CERTIFIED SYSTEM IDENTIFICATION
towards distribution-free results

Marco C. Campi [*] **Balázs Cs. Csáji** [**] **Simone Garatti** [***]
Erik Weyer [****]

[*] *Department of Information Engineering, University of Brescia, Via Branze 38, 25123 Brescia, Italy*
E-mail: marco.campi@ing.unibs.it
Web-site: http://www.ing.unibs.it/campi/
[**] *Department of Electrical and Electronic Engineering, The University of Melbourne, Parkville, VIC 3010, Australia, and Computer and Automation Research Institute, Hungarian Academy of Sciences, Kende utca 13-17, H-1111, Budapest, Hungary*
E-mail: bcsaji@unimelb.edu.au
Web-site: http://www.sztaki.hu/∼csaji/
[***] *Dipartimento di Elettronica e Informazione, Politecnico di Milano, piazza Leonardo da Vinci 32, 20133 Milan, Italy*
E-mail: sgaratti@elet.polimi.it
Web-site: http://home.dei.polimi.it/sgaratti/
[****] *Department of Electrical and Electronic Engineering, The University of Melbourne, Parkville, VIC 3010, Australia*
E-mail: ewey@unimelb.edu.au
Web-site: http://people.eng.unimelb.edu.au/ewey/

Abstract: System identification is the science of constructing models from data. A model is never an exact description of reality, and it is desirable that the identified model comes accompanied by certificates of quality able to describe the level of precision of the model and its domain of validity. This paper is about certified system identification. Our contention is that data contains more information than traditional identification methods can exploit, and, by looking at classical identification problems with new eyes, methods can be developed carrying precise quality guarantees that are valid under general assumptions. Taking the challenge of developing these methods may lead to a paradigm shift in many contexts in which identification is applied.

Keywords: quality certificate, distribution-free results, parameter estimation, prediction, filtering,

1. INTRODUCTION

1.1 Model quality certification

System identification is the science of constructing models from data.

A model is never an exact description of reality, and it is desirable that the identified model comes accompanied by *certificates of quality* able to describe the level of precision of the model and its domain of validity. A certificate of quality can e.g. be a statement of the type

$$\|\theta^o - \hat{\theta}\| \leq 0.1,$$

where θ^o is the parameter being estimated and $\hat{\theta}$ is its estimated value, or of the type

$$y_{t+1} \in \text{region } \hat{Y} \quad \text{with probability } 99\%,$$

when a region $\hat{Y}$ is estimated from data for the purpose of predicting the next value of a signal y. Certificates of quality are relevant to the practice of system identification, and are necessary for its scientific use.

System identification relies on data, and data is the real wealth in a system identification procedure. Data is always a limited resource, that is the data set has always a finite size N. Correspondingly, system identification methods should be able to squeeze out all the relevant information contained in the data, for the purpose of constructing models and of certifying their quality,

$$\text{DATA SET of SIZE N} \Rightarrow \begin{cases} \text{MODEL} \\ \text{CERTIFICATE of QUALITY.} \end{cases}$$

1.2 A common thread of this paper: from deterministic, via probabilistic, to distribution-free certificates

Certificates of quality can be provided in various forms. *Deterministic certificates* do not make any use of probability and assert facts, properties or results that are always valid, given the assumed premises. In contrast, *probabilistic certificates* affirm results that hold true with a given, possibly high, probability, but not always. Typically, a deterministic certificate is established under quite stringent assumptions, so that its applicability requires a deep prior knowledge of the environment in which the identification method is used. A probabilistic certificate does not require as stringent assumptions. However, probabilistic priors are normally used to infer the relevant properties of the estimate.

Assumptions limit the applicability of a method in two respects. First, the method is not applicable if the assumptions are not satisfied. Second, even if the assumptions are satisfied, the user may not know that they are. When knowledge is not sufficient to discriminate whether or not a method is applicable, one may be tempted to use the method anyhow if the assumptions do not seem implausible, but of course this is an awkward approach if one looks for certified guarantees.

Besides deterministic and probabilistic certificates, a third typology of certificates exists, which we here call *distribution-free certificates*. A distribution-free certificate is still a certificate of probabilistic nature, so that the existence of a probability is assumed in the mathematical formulation of the problem. Yet, probabilistic guarantees are obtained without knowledge of the actual values of the probability that describes the mechanism through which data are generated. Thus, the existence of a probability is assumed, but the probability values are not used in the method to derive conclusions.

Our contention in this paper is that identification methods carrying distribution-free certificates can be developed in various domains in which system identification is applied. These certificates are valid under little knowledge of the data generation mechanism and are practically useful. Working out distribution-free certificates, however, requires in many cases a *paradigm shift* and we have to look at traditional identification problems with new eyes. The potential reward is worth the effort, we believe, and this paper contains some preliminary ideas and results, which we hope will stimulate the interest of others and will foster the development of new research directions within the system identification community.

1.3 Paper structure

We shall undertake a journey through some of the core problems in system identification, namely,

- `parameter estimation;`
- `prediction;`
- `filtering.`

These problems will be dealt with in turn in Sections 2, 4, and 5. Ideas will be presented mainly through simple examples. For one thing, we believe simple examples provide a privileged entry point to get acquainted with theories; for another thing, the material presented in this paper is in many cases at the cutting edge of research, and a systematic theory providing a full coverage of the topic is not yet available.

2. PARAMETER ESTIMATION

2.1 A simple example

Consider the problem of estimating a parameter θ^0 from noisy measurements

$$y_t = \theta^o + n_t. \tag{1}$$

10 measurements are provided as follows

t	1	2	3	4	5	6	7	8	9	10
y_t	0.56	−0.66	1.12	1.32	−0.14	2.25	−0.21	0.96	1.28	1.17

what can we say about θ^0?

The *deterministic man* would claim he knows that noise is bounded, say $|n_t| \leq 2$. Since the first measurement is 0.56 and noise is in the interval $[-2, 2]$, θ^0 has to lie in $[-1.44, 2.56]$, see Figure 1. Turning to consider the

Figure 1. The interval compatible with the 1st measurement.

other measurements, one eventually constructs a set Θ by intersecting 10 line segments as depicted in Figure 2, and the claim associated with this construction is that certainly $\theta^0 \in \Theta$. This is a *deterministic certificate*, and

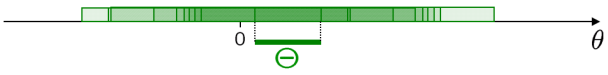

Figure 2. Θ is the interval compatible with all measurements.

the underlying logic of construction is pure intersection of the domains in θ that are compatible with the seen data, given the prior knowledge.

The deterministic man makes his construction on rigid deterministic priors.

Prior knowledge is based on experience that has been accrued in the past, and the *probabilistic man* believes that experience can never set a final word. He thus starts doubting the rigid prior that $|n_t| \leq 2$, and suggests that n_t can possibly take values larger than 2, even though rarely, which leads in his mind to the concept of probabilistic tail. One possible formalization he may suggest is that $n_t \sim \text{Gaussian}(0, 1)$. A unitary standard deviation makes unlikely that $|n_t| > 2$, and Gaussianity of the distribution is because he believes the world has a tendency to be Gaussian. Moreover, he adds the assumption that the noise affecting the various measurements are independent of each other, an assumption that he justifies by e.g. noting that different measurements have been collected with different sensors.

The probabilistic man computes the least squares estimate

$$\hat{\theta}_{LS} = \frac{1}{10}\sum_{t=1}^{10} y_t = 0.76,$$

and observes that

$$\theta^o - \hat{\theta}_{LS} \sim \text{ Gaussian}\left(0, \frac{1}{10}\right).$$

He then suggests constructing a 90% confidence region Θ for a Gaussian distribution with variance $\frac{1}{10}$ centered in $\hat{\theta}_{LS}$, Figure 3. This Θ contains θ^0 with probability 90%.

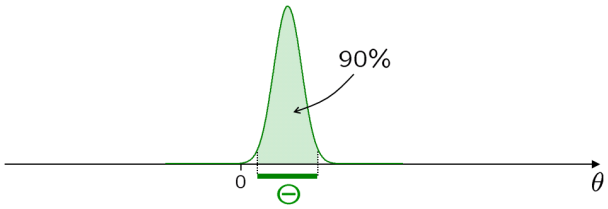

Figure 3. The 90% confidence region constructed by the "probabilistic man".

The deterministic man has introduced rigid hard priors. They have been relaxed by the probabilistic man, who, however, has assumed he knows the whole distribution of noise. This means assigning infinite probability values, one for each segment of the real line.

Let us try a *paradigm shift*. Noise still forms an independent sequence, but now it is assumed to have an *unknown* density with zero mean and a symmetric distribution around zero. Having some prior knowledge about the mean of the noise is necessary to be able to estimate θ^0 since the mean is added to θ^0 in the measurements generation mechanism (1). The density being symmetric is instead a restrictive assumption we make. Thus, noise can e.g. be Gaussian with any variance, or uniform, or triangular, see Figure 4, and the actual density is not known when identification is performed. Moreover, the density of noise

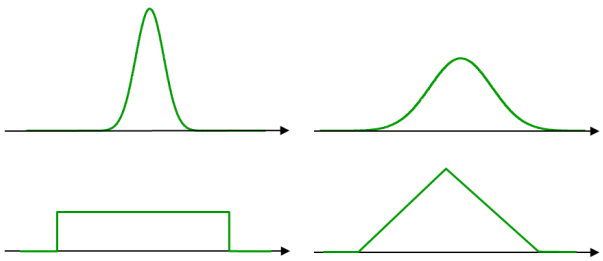

Figure 4. Possible noise densities

is allowed to change through time, so accommodating nonstationary situations, see Figure 5.

To determine a set Θ for θ^0, the following construction is used. First a test parabola is constructed as follows:

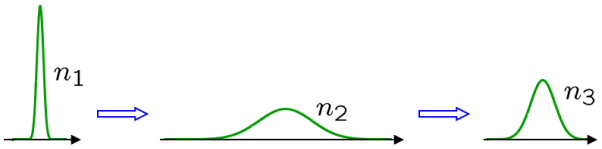

Figure 5. The noise density can vary though time.

$$\text{TEST PARABOLA} = \left[\sum_{t=1}^{10}(y_t - \theta)\right]^2.$$

This test parabola has vertex in $\hat{\theta}_{LS}$. For the data at hand, the test parabola is shown in Figure 6. Next, 9 other

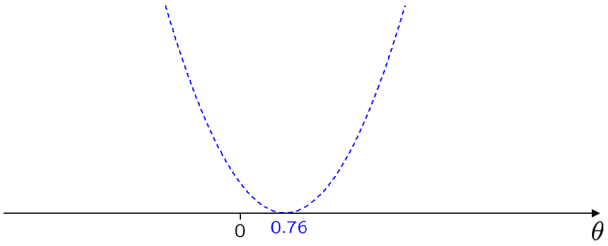

Figure 6. The test parabola.

parabolas are constructed as follows

$$\text{n}^{th} \text{ PARABOLA} = \left[\sum_{t=1}^{10} \pm(y_t - \theta)\right]^2, \quad n = 1, 2, \ldots, 9,$$

where $\pm$ are random signs obtained, for each parabola, by flipping a coin as many times as there are data, moreover, independent coin flippings are used for different parabolas. We did this construction for the data at hand, and obtained the parabolas displayed in Figure 7. The

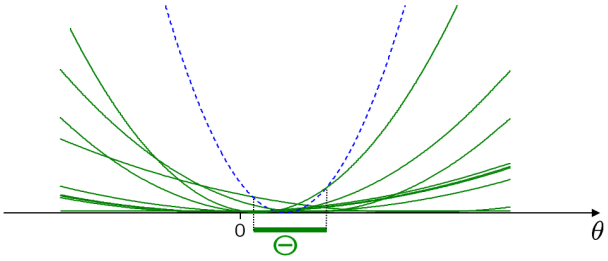

Figure 7. Θ is a (almost) distribution-free 90% confidence region.

interval Θ is where the test parabola is not at top. Figures 2, 3, and 7 are in scale, showing Θ intervals of similar size. The following result holds for this construction.

Theorem 1. $\theta^0 \in \Theta$ with probability 90%, irrespective of the noise density, provided that the noise density is symmetric around zero. ∗

A sketch of the proof is given in Appendix A.

This theorem cannot be claimed to be a true distribution-free result as noise has to be symmetrically distributed

around zero. Yet, it shows that a construction is possible leading to a set Θ that is guaranteed to contain θ^0 with probability 90% with limited knowledge of the noise characteristics. Moreover, 90% is not an upper bound, it is an exact probability, and the construction is therefore not conservative.

Additional properties hold for this construction. Here, we have used $N = 10$ data points. As the number of data points is let increase, $N \to \infty$, one can show that, under mild assumptions, the set Θ shrinks around θ^0 at a rate $O(1/\sqrt{N})$, which is the same as the rate obtained when the noise distribution is known.

This example shows that

> `there is information in the data that traditional methods do not exploit,`

and

> `methods can be conceived that let data speak beyond what traditional methods do.`

2.2 Generalizations

In the previous section, we have presented a simple example of a general theory. The corresponding identification method, which we have called *Sign Perturbed Sums* (SPS), can be applied to generic dynamical systems to construct guaranteed confidence regions for the system parameters. If, for example, the system has the following Finite Impulse Response (FIR) structure

$$y_t = \theta_1^o u_{t-1} + \theta_2^o u_{t-2} + n_t,$$

the same construction as in the static example of the previous section can be applied by simply substituting the parabolas with the paraboloids

$$\text{TEST PARABOLOID} = \left\| \sum_{t=1}^{N} \begin{bmatrix} u_{t-1} \\ u_{t-2} \end{bmatrix} (y_t - \theta_1 u_{t-1} - \theta_2 u_{t-2}) \right\|^2 ;$$

$$\text{n}^{th}\ \text{PARABOLOID} = \left\| \sum_{t=1}^{N} \pm \begin{bmatrix} u_{t-1} \\ u_{t-2} \end{bmatrix} (y_t - \theta_1 u_{t-1} - \theta_2 u_{t-2}) \right\|^2 ,$$

and a result that replicates, *mutatis mutandis*, Theorem 1 holds true. Weighting matrices can also be introduced in the paraboloids to optimize the shape of the region Θ. Moreover, the theory can be carried over to more general dynamical systems, such as AutoRegressive Moving Average (ARMA) systems

$$A^o(z^{-1})y_t = B^o(z^{-1})u_t + n_t,$$

or Box-Jenkins systems

$$y_t = G^o(z^{-1})u_t + H^o(z^{-1})n_t.$$

The reader is referred to Csáji et al. (2012a) and to the article in preparation Csáji et al. (2012b) for more details.

3. MID-PAPER CONCLUSIONS

Let us pause a moment, and analyze what we have seen so far.

In the problem of estimating a parameter θ^0 from noisy measurements, the deterministic man constructs a set Θ and certifies it with the claim that $\theta^0 \in \Theta$. This deterministic certificate is a *set-theoretic* result, no matter what the noise sequence n_t is in its domain of variability characterized as the set where $|n_t| \leq 2$, the result holds true.

Condition $|n_t| \leq 2$ is stiff, and the theory is deeply relying on this condition. Quoting from Mark Twain:

> *"what gets us into trouble is not what we don't know. It's what we know for sure that just ain't so."*

Removing the stringent deterministic condition, the probabilistic man does not aprioristically clip the noise to a given value, and he also allows for noise sequences that take on large values. As a result, it may happen that $\theta^0 \notin \Theta$. Graphically, the situation becomes as illustrated in Figure 8. The set of noise sequences is partly depicted in white,

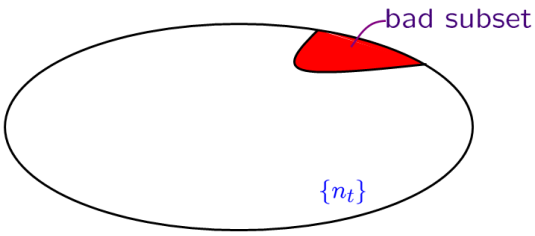

Figure 8. The "bad set".

signifying that $\theta^0 \in \Theta$ if n_t belongs to this white subset, and partly in red, signifying the "bad subset" of noise sequences where $\theta^0 \notin \Theta$. Since the result that $\theta^0 \in \Theta$ is no longer set-theoretic and it only holds in a subset of the set of noise sequences, the probabilistic man is facing a new problem, that of *being quantitative*, he has to measure the extension of the bad set. The tool used for measuring sets is measure theory, and when the measure is interpreted as chance of occurrence, the measure is called probability. Thus, we see that

> `a probability is needed to tackle the challenge of being quantitative.`

The claim he makes is that $\theta^0 \in \Theta$ with probability 90%.

The probabilistic man assumes he has quite a bit of knowledge about the data generation mechanism, to the point that he is able to attribute a probability value to each segment of the real line where noise takes value. Jan Willems, one of the deepest thinkers of the systems theory community, once noticed

> *"where would the numerical values of the probability come from?"*

This question is central in the probabilistic formulations of identification theories.

Moving towards a distribution-free approach, the attempt is to create a new paradigm where the probability that

Θ contains θ^0 is made as independent as possible of the probabilistic characteristics of noise. Prior knowledge is reduced to a minimum, and the identification algorithm is required to squeeze out the information contained in the data, *without a-priori assuming what the data have to tell us*. In this context, a probability is assumed to exist, but the probability values cease to play a role in the algorithm, which only uses data.

4. PREDICTION

4.1 Interval prediction models: an example

We are presented with the $N = 19$ data points shown in Figure 9 taken at random from a population of points in $\mathbb{R}^2$. Mathematically, the 19 data points are samples of

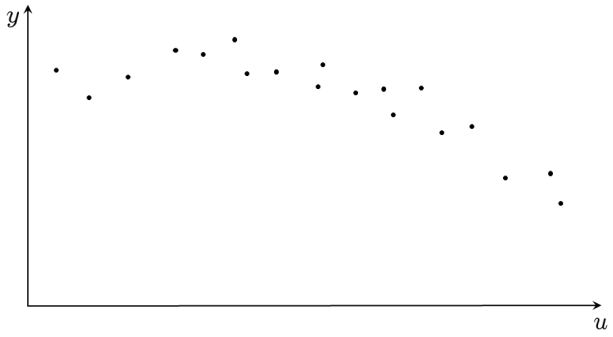

Figure 9. The data points.

variables independent and identically distributed (i.i.d.) according to an unknown probability distribution on $\mathbb{R}^2$. From an application perspective, instead, each point represents a member of a population described by two attributes, u and y, whose interpretation varies with the application and can e.g. be u = "height" and y = "weight", u = "investment" and y = "return", u = "medical test result" and y = "level of a disease". We want to see u as a variable we use to predict y, and a prediction model has to be constructed from the seen 19 observations.

Given the next value of u, say $\bar{u}$ in Figure 10, our prediction is given as an interval to which the next value $\bar{y}$ is expected to belong. How can we construct the prediction interval? and, what kind of guarantee can we attach to our construction?

Along a deterministic approach, one would argue that, without additional information and unless we uselessly take as prediction interval the whole real line of y, any construction can be invalidated by the next value of y, that is, our prediction is wrong. This is indeed true, as nothing prevents $\bar{y}$ from being outside the prediction interval, whatever prediction interval we exhibit. The way out of this problem he suggests is to introduce prior knowledge on the underlying data generation mechanism. Among many possibilities, let us assume for example that we know that the data are generated according to the mechanism

$$y = f(u) + n,$$

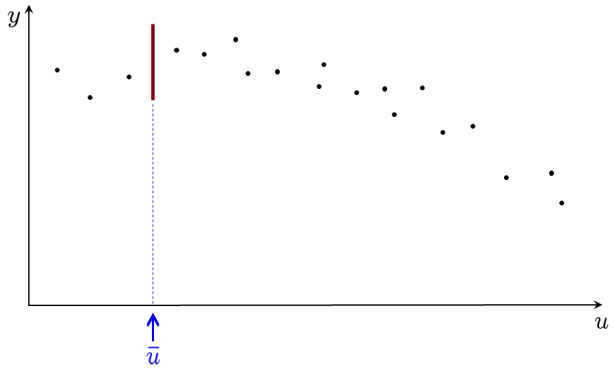

Figure 10. The prediction interval for $u = \bar{u}$.

where function f has limited slope, say $\left|\frac{\partial f}{\partial u}\right| \leq 1$, and n is bounded, say $|n| \leq 0.5$. Then, an interval prediction model is constructed as shown in Figure 11, and the predicted interval in correspondence of $\bar{u}$ is given by the intersection of the vertical line starting from $\bar{u}$ with the interval prediction model. This prediction is always

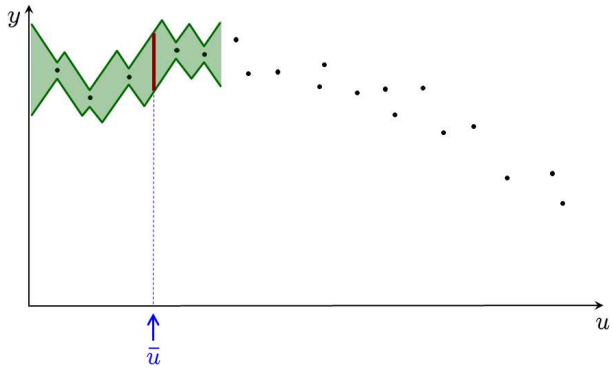

Figure 11. The interval prediction model constructed along a deterministic approach.

correct, provided the assumptions are true.

The deterministic assumptions are demanding, and, correspondingly, the conclusion that y is certainly in the interval is quite assertive. We next want to investigate whether the deterministic assumptions can be relaxed by the adoption of a probabilistic approach.

Can we make any sensible probabilistic claim with no restrictive assumptions on the underlying probabilistic set-up according to which data are generated? To rapidly gain insight on this question, let us directly move to the extreme point that the data are sampled from a population whose probability distribution is *completely unknown* to us. That is, let us take a *distribution-free* approach. The following optimization program determines the thinner layer centered around a tunable parabola and that contains all the seen data

$$\begin{aligned} &\min_{r,\alpha,\beta,\gamma} \quad r \\ &\text{subject to:} \quad \left|y_i - [\alpha + \beta u_i + \gamma u_i^2]\right| \leq r, \quad i = 1, \ldots, N. \end{aligned}$$

For the data at hand, the layer is shown in Figure 12, this layer is our interval prediction model. Tomorrow, a new

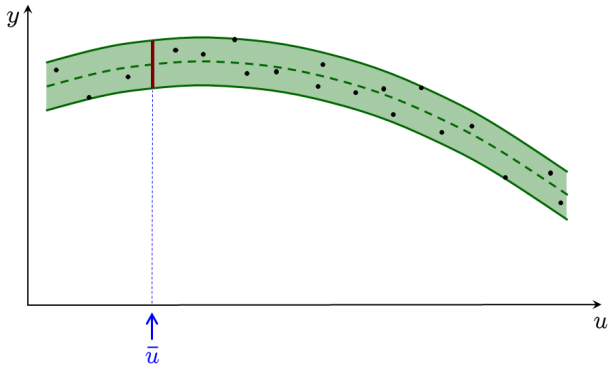

Figure 12. The probabilistic layer.

member $(\bar{u}, \bar{y})$ is sampled from the population according to the same probability distribution as the seen data. We are given $\bar{u}$, and we are asked to predict $\bar{y}$. As before, the prediction interval is obtained by intersecting the vertical line starting from $\bar{u}$ with the interval prediction model. What can we claim about this construction? The following theorem provides a crisp answer.

Theorem 2. Suppose that the probability distribution on $\mathbb{R}^2$ admits density. Then, the prediction is correct, that is, $\bar{y}$ is in the interval, with probability 80%, whatever the probability density on $\mathbb{R}^2$ according to which data are generated is. ∗

This theorem follows from the following more general result whose proof is sketched in Appendix B.

Theorem 3. Let the centerline of the layer be linearly parameterized in terms of k parameters, and suppose that N data points are sampled in an i.i.d. fashion according to a probability distribution on $\mathbb{R}^2$ that admits density. Then, the thinner layer that contains the data points predicts correctly with exact probability p if

$$N = \frac{k+p}{1-p},$$

whatever the probability density on $\mathbb{R}^2$ according to which data are generated is. ∗

In our example, $N = 19$ equals $\frac{k+p}{1-p} = \frac{3+0.8}{1-0.8}$.

4.2 Discussion

Theorem 3 is a distribution-free result applicable to all distributions that admit a density, and shows that the probability of an incorrect prediction is independent of the probability density of the data. In other words, Theorem 3 provides a *universal* result.

To appreciate more concretely the significance of Theorem 3, suppose we are regressing the height of the individuals belonging to a certain population against their weight. Given the weight and the height of N individuals, the layer constructed on these N observations is guaranteed to correctly predict the height of the next individual given his/her weight with a probability p. This probability is *exact*, not a lower bound, so that no conservatism is present in this probabilistic evaluation, and no a-priori knowledge is required for this result to hold. Thus, knowing for instance that the population is solely formed by female, or solely by male, or that children are, or are not, present does not help to improve the result.

One can wonder about the reason that makes this result possible. The reason is that what is used for predicting, a layer centered e.g. around a parabola or around a cubic polynomial, is a simple object but, despite its simplicity, it can reliably predict (i.e. it predicts correctly with high probability) even complex data generation mechanisms. An example is given in Figure 13, where the layer is centered around a cubic polynomial. In Figure 13 (a),

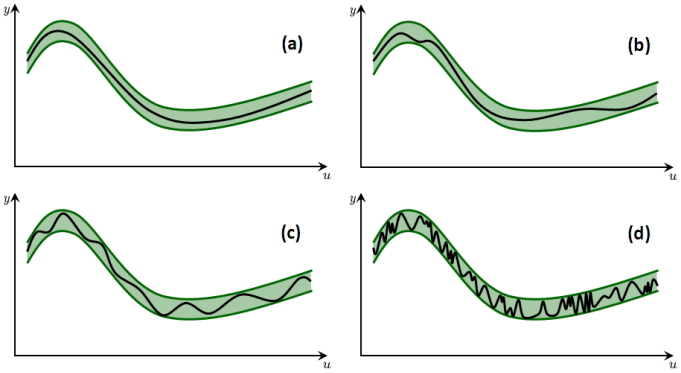

Figure 13. A single layer is reliable for a wide class of functions with increasing level of complexity.

the layer is always correct in predicting the output of the function in black, as it can be seen by the fact that the function is entirely contained in the layer. The same layer, however, can predict with no error the output of the functions in Figures 13 (b)-(d) that show an increasing level of complexity. The fact that a simple layer can reliably describe complex data generation mechanisms is represented in Figure 14 by the arrow connecting "low complexity" with "reliability". The other arrow in this

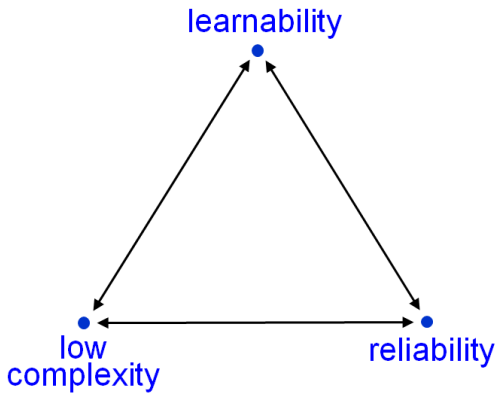

Figure 14. Links between "low complexity", "reliability", and "learnability".

figure connecting "low complexity" with "learnability" simply means that learning from a finite data set is possible for model classes having low complexity. Putting together

these two arrows we obtain a link between "reliability" and "learnability", which is represented by the third arrow in the figure. This third arrow is interpreted that, in this scheme, reliable models can be learned even for complex data generation mechanisms.

As we have seen, prior knowledge does not count in the reliability result announced in Theorem 3. So, does this theory claim that *prior knowledge* is of no use? Certainly, this is not so. Besides reliability, a layer has a second, equally important, attribute, its *thickness*. A thick layer provides wide prediction intervals, which are of little use in practice. Theorem 3 only provides answers about reliability, while it says nothing about thickness. In the theorem, the centerline of the layer is required to be linearly parameterized in terms of k parameters, while the regression functions are left free, and the user should spend his prior knowledge to determine suitable regression functions so that the resulting layer is as thin as possible.

Further developing on the concepts of reliability and thickness, notice that reliability is not a property of the model since a model can be reliable for a data generation mechanism and not reliable for another data generation mechanism. Thus,

```
reliable is a property with two arguments:
the model and the data generation mechanism.
```

Instead,

```
thickness is a property with one argument
only, the model.
```

As a consequence, thickness can be inspected once the model has been constructed, reliability can not. For this very reason it is well acceptable that priors impact on the thickness, which we can assess before we "buy" the model, whereas it is important that reliability is guaranteed by a theory that holds under the most general possible assumptions, what we call a distribution-free theory.

4.3 Generalizations

The approach illustrated in the previous section can be generalized in many directions, and the reader is referred to Campi et al. (2009a) for an overview presentation. We here limit ourselves to say that interval prediction models can be constructed that have more general structure than that of a layer, allowing e.g. for intervals whose width is varying with the input, as shown in Figure 15. Moreover, prediction with more inputs $u_1, \ldots, u_p$ can be accommodated within this framework. Also, observations that are showing little conformity to the other observations (outliers) can be excluded from the interval prediction model as illustrated in Figure 16, and the theory assuring the reliability of the model can be extended to cover this situation, see Campi and Garatti (2011).

Another notable generalization is to classification problems. Supposing e.g. that the output is binary, 0 or 1, the equivalent of an interval prediction model in this context is a model that can output 0, 1, or $\{0, 1\}$, where $\{0, 1\}$ corresponds to abstention from a classification. The reader is referred to Campi (2010) for a presentation of how classification problems can be dealt with along the approach

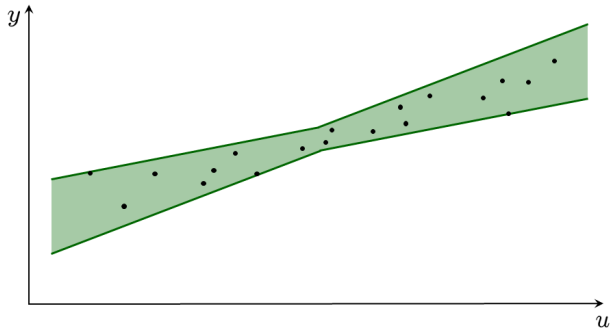

Figure 15. An interval prediction model whose width is varying with the input.

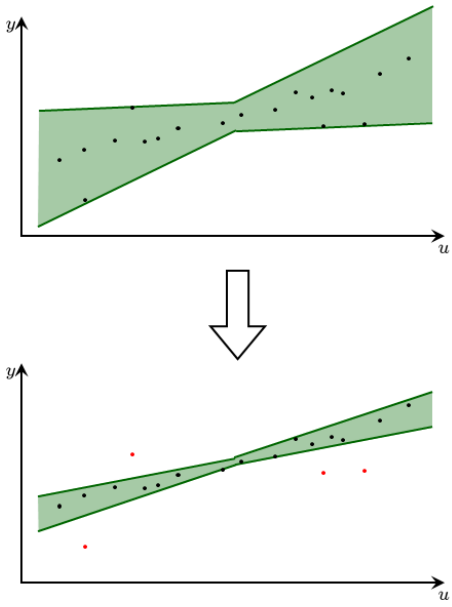

Figure 16. Outliers removal.

of this section, which leads to classifiers having a known and exact probability of correct classification.

The results outlined above hold for independent data. Extending these results to correlated data generated by a dynamical system presents interesting theoretical challenges. Despite some preliminary results are given in Campi et al. (2009a), the problem of working out distribution-free non-conservative results in the correlated context is currently an open issue.

5. FILTERING

5.1 An example of filtering for a system with scalar state

In this section, we sketch some ideas for constructing confidence regions for the state of a system without knowing the variances of the system noises.

Consider the following first order system

$$\begin{aligned} x_{t+1} &= f x_t + v_t \\ y_t &= x_t + w_t, \end{aligned} \tag{2}$$

where x_t, y_t, v_t, w_t are scalars, f is known, and v_t and w_t are mutually independent sequences of independent and identically distributed (i.i.d.) Gaussian random variables. Assuming that the noise variances σ_v^2 and σ_w^2 are known, and so is an initial Gaussian distribution of the system state, the problem of estimating x_{t+1} from output measurements is optimally solved by the Kalman filter. Moreover, based on the state estimate and the estimation error variance, which is obtained by solving a Riccati equation, a confidence ellipsoid can be constructed carrying a predefined probability to contain the true state.

Suppose instead that σ_v^2 and σ_w^2 are not known. Can we still construct a region that contains x_{t+1} with an *exact probability*? In addressing this question, we make an attempt to move a step towards a novel filtering approach that carries precise probabilistic guarantees under more general assumptions than Kalman filtering.

Let us introduce the assumption that $|f| < 1$, and that system (2) operates in steady-state, so that x_t is a stationary process. x_t can also be represented by a state equation running backwards in time, see e.g. Lemma 5.4.4 in Kailath et al. (2000),

$$x_t = f x_{t+1} + v'_t, \tag{3}$$

where v'_t is a new sequence of i.i.d. Gaussian random variables with variance σ_v^2. Using (3) and (2), the outputs observed between time 1 and t can be expressed in terms of x_{t+1} as follows

$$\begin{bmatrix} y_t \\ y_{t-1} \\ y_{t-2} \\ \vdots \\ y_1 \end{bmatrix} = \begin{bmatrix} f \\ f^2 \\ f^3 \\ \vdots \\ f^t \end{bmatrix} x_{t+1} + \\ + \begin{bmatrix} 1 & 0 & 0 & \cdots \\ f & 1 & 0 & \cdots \\ f^2 & f & 1 & \cdots \\ \vdots & \vdots & \vdots & \ddots \\ f^{t-1} & f^{t-2} & f^{t-3} & 1 \end{bmatrix} \begin{bmatrix} v'_t \\ v'_{t-1} \\ v'_{t-2} \\ \vdots \\ v'_1 \end{bmatrix} + \begin{bmatrix} w_t \\ w_{t-1} \\ w_{t-2} \\ \vdots \\ w_1 \end{bmatrix}. \tag{4}$$

Next we transform (4) such that the transformed noise forms an independent sequence. Let

$$B = \begin{bmatrix} 1 & 0 & 0 & \cdots \\ f & 1 & 0 & \cdots \\ f^2 & f & 1 & \cdots \\ \vdots & \vdots & \vdots & \ddots \\ f^{t-1} & f^{t-2} & f^{t-3} & 1 \end{bmatrix},$$

$$C = B^{-1} = \begin{bmatrix} 1 & 0 & 0 & \cdots \\ -f & 1 & 0 & \cdots \\ 0 & -f & 1 & \cdots \\ \vdots & \vdots & \ddots & \ddots \\ & & & -f \ 1 \end{bmatrix},$$

and let the singular value decomposition of CC^T be given by

$$VDV^T = CC^T,$$

where V is a $t \times t$ matrix with the property that $V^TV = I$ (the identity matrix), and D is a $t \times t$ diagonal matrix. It follows that

$$V^TCC^TV = D.$$

By premultiplying (4) by V^TC we obtain the equation

$$\begin{bmatrix} s_t \\ s_{t-1} \\ s_{t-2} \\ \vdots \\ s_1 \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_t \end{bmatrix} x_{t+1} + \begin{bmatrix} n_t \\ n_{t-1} \\ n_{t-2} \\ \vdots \\ n_1 \end{bmatrix}, \tag{5}$$

where

$$\begin{bmatrix} s_t \\ s_{t-1} \\ s_{t-2} \\ \vdots \\ s_1 \end{bmatrix} = V^TC \begin{bmatrix} y_t \\ y_{t-1} \\ y_{t-2} \\ \vdots \\ y_1 \end{bmatrix}, \tag{6}$$

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_t \end{bmatrix} = V^TC \begin{bmatrix} f \\ f^2 \\ f^3 \\ \vdots \\ f^t \end{bmatrix},$$

$$\begin{bmatrix} n_t \\ n_{t-1} \\ n_{t-2} \\ \vdots \\ n_1 \end{bmatrix} = V^T \begin{bmatrix} v'_t \\ v'_{t-1} \\ v'_{t-2} \\ \vdots \\ v'_1 \end{bmatrix} + V^TC \begin{bmatrix} w_t \\ w_{t-1} \\ w_{t-2} \\ \vdots \\ w_1 \end{bmatrix}.$$

$n_1, \ldots, n_t$ are zero mean Gaussian random variables. It turns out that these variables are also independent. The proof of this fact is straightforward: the covariance matrix of $[n_t, \ldots, n_1]^T$ is given by the diagonal matrix $\sigma_v^2 I + \sigma_w^2 D$, showing that $[n_t, \ldots, n_1]^T$ has uncorrelated components. Independence of $n_1, \ldots, n_t$ follows from the fact that $[n_t, \ldots, n_1]^T$ is a Gaussian vector. Notice, however, that $n_1, \ldots, n_t$ are not identically distributed.

From the backward representation (3) it is seen that x_{t+1}, which is constructed from v'_i, $i \geq t+1$, is independent of $v'_1, \ldots, v'_t$. Since x_{t+1} is also independent of $w_1, \ldots, w_t$, it follows that x_{t+1} is independent of $n_1, \ldots, n_t$. As a result, x_{t+1} in equation (5) can be regarded as though it were a deterministic quantity, which, in mathematical terms, is achieved by conditioning (5) with respect to x_{t+1}. Thus, the problem of estimating x_{t+1} in (5) is an identification problem where the parameter to be estimated is x_{t+1}, $s_1, \ldots, s_t$ are the observations whose value is known from equation (6) (notice that V and C only depend on the known system state parameter f), and $a_1, \ldots, a_t$ are also known quantities. To this problem, the same technique as that described in Section 2 can be applied in order to construct an interval that contains x_{t+1} with guaranteed probability. In this context, the test parabola is

$$\left[\sum_{i=1}^{t} a_i(s_i - a_i x)\right]^2,$$

and the other parabolas are constructed by adding $\pm$ random signs. Adding a_i in front of $(s_i - a_i x)$ serves the purpose of making all signs of x in the test parabola negative, so that the test parabola has a curvature narrower than the other parabolas and the interval for x_{t+1} is bounded. The fact that the noise $n_1, \ldots, n_t$ has varying intensity does not represent a difficulty, as it was already pointed out in Section 2.

There is a conceptual difference in the results that are achieved along this approach as compared to Kalman filtering that is worth noting. Assuming that σ_v^2 and σ_w^2 are known, with the Kalman filter a confidence ellipsoid can be constructed that contains the system state with a given probability. The result that is obtained with the approach of this section is of different nature. The result is that, *regardless of what the true state is*, the constructed confidence set will contain the true state with a given probability. Thus, even states x_{t+1} that are only rarely touched in the system evolution, if touched, they will fall in the confidence set with the given probability. This type of result is particularly relevant for monitoring applications, where we are much more concerned about obtaining reliable confidence sets when the state takes on particular values, e.g. when the state is in or close to a dangerous operating region.

In this section, we have briefly outlined some ideas of a broad novel approach to filtering that is under construction. The reader is referred to Weyer and Campi (2011b) for further comments and simulation results on the scalar state estimation problem. The general case of multidimensional state is currently under consideration, and some preliminary results can be found in Weyer and Campi (2011a). Many problems remain open at the present stage of knowledge regarding the weighting of the observations, as well as the use of the backward representation in the multidimensional case.

6. BIBLIOGRAPHICAL NOTES

Deterministic identification has been pursued along many lines of research e.g. in Milanese and Vicino (1991), Bai et al. (1995, 1996), Vicino and Zappa (1996), Giarré et al. (1997), Garulli et al. (2000, 2002). The developed methods also cover the presence of unmodelled dynamics, and the treatment of noise without requiring an explicit description of the noise model, provided that upper bounds are available on all the unknown quantities.

The literature on stochastic system identification is truly vast. The reader is referred to the books Söderström and Stoica (1989) and Ljung (1999) for a general presentation, while the following is just a selection among many papers on the subject, Ljung (1978, 1985), Hjalmarsson and Ljung (1992), Goodwin et al. (1992), Ninness and Goodwin (1995), Hakvoort and Van den Hof (1997), Pintelon et al. (1997), Ninness et al. (1999), Gevers et al. (2001), Ninness and Hjalmarsson (2004), Garatti et al. (2004, 2006), Hjalmarsson and Mårtensson (2011). The interest in establishing finite sample results valid under general assumptions was recognized as early as in Gosset (Student) (1908). On the other hand, in stochastic system identification almost all existing results leverage on asymptotic results from statistics and are therefore applicable to data sets of diverging size. Early works of the authors of this paper on distribution-free finite sample results in system identification are Campi and Weyer (2005), Dalai et al. (2007), Campi et al. (2009b), Campi and Weyer (2010).

Deterministic prediction is the subject of e.g. Milanese and Novara (2004, 2005).

Filtering is the subject of the textbooks Jazwinski (1970), Anderson and Moore (1976), Kailath et al. (2000), while the approach discussed in Section 5 has been introduced in Weyer and Campi (2011a,b).

Appendix A. PROOF OF THEOREM 1

We first assume that, corresponding to $\theta = \theta^0$, no tie occurs, that is, no two parabolas have the same value.

For $\theta = \theta^0$, the test parabola writes

$$\left[\sum_{t=1}^{10}(y_t - \theta^o)\right]^2 = \left[\sum_{t=1}^{10}(\theta^o + n_t - \theta^o)\right]^2 = \left[\sum_{t=1}^{10} n_t\right]^2,$$

while the nth parabola is

$$\left[\sum_{t=1}^{10} \pm n_t\right]^2.$$

Since n_t is an independent sequence with symmetric distribution, $\left[\sum_{t=1}^{10} n_t\right]^2$ and $\left[\sum_{t=1}^{10} \pm n_t\right]^2$ have the same probability distribution, and each of the 10 parabolas has the same chance to be at top as any other one. As a result, setting Θ to be the region where the test parabola is not at top leaves us with a probability 90% that $\theta^0 \in \Theta$.

We have assumed that, corresponding to $\theta = \theta^0$, no tie occurs. A tie occurs when two parabolas have the same or opposite random sign $\pm$ sequence, which can be avoided by dropping a $\pm$ sequence when it occurs to be identical to or opposite of a previously constructed sequence. Even for different and not opposite $\pm$ sequences, a tie occurs if summing noise terms with different signs leads to the same value. Such circumstance has however probability zero if the noise admits density, and it does not affect the result.

This issue of dealing with ties can be treated in other ways, beyond what has been described here. However, this issue only plays a marginal conceptual and applicative role, and we do not dwell on further discussing it here.

Appendix B. PROOF OF THEOREM 3

Consider a set of $N+1$ points extracted one independently of the others according to the probability density on $\mathbb{R}^2$, and construct the thinner layer that contains these $N+1$ points. It can be shown, e.g. Garatti and Campi (2009), that the number of points that touch the boundary of the layer is equal to $k+1$ with probability 1. For instance, for a

parabolic centerline as in Figure 12, the number of points that touch the boundary is 4. If one of the $k+1$ points that touch the boundary is eliminated and the thinner layer that contains the other points is constructed, this constructed layer is thinner than the layer containing all the points, and the eliminated point falls outside the layer. If instead one of the points that do not touch the boundary is eliminated and the thinner layer that contains the other points is constructed, the eliminated point falls in the layer.

We thus have

$$
\begin{aligned}
&\text{Probability of correct prediction}\\
&= [\text{let } p_i = (u_i, y_i), i = 1, \dots, N; p_{N+1} = (\bar{u}, \bar{y})]\\
&= \text{Prob}\{p_{N+1} \in \text{ layer containing } p_1, \dots, p_N\}\\
&= [\mathbb{1}(\cdot) = \text{ indicator function}]\\
&= \int \mathbb{1}(p_{N+1} \in \text{ layer containing } p_1, \dots, p_N) d\text{Prob}^{N+1}\\
&= [\text{since each point is equivalent to any other point}]\\
&= \frac{1}{N+1} \int \sum_{i=1}^{N} \mathbb{1}(p_i \in \text{ layer containing } p_1, \dots, p_{i-1},\\
&\quad p_{i+1}, \dots, p_{N+1}) d\text{Prob}^{N+1}\\
&= [\text{use the fact that } k+1 \text{ points, if eliminated, fall}\\
&\quad \text{outside the layer that contains the other points}]\\
&= \frac{1}{N+1} \int [(N+1) - (k+1)] d\text{Prob}^{N+1}\\
&= \frac{N-k}{N+1}. \qquad \text{(B.1)}
\end{aligned}
$$

Letting

$$\text{Probability of correct prediction} = p,$$

and making (B.1) explicit with respect to N, the result in Theorem 3 follows.

REFERENCES

B.D.O. Anderson and J.B. Moore. *Optimal Filtering*. Prentice Hall, 1976.

E.W. Bai, K.M. Nagpal, and R. Tempo. Membership set estimators: size, optimal inputs, complexity and relations with least squares. *IEEE Transactions on Circuits and Sytems*, 42:266–277, 1995.

E.W. Bai, K.M. Nagpal, and R. Tempo. Bounded-error parameter estimation: noise models and recursive algorithms. *Automatica*, 32:985–999, 1996.

M.C. Campi. Classification with guaranteed probability of error. *Machine Learning*, 80:63–84, 2010.

M.C. Campi and S. Garatti. A sampling-and-discarding approach to chance-constrained optimization: Feasibility and optimality. *Journal of Optimization Theory and Applications*, 148:257–280, 2011.

M.C. Campi and E. Weyer. Non-asymptotic confidence sets for the parameters of linear transfer functions. *IEEE Transactions on Automatic Control*, 55:2708–2720, 2010.

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

M.C. Campi, G. Calafiore, and S. Garatti. Interval predictor models: identification and reliability. *Automatica*, 45:382–392, 2009a.

M.C. Campi, S. Ko, and E. Weyer. Non-asymptotic confidence regions for model parameters in the presence of unmodelled dynamics. *Automatica*, 45:2175–2186, 2009b.

B.C. Csáji, M.C. Campi, and E. Weyer. Non-asymptotic confidence regions for the least-squares estimate. In *Proceedings of the 16th IFAC Symposium on System Identification*, Bruxelles, Belgium, 2012a.

B.C. Csáji, M.C. Campi, and E. Weyer. Sign-perturbed sums (SPS): A method for constructing exact finite-sample confidence regions for general linear systems. *Article in preparation*, 2012b.

M. Dalai, E. Weyer, and M.C. Campi. Parameter identification for nonlinear systems: guaranteed confidence regions through LSCR. *Automatica*, 43:1418–1425, 2007.

S. Garatti and M.C. Campi. L_∞ layers and the probability of false prediction. In *Proceedings of the 15th IFAC Symposium on System Identification*, Saint Malo, France, 2009.

S. Garatti, M.C. Campi, and S. Bittanti. Assessing the quality of identified models through the asymptotic theory - when is the result reliable? *Automatica*, 40: 1319–1332, 2004.

S. Garatti, M.C. Campi, and S. Bittanti. The asymptotic model quality assessment for instrumental variable identification revisited. *Systems & Control Letters*, 55:494–500, 2006.

A. Garulli, A. Vicino, and G. Zappa. Conditional central algorithms for worst-case set membership identification and filtering. *IEEE Transactions on Automatic Control*, 45:14–23, 2000.

A. Garulli, L. Giarré, and G. Zappa. Identification of approximated Hammerstein models in a worst-case setting. *IEEE Transactions on Automatic Control*, 47: 2046–2050, 2002.

M. Gevers, L. Ljung, and P. Van den Hof. Asymptotic variance expressions for closed-loop identification. *Automatica*, 37:781–786, 2001.

L. Giarré, B.Z. Kacewicz, and M. Milanese. Model quality evaluation in set membership identification. *Automatica*, 33:1133–1139, 1997.

G.C. Goodwin, M. Gevers, and B. Ninness. Quantifying the error in estimated transfer functions with application to model order selection. *IEEE Transactions on Automatic Control*, 37:913–928, 1992.

W.S. Gosset (Student). The probable error of a mean. *Biometrika*, 6:1–25, 1908.

R. Hakvoort and P.M.J. Van den Hof. Identification of probabilistic system uncertainty by explicit evaluation of bias and variance errors. *IEEE Transactions on Automatic Control*, 42:1516–1528, 1997.

H. Hjalmarsson and L. Ljung. Estimating model variance in the case of undermodeling. *IEEE Transactions on Automatic Control*, 37:1004–1008, 1992.

H. Hjalmarsson and J. Mårtensson. A geometric approach to variance analysis in system identification. *IEEE Transactions on Automatic Control*, 56:983–997, 2011.

A.H. Jazwinski. *Stochastic processes and filtering theory*. Academic Press, 1970.

T. Kailath, A.H. Sayed, and B. Hassibi. *Linear estimation*. Prentice Hall, 2000.

L. Ljung. *System Identification: Theory for the User*. Prentice-Hall, Upper Saddle River, 2nd edition, 1999.

L. Ljung. Convergence analysis of parametric identification methods. *IEEE Transactions on Automatic Control*, 23:779–783, 1978.

L. Ljung. Asymptotic variance expressions for identified black-box transfer function models. *IEEE Transactions on Automatic Control*, 30:834–844, 1985.

M. Milanese and C. Novara. Set-membership identification of nonlinear systems. *Automatica*, 40:957–975, 2004.

M. Milanese and C. Novara. Set-membership prediction of nonlinear time series. *IEEE Transactions on Automatic Control*, 50:1655–1669, 2005.

M. Milanese and A. Vicino. Optimal estimation theory for dynamic systems with set membership uncertainty: an overview. *Automatica*, 27:997–1009, 1991.

B. Ninness and G. Goodwin. Estimation of model quality. *Automatica*, 31:1771–1795, 1995.

B. Ninness and H. Hjalmarsson. Variance error quantifications that are exact for finite model order. *IEEE Transactions on Automatic Control*, 49:1275–1291, 2004.

B. Ninness, H. Hjalmarsson, and F. Gustafsson. The foundamental role of orthonormal bases in system identification. *IEEE Transactions on Automatic Control*, 44: 1384–1406, 1999.

R. Pintelon, J. Schoukens, and G. Vandersteen. Frequency domain system identification using arbitrary signals. *IEEE Transactions on Automatic Control*, 42: 1717–1720, 1997.

T. Söderström and P. Stoica. *System Identification*. Prentice Hall, 1989. ISBN 0-13-881236-5.

A. Vicino and G. Zappa. Sequential approximation of feasible parameter sets for identification with set membership uncertainty. *IEEE Transactions on Automatic Control*, 41:774–785, 1996.

E. Weyer and M.C. Campi. Prediction, filtering and smoothing using LSCR: State estimation algorithms with guaranteed confidence sets. In *Proceedings of the 50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC)*, Orlando, Florida, USA, 2011a.

E. Weyer and M.C. Campi. State estimation algorithms with guaranteed confidence intervals for first order systems. In *Proceedings of the 9th IEEE International Conference on Control and Automation (ICCA)*, Santiago, Chile, 2011b.