'(\theta^o)]^T I_{id}^{\dagger}(\theta^o)\mathcal{J}'(\theta^o)\Big).\end{aligned}$$

Furthermore, the rank of $\mathcal{J}'(\theta^o)$ for any such quantity is at most Rank $I_{id}(\theta^o)$.

Appendix III

Proof of Theorem 3.1

We start with proving that $\mathcal{E}_{id}(\Gamma) \subseteq \mathcal{E}_{app}$ if and only if

$$\Gamma\big(\Gamma^T R^{\dagger}\Gamma\big)^{\dagger}\Gamma^T \geq \frac{\gamma\chi_{\alpha}^2(m)}{N} V''_{app}(\theta^o). \tag{87}$$

This follows from the following lemma.

Lemma 3.1: *Let* $T_1 \geq 0$ *and* $T_2 \geq 0$. *Then*

$$\theta^T T_1 \theta \leq 1 \quad \Rightarrow \quad \theta^T T_2 \theta \leq 1 \tag{88}$$

and

$$T_1 \geq T_2$$

are equivalent.

Proof: The proof of the lemma is almost trivial. It is also a direct consequence of the S-procedure [72]. However, since it is important for our arguments we will prove the result.

$\Rightarrow$: Consider first a $\theta = \tilde{\theta}$ such that $T_1\tilde{\theta} = 0$ (if such a θ exists). Suppose that for $\tilde{\theta}, \tilde{\theta}^T T_2 \tilde{\theta} = v > 0$. Then take $\theta = 2\tilde{\theta}/\sqrt{v}$. Then $\theta^T T_1 \theta = 0 \leq 1$ but $\theta^T T_2 \theta = \frac{4}{v}\tilde{\theta}^T T_2 \tilde{\theta} = 4 > 1$. Thus (88) is not satisfied. Hence we have shown that T_2 must share all eigenvectors of T_1 corresponding to zero eigenvalues.

Consider now a $\theta = \tilde{\theta}$ for which $T_1\tilde{\theta} \neq 0$. Suppose now that for $\tilde{\theta}, \tilde{\theta}^T(T_1 - T_2)\tilde{\theta} < 0$. This implies that $\tilde{\theta}^T T_1 \tilde{\theta} < \tilde{\theta}^T T_2 \tilde{\theta}$. Now let θ be a scaled version of $\tilde{\theta}$ such that $\theta^T T_1 \theta = 1$, but then the preceding observation implies $\theta^T T_2 \theta > 1$ and hence (88) is violated. Thus for (88) to hold for θ not in the kernel of T_1, we must have $\theta^T(T_1 - T_2)\theta \geq 0$.

Summing up the two results above, we have shown that (88) implies (89).

$\Leftarrow$: Suppose that θ is such that the left part of (88) holds. Then $\theta^T T_2 \theta = \theta^T T_1 \theta + \theta^T(T_2 - T_1)\theta \leq 1 + 0 \leq 1$, i.e. the right part of (88) holds. This concludes the proof. ■

To simplify the notation and the derivations, we assume that $\widetilde{\Gamma}\widetilde{M}\widetilde{\Gamma}^T$ is an eigenvalue decomposition of $\frac{\gamma\chi_{\alpha}^2(m)}{N} V''_{app}(\theta^o)$ with $\widetilde{\Gamma}^T\widetilde{\Gamma} = I$ and $\widetilde{M} > 0$. Notice that the rank of $\widetilde{\Gamma}$ equals the rank of Γ in the theorem. Thus (87) is equivalent to

$$\widetilde{\Gamma}\left(\left(\widetilde{\Gamma}^T R^{\dagger}\widetilde{\Gamma}\right)^{\dagger} - \widetilde{M}\right)\widetilde{\Gamma}^T \geq 0. \tag{90}$$

Now we proceed and observe that (90) is equivalent to

$$\left(\widetilde{\Gamma}^T R^{\dagger}\widetilde{\Gamma}\right)^{\dagger} - \widetilde{M} \geq 0. \tag{91}$$

Clearly (91) implies (90) and multiplying (90) from left by $\widetilde{\Gamma}^T$ and from right by $\widetilde{\Gamma}$ we see that (90) implies (91). Next observe that $\widetilde{M}$ is non-singular and hence (91) implies that $\left(\widetilde{\Gamma}^T R^{\dagger}\widetilde{\Gamma}\right)^{\dagger}$ is nonsingular as well, which in turn is equivalent to that

$$\widetilde{\Gamma}^T R^{\dagger}\widetilde{\Gamma} > 0. \tag{92}$$

This inequality is equivalent to that $\widetilde{\Gamma}$ is in the range space of R. To see this let $R = EDE^T$ be an eigenvalue decomposition of R with $E^TE = I$ and $D > 0$, so that $R^{\dagger} = ED^{-1}E^T$. Now, for (92) to hold it is clear that $\widetilde{\Gamma}$ must be in the range of E which is the range of R. But $\widetilde{\Gamma}$ being in the range of R is equivalent to $\widetilde{\Gamma}^T(I - RR^{\dagger}) = 0$ since

$$\begin{aligned} 0 = \widetilde{\Gamma}^T(I - RR^{\dagger}) &= \widetilde{\Gamma}^T - \widetilde{\Gamma}^T EDE^T ED^{-1}E^T \\ &= \widetilde{\Gamma}^T - \widetilde{\Gamma}^T EE^T \\ &\Leftrightarrow \\ \widetilde{\Gamma} &= EE^T\widetilde{\Gamma}, \end{aligned}$$

and EE^T is the projection operator onto the range of E. Summarizing we have so far shown that (87) is equivalent to

$$\begin{aligned} \widetilde{M}^{-1} - \widetilde{\Gamma}^T R^{\dagger}\widetilde{\Gamma} &\geq 0, \\ R^{\dagger} &\geq 0, \\ \widetilde{\Gamma}^T\big(I - RR^{\dagger}\big) &= 0. \end{aligned}$$

Using Schur complement [72] this is equivalent to

$$\begin{bmatrix} \widetilde{M}^{-1} & \widetilde{\Gamma}^T \\ \widetilde{\Gamma} & R \end{bmatrix} \geq 0,$$

and using Schur complement again gives that this is equivalent to

$$\begin{aligned} R - \widetilde{\Gamma}\widetilde{M}\widetilde{\Gamma}^T &\geq 0, \\ \widetilde{M}^{-1} &\geq 0, \\ \widetilde{\Gamma}\left(I - \widetilde{M}^{-1}\widetilde{M}\right) &= 0, \end{aligned}$$

where the last to conditions are trivially satisfied. Now observing that $\widetilde{\Gamma}\widetilde{M}\widetilde{\Gamma}^T = \frac{\gamma\chi^2_\alpha(m)}{N} V''_{app}(\theta^o)$ concludes the proof.

Appendix IV

Model Selection for Linear Regression Problems

In this appendix we will consider a linear regression problem, with the true system given by

$$Y = \Phi\theta^o + W, \quad Y \in \mathbb{R}^N,\ \theta^o \in \mathbb{R}^n,$$

where $W \sim N(0, \lambda_e I)$. We will assume the noise variance λ_e to be known. We will also assume that the regressor vector Φ is deterministic.

Taking into account that we now use finite data this means that

$$V_{app}(\theta) = \frac{1}{\gamma\chi^2_\alpha(n)\lambda_e}(\theta - \theta^o)\Phi^T\Phi(\theta - \theta^o).$$

We will assume that an "oracle" with complete knowledge of the true system has designed the identification experiment so that the identification criterion matches the performance degradation cost. The question is thus what we can do with the given data.

We start our analysis with defining the prediction error

$$\mathcal{E}(\theta) = Y - \Phi\theta = \Phi(\theta^o - \theta) + W$$

and the corresponding quadratic cost

$$\begin{aligned} V_{LS}(\theta) &:= \mathcal{E}(\theta)^T\mathcal{E}(\theta) \\ &= (\theta - \theta^o)\Phi^T\Phi(\theta - \theta^o) + W^TW + 2(\theta^o - \theta)^T\Phi^TW \\ &= \gamma\chi^2_\alpha(n)\lambda_e V_{app}(\theta) + W^TW + 2(\theta^o - \theta)^T\Phi^TW. \end{aligned} \tag{93}$$

Thus for a given θ, $V_{LS}(\theta)$ provides an observation of the corresponding performance degradation cost, but a noisy one. Since we would like to make V_{app} small it is natural to use the "observer" V_{LS} for the optimization. However, then care has to be exercised since as soon as we start choosing θ depending on what we observe, we introduce correlations since V_{LS} is a noisy observation. Since we are interested in minimizing the performance degradation cost it is natural to use the parameter that minimizes V_{LS}. We will consider different model orders so first we make the partitions

$$\begin{aligned} \Phi &= \begin{bmatrix} \widetilde{\Phi} & \Phi_e \end{bmatrix}, \quad \widetilde{\Phi} \in \mathbb{R}^{N\times\tilde{n}}, \\ \theta^o &= \begin{bmatrix} \widetilde{\theta}^o & \theta^o_e \end{bmatrix}, \quad \widetilde{\theta}^o \in \mathbb{R}^{\tilde{n}}. \end{aligned}$$

The least-squares estimate of η in the model

$$Y = \widetilde{\Phi}\eta$$

is given by

$$\begin{aligned} \hat{\eta} &= \tilde{R}^{-1}\widetilde{\Phi}^TY \\ &= \tilde{R}^{-1}\widetilde{\Phi}^T\left[\begin{bmatrix} \widetilde{\Phi} & \Phi_e \end{bmatrix}\begin{bmatrix} \widetilde{\theta}^o & \theta^o_e \end{bmatrix} + W\right] \\ &= \widetilde{\theta}^o + \tilde{R}^{-1}\widetilde{\Phi}^T\Phi_e\theta^o_e + \tilde{R}^{-1}\widetilde{\Phi}^TW, \end{aligned}$$

where $\tilde{R} = \left[\widetilde{\Phi}^T\widetilde{\Phi}\right]$.

Introducing

$$\theta(\hat{\eta}) = \begin{bmatrix} \hat{\eta} \\ 0 \end{bmatrix},$$

we obtain

$$\theta^o - \theta(\hat{\eta}) = \begin{bmatrix} -\tilde{R}^{-1}\widetilde{\Phi}^T\Phi_e\theta^o_e - \tilde{R}^{-1}\widetilde{\Phi}^TW \\ \theta^o_e \end{bmatrix}$$

and

$$\begin{aligned} &(\theta^o - \theta(\hat{\eta}))^T\Phi^TW = \\ &\begin{bmatrix} -(\theta^o_e)^T\Phi^T_e\widetilde{\Phi}\tilde{R}^{-1} - W^T\widetilde{\Phi}\tilde{R}^{-1} & (\theta^o_e)^T \end{bmatrix}\begin{bmatrix} \widetilde{\Phi}^T \\ \Phi^T_e \end{bmatrix}W \end{aligned}$$

in turn. Plugging these expressions into (93) gives

$$\begin{aligned} &V_{LS}(\theta(\hat{\eta})) \\ &= \gamma\chi^2_\alpha(n)\lambda_e V_{app}(\theta(\hat{\eta})) + W^TW + 2(\theta^o - \hat{\theta})^T\Phi^TW \\ &= \gamma\chi^2_\alpha(n)\lambda_e V_{app}(\theta(\hat{\eta})) + W^TW - 2W^T\widetilde{\Phi}\widetilde{R}^{-1}\widetilde{\Phi}^TW \\ &+ 2(\theta^o_e)^T\Phi^T_e\left[I - \widetilde{\Phi}\widetilde{R}^{-1}\widetilde{\Phi}^T\right]W \\ &= \gamma\chi^2_\alpha(n)\lambda_e V_{app}(\theta(\hat{\eta})) - W_1 + W_2 + W_3, \end{aligned}$$

where

$$\begin{aligned} W_1 &= W^T\widetilde{\Phi}\widetilde{R}^{-1}\widetilde{\Phi}^T W,\\ W_2 &= W^T W - W_1,\\ W_3 &= 2(\theta_e^o)^T\Phi_e^T\Big[I - \widetilde{\Phi}\widetilde{R}^{-1}\widetilde{\Phi}^T\Big]W. \end{aligned}$$

Notice that

$$\begin{aligned} &\frac{1}{\lambda_e}W_1 \sim \chi^2(\tilde{n}),\\ &\frac{1}{\lambda_e}W_2 \sim \chi^2(N-\tilde{n}),\\ &\quad W_3 \sim N\Big(0, 4\lambda_e(\theta_e^o)^T\Phi_e^T\Big[I - \widetilde{\Phi}\widetilde{R}^{-1}\widetilde{\Phi}^T\Big]\Phi_e\theta_e^o\Big), \end{aligned}$$

and that all these three variables are independent of each other. Notice also that the variance of W_3 can be expressed as

$$4\lambda_e V_{app}(\theta(\eta^*)),$$

where $\theta(\eta^*)$ is the minimizer of $V_{app}(\theta(\eta))$.

Thus $V_{LS}(\hat{\theta})$ is an observation of $\gamma\chi_\alpha^2(n)\,V_{app}(\hat{\theta})$ in noise with mean value $\lambda_e(N-\tilde{n}) - \lambda_e\tilde{n}$. Thus an unbiased estimate of $V_{app}(\hat{\theta})$ is given by

$$\hat{V}_{app}(\hat{\theta}) = \frac{V_{LS}(\hat{\theta}) + 2\lambda_e\tilde{n} - N\lambda_e}{\gamma\chi_\alpha^2(n)\lambda_e}.$$

Notice that the model order dependent term is

$$V_{LS}(\hat{\theta}) + 2\lambda_e\tilde{n}$$

and that this expression corresponds to AIC (when, as we assume, the noise variance λ_e is known). Thus AIC will provide us with an estimate of which least-squares estimate in a sequence of of increasing order that will give the smallest performance degradation cost!

Appendix V

Minimizing the Experimental Cost

An eigenvalue decomposition gives

$$V''_{app}(\theta^o) = EDE^T = E_a D_a E_a^T + E_\Delta D_\Delta E_\Delta^T,$$

where D_a and D_Δ are diagonal matrices containing the eigenvalues of $V''_{app}(\theta^o)$ and where $[E_a \quad E_\Delta]$ $[E_a \quad E_\Delta]^T = I$. Now take

$$\tilde{V}''_{app}(\theta^o) := E_a D_a E_a^T$$

and design an experiment such that (37) holds with $V''_{app}(\theta^o)$ replaced by $\tilde{V}''_{app}(\theta^o)$ and where m is the rank of $\tilde{V}''_{app}(\theta^o)$. This will ensure that $\hat{\theta}_N$ will end up in the set

$$(\theta - \theta^o)^T\tilde{V}''_{app}(\theta^o)(\theta - \theta^o) \leq \frac{1}{\gamma}$$

with given probability α. But our concern is with what we can say about

$$(\hat{\theta}_N - \theta^o)^T V''_{app}(\theta^o)(\hat{\theta}_N - \theta^o).$$

Now the experiment design has not taken V_{app} into account which means that the model may have tried to pick up system properties from data which are relevant for V_{app} but not for $\tilde{V}_{app}$, and that this has been done poorly since the experiment was not designed for V_{app}. To overcome this problem one should project the estimate on the range space of $\tilde{V}''_{app}(\theta^o)$, i.e. one should use

$$\hat{\theta}_N^{proj} = E_a E_a^T\hat{\theta}_N.$$

For this estimate we get

$$\begin{aligned} &(\hat{\theta}_N^{proj} - \theta^o)^T V''_{app}(\hat{\theta}_N^{proj} - \theta^o)\\ =&\Big(E_a E_a^T(\hat{\theta}_N - \theta^o) - E_\Delta E_\Delta^T\theta^o\Big)^T\left(E_a D E_a^T + E_\Delta D_\Delta E_\Delta^T\right)\\ &\Big(E_a E_a^T(\hat{\theta}_N - \theta^o) - E_\Delta E_\Delta^T\theta^o\Big)\\ =&(\hat{\theta}_N - \theta^o)^T\tilde{V}''_{app}(\theta^o)(\hat{\theta}_N - \theta^o) + (\theta^o)^T E_\Delta D_\Delta E_\Delta^T\theta^o. \end{aligned}$$

Thus we have that

$$\begin{aligned} &(\hat{\theta}_N^{proj} - \theta^o)^T V''_{app}(\theta - \theta^o)(\hat{\theta}_N^{proj} - \theta^o)\\ &\leq \frac{1}{\gamma} + (\theta^o)^T E_\Delta D_\Delta E_\Delta^T\theta^o \end{aligned} \tag{94}$$

with probability α. Clearly $E_\Delta D_\Delta E_\Delta^T$ should be selected so that the second term in (94) is minimized. Formally, we would like to approximate $V''_{app}(\theta^o)$ with a symmetric positive semi-definite matrix of some rank m (this is EDE^T) such that $(\theta^o)^T(V''_{app}(\theta^o) - EDE^T)\theta^o$ is minimized. Writing $\theta^o = [E \quad E_\Delta]\,\alpha$ gives

$$(\theta^o)^T V''_{app}(\theta^o)\theta^o = \alpha^T\begin{bmatrix} D & \\ 0 & D_\Delta\end{bmatrix}\alpha = \sum_{k=1}^n \alpha_k^2\lambda_k$$

from which we see that if we order the eigenvalue decomposition $V''_{app}(\theta^o)$ in descending order of $\alpha_k^2\,\lambda_k$

$$(\theta^o)^T E_\Delta D_\Delta E_\Delta^T\theta^o = \sum_{k=m+1}^n \alpha_k^2\,\lambda_k$$

is minimized. It is now straightforward to derive the properties of the algorithm in Section 4.3.

Appendix VI

A Separation Principle

A fundamental question related to how to identify complex systems is whether it is necessary to model the full system. Sometimes it is argued that it is better to use a biased model when it is possible to reduce the variance error more than the bias error increases, resulting in a net decrease of the mean square error.

Let $\hat{\theta}_{ML}$ be the ML-estimate of $\theta \in D_{\mathcal{M}} \in \mathbf{R}^n$ and let $f : D_{\mathcal{M}} \to \Omega \subset \mathbf{R}^m$ with $m \leq n$. It then holds that $f(\hat{\theta}_{ML})$ is the ML-estimate of $f(\theta)$. This is the so called invariance principle for ML-estimation [66] (Theorem 5.1.1 in [94]). Hence, it follows under very general conditions on f that if $\hat{\theta}_{ML}$ is asymptotically efficient, i. e. it is consistent and its asymptotic covariance matrix reaches the Cramér-Rao lower limit [7], then $f(\hat{\theta}_{ML})$ is also asymptotically efficient.

Consider a model set parametrized by η, which does not contain the true system. Using data, an estimate $\eta(Z^N)$ is produced which converges to some point η^* as $N \to \infty$. When N is sufficiently large, the performance degradation cost associated with using $\eta(Z^N)$ can be approximated by

$$\begin{aligned} \mathrm{E}\left[J(\eta(Z^N))\right] &\approx J(\eta^*) + J'(\eta^*)\mathrm{E}\left[(\eta(Z^N) - \eta^*)\right] \\ &+ \frac{1}{2}\mathrm{E}\left[(\eta(Z^N) - \eta^*)^T J''(\eta^*)(\eta(Z^N) - \eta^*)\right] \\ &\approx J(\eta^*) + \frac{1}{2} Tr J''(\eta^*)\mathrm{E}\left[(\eta(Z^N) - \eta^*)(\eta(Z^N) - \eta^*)^T\right]. \end{aligned}$$

Notice that since $\eta(Z^N)$ depends on the data which in turn depends on the true parameter vector θ^o, it holds that $\eta^* = \eta^*(\theta^o)$. Thus we can see η^* as a system property that we are trying to estimate consistently. But then the invariance principle gives that the estimate with smallest covariance matrix is given by $\eta^*(\hat{\theta}_{ML})$, at least as the sample size becomes large. This implies

$$\lim_{N\to\infty} N\Big(\mathrm{E}\left[J(\eta(Z^N))\right] - \mathrm{E}\left[J(\eta^*(\hat{\theta}_{ML}))\right]\Big) \geq 0,$$

The interpretation is that the performance of an application based on a model set of restricted complexity can be improved by first doing maximum likelihood estimation of a full order model and then using this model as if it is the true system for computing what the limit estimate η^* of $\eta(Z^N)$ would be, and then using this estimate in the application.

The invariance principle can thus be seen as a separation principle where the estimation problem is separated from the application dependent part of the problem.