Estimation of Model Quality

Brett Ninness and Graham Goodwin

March 1, 1995

ABSTRACT

This paper gives an introduction to recent work on the problem of quantifying errors in the estimation of models for dynamic systems. This is a very large field. We therefore concentrate on approaches that have been motivated by the need for reliable models for control system design. This will involve a discussion of efforts which go under the titles of 'Estimation in H_∞', 'Worst Case Estimation', 'Estimation in ℓ_1', 'Information Based Complexity', and 'Stochastic Embedding of Undermodelling'. A central theme of this survey is to examine these new methods with reference to the classic bias/variance tradeoff in model structure selection.

Technical Report EE9437

Centre for Industrial Control Science and Department of Electrical and Computer Engineering, University of Newcastle, Callaghan 2308, AUSTRALIA

1 Introduction

Our aim in this paper is to survey an area of research which has flourished in recent years. The common denominator of this work is that of finding system identification methods that not only provide models, but also provide bounds on the accuracy (or quality) of those models.

A key idea is that the purpose of deriving models is usually to facilitate some sort of design procedure. In this case an estimate of the accuracy of the model is required if a reliable design is to be obtained. As a simple example, resistors used in electronic circuits are always specified with an associated accuracy. This accuracy is crucial in ensuring that a given design meets the intended specification. As an even simpler example, note that it is of little help to be told that the a-priori expected value of the winning ticket in a 1,000,000 ticket lottery is 500,000.

In the setting of this paper, namely the area of systems science, it is also important that any model be accompanied by some form of error quantification. For example, in control system design a difference between the real system and the model may not only mean that the design specification is not met but that the closed loop system is actually unstable. Indeed, this is the motivation for the widespread study of robust control theory.

The model to be used in such a control system design can be obtained via a first principles physical modelling of the plant and subsequent measurement of physical quantities such as masses and temperatures. Alternatively, the input-output response of the plant may be observed and recorded and a black-box model may be chosen that most accurately describes this observed response in terms of some suitable measure.

If this latter system identification approach is taken it is intuitively reasonable that the details of its implementation should depend on the control system design tools available. However, as evidenced by the publication dates of available monographs [80, 133, 142, 24], until recently the two schools of robust control and (to a lesser extent) system identification were still in a developmental phase and, as is natural, progressed relatively independently. Control theorists assumed knowledge of plant information in a certain format and identification theorists provided plant information in a format they assumed would be useful.

The discovery that these two formats were not compatible has provided a catalyst over the last five years for an attempt to marry system identification and robust control methods. This effort has seen the genesis of new methods that seek to retain existing robust control theory and develop new estimation theory. This generally involves avoiding stochastic assumptions in favour of non-stochastic magnitude bounded disturbance models and then considering the worst possible estimation error.

This may at first seem surprising since a stochastic model for disturbances is certainly well entrenched in systems science and has enjoyed a long and successful history of application. In particular, when a stochastic model is valid, the efforts of system identification researchers over the last several decades have resulted in a large and complete body of theory to explain the characteristics of the estimated model, including a quantification of the errors in the model.

The motivation for discarding this well developed and sophisticated theory in favour of

deterministic descriptions for residuals (the difference between observed data and predictions of a model) therefore needs to be thoughtfully examined. Indeed, the work we survey makes a careful case for such a choice, and the arguments put forward can be summarised as follows:

1. The residuals are likely to be the result of complex dynamics which are not captured in the model. Residuals arising from this source are not well described by the stationary stochastic process model paradigm, but might be better modelled deterministically as being simply bounded in magnitude.

2. Bounded magnitude models for disturbances lead to hard (non-probabilistic) bounds on estimation error and this is in keeping with the zeitgeist surrounding modern robust control.

3. It is felt that the more conventional stochastic process model requires too much prior information such as correlation structure and probability distribution to be specified.

These are compelling arguments, and we will be making frequent reference to them. In particular, we believe item 1 is of central importance to the development of new ideas in system identification. For example, in response to this concern one could suggest a strategy of merely increasing the model complexity. However, a central maxim of 'conventional' system identification theory is that this is often disallowed due to the overwhelming effects of measurement noise; the so-called bias/variance tradeoff.

Put most simply, what we mean by this tradeoff is that the signal to noise ratio in the available data sets an upper limit on the complexity of the model that can be fitted to it before the corrupting influence of the noise (variance error) becomes intolerable. It may well be that at this level of limited complexity, the residuals are not well described by a stochastic process (bias error), and so any error analysis based on stochastic assumptions will be bogus. Error analysis based on a deterministic model for residuals may not suffer the same fate at this bias versus variance error tradeoff point. With these ideas in mind, the literature we will survey can be broadly classified according to how residuals are described as:

Deterministic and bounded in magnitude in the frequency domain: This obviously has the flavour of H_∞ design methods and so has become colloquially termed 'Estimation in H_∞' [105, 56, 87, 107, 41, 42, 63]. To be slightly more specific, the method deals with using observed frequency response data to provide models in H_∞ with $\|\cdot\|_\infty$ norm error bounds. Such models can obviously be bolted directly to existing H_∞ control design theory.

Deterministic and Bounded in Magnitude in the time domain: These methods can provide more explicit information in the form of non-probabilistic frequency dependent bounds [139, 66, 65, 67, 144] or can provide less specific norm bounds on frequency responses [62, 86, 132, 112, 49, 85, 64, 69]. The former methods generally, but not always, use ideas from so-called 'set estimation' or 'worst case' estimation

theory that were developed in the early 1980's. The latter methods we mention are often collectively referred to as 'ℓ_1 estimation theory'.

Stochastic, non-stationary and correlated with the input signal. This is the so called 'stochastic embedding' approach [117, 39, 34]. A motif of this work is that stochastic ideas are not avoided, but rather an effort is made to extended to be able to describe undermodelling induced error.

It is here that we must make the apology that is obligatory in any survey paper to the work which is undeniably relevant but which must be left out to avoid a prohibitively long treatment. We will be conducting our survey under the above general headings in order to try to provide some sort of unifying framework for the new material, and so we elect to not comment on work which doesn't neatly fit into this structure. Among this material is

Mixtures of Frequency Domain and Time Domain Techniques. These methods can roughly be summarised as consisting of a first part in which an empirical transfer function estimate (ETFE) [80] is calculated from time domain data and then a second part in which a second model is fitted to the ETFE in the frequency domain. Indeed, one of the earliest papers in the field we are surveying [72] which later appeared as [73] falls into this category. More recent work related to this area is [136, 137, 47, 6, 7].

Deterministic Formulations for Model Validation. The area is not strictly involved with estimation of model quality, although it does seek to marry estimation and robust control theory. The effort here is to come up with model validation methods that assume deterministic models for disturbances and are tailored to robust control performance and stability criterion. Work in this area includes [123, 114, 75].

Iterative Identification and Control Design. This is a very active area that was begun by Bitmead, Gevers and Wertz in [10, 11, 9, 32] and has now been taken up by many other authors [76, 78, 77, 118, 119]. The idea here is not so much to quantify the model quality, but to be aware of how it's nature may be shaped by appropriate experiment design and data prefiltering. This leads to methods in which the controller design and the system identification design complement one-another. Interestingly, the most recent work in this area suggests avoiding the system identification phase altogether [59].

In order to provide a reference point for our discussion, we will begin the paper proper with a brief review of some simple stochastic estimation theory that predates these new methods. This will allow us to explain the motivation for the new methods via the bias/variance tradeoff.

Each approach we survey we will profile via a running simulations example. The common denominator in our simulations is not meant to be the exact measurement set up. Instead, we consider a given unknown plant, with a given level of noise corruption in measurements. The issue then is to decide on an experimentation strategy to extract information about the plant in a format suitable for subsequent control system design. The

methods we survey suggest very different strategies. Hence our profiling of them via simulation example sometimes involves different inputs, or different amounts of data collection. However, the plant and (within minor variations) the noise corruption remain the same.

In adopting this approach, we have chosen a special vantage point of asking the question 'suppose one limits consideration to the simplest case of linear time invariant (LTI) systems with measurement noise - what are the unsolved issues in this setting and how does the new theory relate to these issues ?'. We recognise this LTI problem as being highly idealised, but historically it has been the impetus for the work we are about to review. Perhaps if the LTI problem is solved, a prototype for the solution of more challenging and realistic problems will be available.

Other papers in this issue [89] and references in this paper [94] take a wider view and deal with perhaps more fundamental and global issues such as the exact nature of uncertainty and how it should be modelled. Readers seeking to progress to detail beyond the survey level here and those seeking opinions alternative to that presented here will find the essays and papers stemming from the 1992 Santa Barbara Workshop [122] of interest, and the work in the special issue [1] of relevance. As well, the paper [89] which is also in this volume provides an excellent review of current thinking in the area.

2 Conventional System Identification

Before examining new approaches to system identification we first reflect on the nature of existing 'classical' approaches which are based on stochastic theory. This will allow us to appreciate the motivation for alternative methods. The generic problem considered in these older approaches is one where an input sequence $\{u_k\}$ is applied to a linear system $G_T(q)$ and measurements $\{y_k\}$ of the response are collected which are corrupted by a noise sequence $\{\nu_k\}$:

$$y_k = G_T(q)u_k + \nu_k. \tag{1}$$

There are many means available to provide an estimate $\widehat{G}$ of G_T based on the observations $\{y_k\}$,$\{u_k\}$ when $\{\nu_k\}$ is described via a stochastic process. See [80, 133, 37] for a comprehensive survey of methods. One of the most popular ideas is the so-called 'least squares' technique where a model $G(q,\theta)$ parameterised by a vector θ of d parameters is fitted to the observed data by minimizing the quadratic norm of the prediction error. What we mean is, the estimate $\widehat{G}$ is formed as $G(q,\widehat{\theta}_N)$ where

$$\widehat{\theta}_N = \arg\min_{\theta} \left\{ \sum_{k=0}^{N-1} \left(y_k - G(q,\theta)u_k\right)^2 \right\}. \tag{2}$$

This calculation depends on the model structure chosen, but the simplest case occurs when θ parameterises $G(q,\theta)$ linearly

$$G(q,\theta)u_k = \phi_k^T\theta. \tag{3}$$

As an example, if the recently popular Laguerre model structure [138] is employed

$$G(q,\theta) = \sum_{k=0}^{d-1} \theta_k \mathcal{B}_k(q,\xi), \qquad \mathcal{B}_n(q,\xi) \triangleq \frac{\sqrt{1-\xi^2}}{(q-\xi)} \left(\frac{1-q\xi}{q-\xi}\right)^{n-1}$$

then ϕ_k^T is taken as

$$\phi_k^T = \left[\mathcal{B}_0(q,\xi)u_k, \cdots, \mathcal{B}_{d-1}(q,\xi)u_k\right].$$

We will use this model structure in future simulations since it allows the incorporation of prior information about G_T by appropriate choice of the free parameter ξ.

In any event, if the model structure admits the linear expression (3) then it is well known (2) can be solved in closed form as

$$\widehat{\theta}_N = P_N \sum_{k=0}^{N-1} \phi_k y_k, \qquad P_N^{-1} \triangleq \sum_{k=0}^{N-1} \phi_k \phi_k^T. \tag{4}$$

This solution is often written more compactly using vector notation as

$$\widehat{\theta}_N = P_N \Phi Y, \qquad P_N^{-1} = \Phi\Phi^T, \tag{5}$$

where

$$Y^T \triangleq [y_0, \cdots, y_{N-1}], \qquad \Phi^T \triangleq [\phi_0, \cdots, \phi_{N-1}]. \tag{6}$$

Using this notation, if $\{\nu_k\}$ can be modelled as a realisation of a zero mean stationary stochastic process with covariance matrix

$$C_\nu = \mathsf{E}\left\{VV^T\right\}, \qquad V^T \triangleq [\nu_0, \cdots, \nu_{N-1}],$$

then the accuracy of the estimate $\widehat{\theta}_N$ can be judged via its covariance

$$C_\theta \triangleq \mathrm{Cov}\left\{\widehat{\theta}_N\right\} = P_N \Phi^T C_\nu \Phi P_N.$$

This covariance can be made smaller by first weighting the data in inverse proportion to it's uncertainty. In this case the least squares estimate is known as the Markov estimate and instead of the formulation in (5) is given by

$$\widehat{\theta}_N = P_N \Phi^T C_\nu^{-1} Y, \qquad P_N^{-1} = \Phi^T C_\nu^{-1} \Phi \tag{7}$$

and the variance of this estimate, which is the smallest obtainable for an unbiased estimator is $C_\theta = P_N$. If $\{\nu_k\}$ is a Gaussian distributed process, then this Markov estimate is also the Maximum Likelihood Estimate. Further still, if θ is a-priori assumed to have a Gaussian distribution of very large (tending to infinity) variance, then the Markov estimate is also the Bayesian maximum a-postiori (MAP) estimate [1].

[1] See section 5.2 for a discussion of Bayesian estimation

If the frequency response $G_T(e^{j2\pi f})$ is a linear function of the parameters:

$$G_T(e^{j2\pi f}) = \Gamma(f)\theta_0, \qquad \Gamma(f) \triangleq [\mathcal{B}_0(e^{j2\pi f}, \xi), \cdots, \mathcal{B}_{d-1}(e^{j2\pi f}, \xi)], \tag{8}$$

then the accuracy of the resultant frequency response estimate can be judged using the variance

$$\mathbf{E}\left\{|G_T(e^{j2\pi f}) - G(e^{j2\pi f}, \widehat{\theta}_N)|^2\right\} = \Gamma(f) P_N \Gamma^\star(f).$$

These results can be made slightly more precise by using the fact that under mild regularity conditions [80]

$$\frac{1}{\sigma_\nu} P_N^{1/2} \left(\widehat{\theta}_N - \theta_0\right) \xrightarrow{D} \mathcal{N}(0, I) \qquad \text{as } N \to \infty$$

so that writing

$$g^T \triangleq \left[\mathrm{Re}\left\{G_T(e^{j2\pi f}) - G(e^{j2\pi f}, \widehat{\theta}_N)\right\}, \mathrm{Im}\left\{G_T(e^{j2\pi f}) - G(e^{j2\pi f}, \widehat{\theta}_N)\right\}\right], \tag{9}$$

$$\Omega(f)^T \triangleq \left[\mathrm{Re}\left\{\Gamma(f)\right\}^T, \mathrm{Im}\left\{\Gamma(f)\right\}^T\right] \tag{10}$$

gives

$$g^T \left(\sigma_\nu^2 \Omega(f) P_N \Omega(f)^T\right)^{-1} g \xrightarrow{D} \chi_2^2 \qquad \text{as } N \to \infty \tag{11}$$

and this allows us to draw confidence region ellipses on estimated Nyquist diagrams[2]. Collectively, these ideas seem a fairly complete answer. Both a nominal model and error bounds about this model can be supplied from observed data, and this sort of information could be used for robust controller design.

The objection could be raised that the bounds are probabilistic and current robust control design methods require hard bounds to guarantee closed loop performance. In answer

- Robust control design methods will still guarantee performance with probability that of the true system being within the bounds. As the bounds are tightened the closed loop performance may improve, but with increasing chance of instability. Indeed, we suggest that real world control problems are nearly always solved by aiming for high performance in the belief that the set of pathological conditions associated with extreme bounds will rarely, if ever, occur. Control engineers always work with a tradeoff of uncertainty versus performance.

- Bounds on system models will always be probabilistic since the prior information necessary to generate the bounds can never be known with complete certainty.

A more serious criticism of bounds such as (11) is that they are derived on the assumption that the vector θ_0 is capable of parameterizing $G_T(q)$ completely. What if there is some

[2]If the ν_k are Gaussian distributed then the confidence regions are exact for finite data, otherwise they are only asymptotically correct.

undermodelling $G_\Delta(q) = G_T(q) - G(q, \theta_0)$ involved with the parameterisation? What we mean is, what if a realistic description for the data in fact is

$$y_k = \phi_k^T \theta_0 + G_\Delta(q) u_k + \nu_k. \tag{12}$$

Now our confidence regions are meaningless because their calculation was predicated on the assumption that a stochastic process $\{\nu_k\}$ could describe anything not captured by the model structure $\phi_k^T \theta_0$ and this assumption is not true since (12) contains a deterministic error sequence component $\{G_\Delta(q)u_k\}$.

The question then arises as to how we can handle this deterministic sequence so that the effects of $G_\Delta(q)$ can be factored into our calculation of error bounds for estimated models. The new estimation methods we will be surveying provide an answer by avoiding stochastic assumptions in favour of deterministic assumptions.

Before we delve into these new methods, we have to answer an obvious question. Why not just increase the dimension of θ so that the model structure is so rich that G_Δ is small enough to be ignored? The answer is that the signal to noise ratio in the data may not be large enough to support the accurate estimation of such a high dimension θ. This phenomenon is known as the bias/variance tradeoff.

3 Bias/Variance Tradeoff

The work we review has arisen either from a desire to quantify undermodelling induced errors with LTI systems, or has been applied to this problem after the fact. Because of this we believe it is essential to examine why undermodelling arises in the context of LTI system identification in the first place.

The only reason we know of is the bias versus variance tradeoff, since if it did not exist, one could simply increase the model complexity until there was no undermodelling induced error that needed quantifying. Of course, in more complicated settings, such as non-linear and time varying systems, there will be reasons beyond this one, but we have already explained why we eschew these more complicated cases.

We should also point out that, particularly in the realm of set membership estimation and the work inspired by the information based complexity framework, the original motivation that inspired the development of the subject has not come from this undermodelling quarter. Rather is has come from more fundamental system theoretic considerations, and it is only in the sequel that the resulting ideas have been pressed into the service of quantifying undermodelling induced errors. Since the latter was never the primary motivating factor, the linking of a discussion bias/variance tradeoffs with these new schemes may seem discordant. However, from the perspective of this paper, where we try to argue that the problem of quantifying undermodelling induced errors in LTI systems identification is still open, we believe the bias/variance tradeoff is fundamental.

The reality of this tradeoff is part of the folk wisdom of the system identification field and so is often taken for granted. It has been rigorously shown to exist for a range of model structures in [83, 81, 138, 134]. Indeed, some of these results could be considered the earliest

attempts at addressing the issues that have driven the new work we are surveying. Since the arguments and motivations in our paper depend so crucially on this tradeoff, we pause to illustrate it with a simulation example that we will carry through the rest of the paper in an attempt to make our review as concrete as possible.

To begin with, the bias/variance tradeoff stems from the fact that estimation error consists of two parts. One component is due to the fact that the true system is not within the model structure that has been chosen, and this is called bias error. The second component of the total error is due to the noise corruption of the observed data. This is termed variance error.

Now, it is intuitively obvious that the bias error drops with increasing model order since the generality of the model structure is increased. The contribution of [83, 81, 138, 134] is to show that the variance error increased with model order, although it also decreases with number of observed data points. The net result is that for a given amount of data, there is an optimum model order that balances the decrease of bias error with the increase in variance error and gives the smallest total error.

It may so happen that for short data lengths that are badly noise corrupted the tradeoff occurs at a model order $\dim\{\theta\}$ that is so low that no matter what choice of θ is made in the model structure $G(q,\theta)$, the prediction residuals can not be well described by a stochastic process. In this case, confidence intervals will not accurately reflect the estimation errors, and it is at this point that some of the new methods we mean to survey may be of interest.

To fix this idea, suppose we have a true system which is linear and has the transfer function description

$$G_T(s) = \frac{e^{-2s}}{(s+1)(10s+1)}. \tag{13}$$

Suppose also that the available data about this system is available in the form of its sampled input-output response. The sampling rate used is 1Hz and the data record available is 500 seconds long. In fact, the available data is also noise corrupted by an uncorrelated zero mean and Gaussian distributed process with a variance of 0.005. The input excitation to the system is a unit amplitude 0.02Hz fundamental frequency square wave with a 0.2V dc offset. Finally, suppose that we try to estimate the frequency response by fitting Laguerre basis function models of various orders.

These Laguerre models (which were introduced in the previous section) are selected to have a pole position ξ corresponding to 0.2 radians per second in continuous time. Because this is only a simulation study we have the luxury of conducting this hypothetical problem in a Monte-Carlo fashion in such a way that for each model order from 1 up to 10 we perform 50 simulation experiments with the same input sequence, but a different noise realization. We can then average the resultant frequency response estimate over these 50 different noise realizations for each of the ten model orders in order to estimate the variance error. Since we also know the true system, we can calculate the average L_2 frequency response estimation error versus model order.

This has been done with the results plotted in the left hand diagram of figure 1. The solid line is the total error. The dash-dot line is the bias error that was obtained by

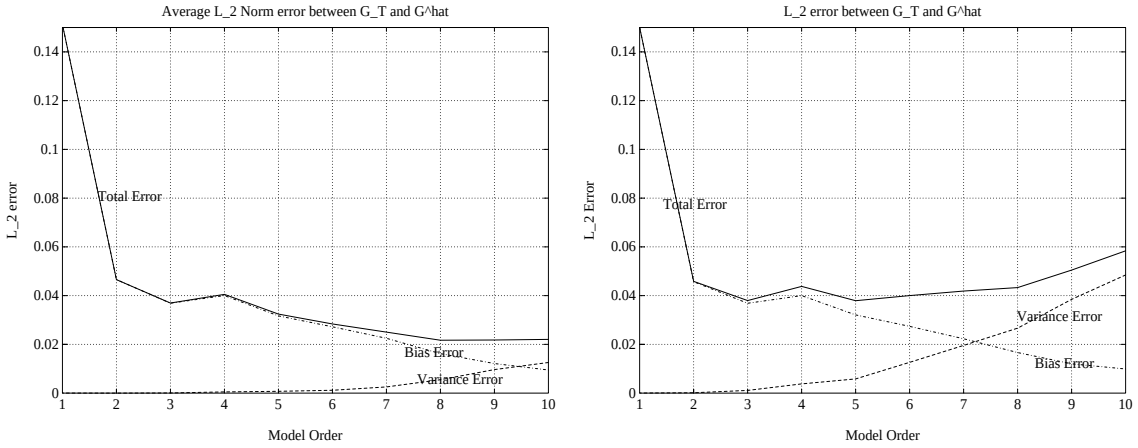

Figure 1: Bias Error, Variance Error and Total error for different model orders. On the left is the case of 500 observed data points corrupted by zero mean Gaussian noise of variance 0.005, on the right is the case of only 50 data points, and hence a smaller signal to noise ratio.

examining one run of the simulation with no measurement noise. Note that we have been careful to remove the effect of initial conditions on this error by waiting until all initial conditions in the regressor filters have essentially died out before collecting data. The reader may wonder why the bias error does not decrease monotonically. The answer is that the frequency response estimate minimises a norm weighted with the input spectral density, but the L_2 error we calculate is unweighted, and hence does not necessarily decrease when the weighted one does.

As can be seen, with respect to an L_2 frequency response error criterion the optimum model order for the 500 data point record is an 8th order one. Furthermore, at this point the bias error is comparable with the variance error. It would therefore seem roughly appropriate to follow the suggestion of Ljung et al. [82] and form total error bounds for the estimated model by simply calculating the variance error as per (2) or (9)-(11) and then doubling it. The results of this approach are shown in figure 2, and in a control design setting they may serve as a perfectly acceptable engineering solution to error quantification. Now let us consider a more difficult problem where there is significantly less data available; only 50 data points. The Monte-Carlo simulation results corresponding to this case are shown in the right hand diagram of figure 1. The situation is now quite different. From figure 1 the best model to choose from an L_2 error perspective is a third order one for which the bias error completely overwhelms the variance error. This is shown on the left of figure 3. On the right of this same figure, the results when an eighth order model is used are presented.

It is here that we believe an open problem lies. In the right of figure 3, the noise has increased the variance error component to a point were we can only obtain an unreliable estimate. On the left, we can improve the accuracy of the estimate by reducing the model order to lower the noise induced variance error. However, a concomitant feature is that the bias error now dominates. How this error should be quantified is then an open question

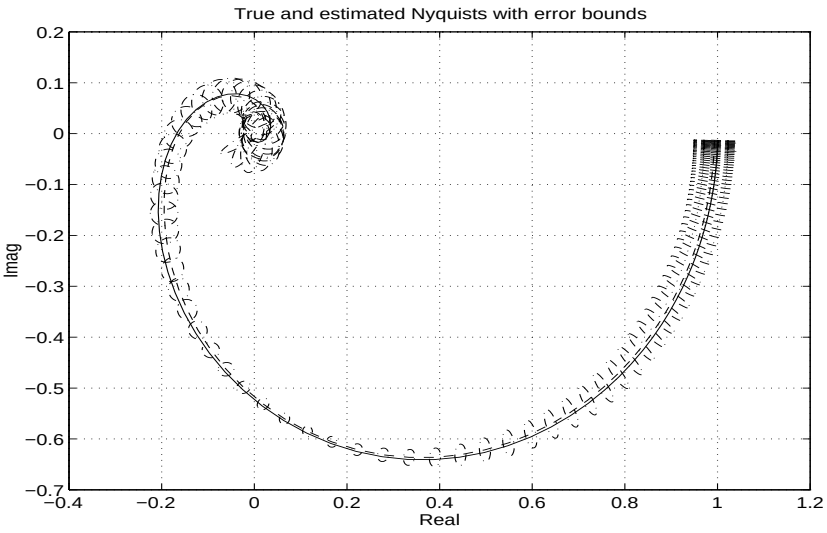

Figure 2: True and estimated Nyquist plots. Solid line is true system. Dashed line is estimate. The estimate is formed by fitting an 8th order Laguerre model via least squares to 500 observations of square wave excitation data. Dashed ellipses on the right are 99.99% confidence regions

that the work we are about to survey is aimed at redressing. We do not pretend that this short data record/bias error dominated problem is the only one that the surveyed work is relevant to. Rather we choose it as a convenient motivator.

4 Model Validation

Before we proceed with our survey, it is essential to discuss the rôle that the fundamental system identification principle of model validation plays in this work. This is so, since in any serious identification study, a validation step is performed to determine the complexity of the model used, and hence the undermodelling induced error that needs to be quantified.

The issue of model validation in this context has recently been explicitly addressed by Guo and Ljung in [45]. There they argue that typically the variance error dominates the bias error in at least two cases,

1. the model is chosen to minimize the total error, or

2. the model is chosen so as to pass a model validation test.

Since this contradicts the running simulation example we propose in which a model structure is chosen such that the bias error does in fact dominate the variance error, we need to provide some clarification.

Firstly, with respect to the first case of the model chosen to minimize the total error, the contradiction is resolved by noting that we have provided an example that is atypical according to the assumptions in [45]. In that work there is an assumption of certain relative

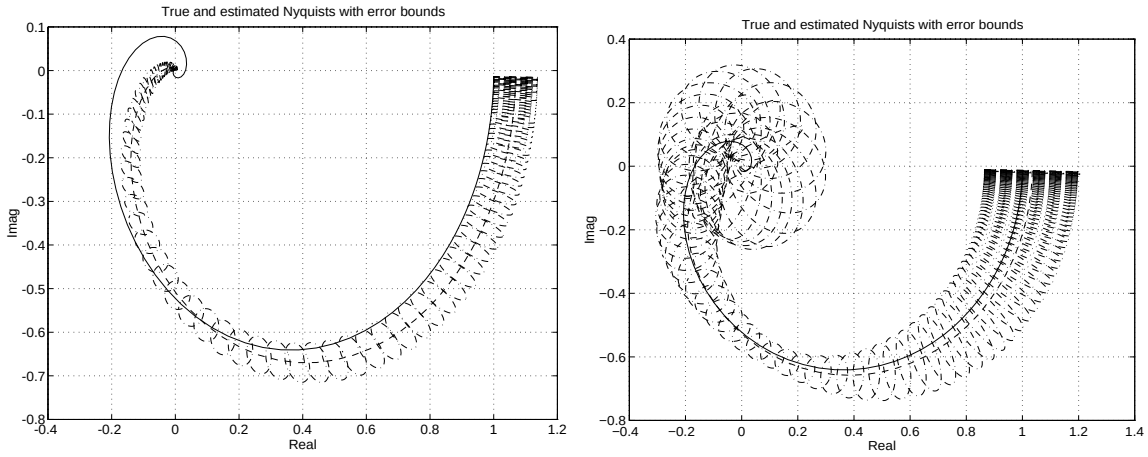

Figure 3: True and estimated Nyquist plots together with stochastic error bounds. Solid line is true system. Dash dot line is estimate. Data used to build the model was 50 observations with a 0.02Hz fundamental square wave input. Data was corrupted by zero mean Gaussian distributed noise of variance 0.005. Dashed ellipses are 99.99% confidence regions. On the left an eighth order Laguerre model was used and on the right a second order model was used.

rates of change of bias and variance errors with change in model complexity, and we violate them by choosing a small, highly noisy data set in which the variance error increases with model order faster than supposed in [45].

With respect to the second case, and in the context of model validation more generally, we note that such schemes typically involve setting up some sort of hypothesis testing problem in order to decide on a model structure. This does not necessarily lead to the choice that is best for control system design. It is better adapted to providing a choice that is appropriate for a prediction application.

To put another way, we argue that the most suitable model for control system design may fail the standard validation tests. We note that this latter consideration has recently led to some workers reconsidering the model validation problem itself with the aim of developing new schemes that are more appropriate in a control design setting [123, 114, 75] than existing schemes [80] are.

One of the motivations for this latter control relevant work, is that if the modelling residuals arise due to non-linear effects, and so are highly correlated with the input, then it is possible that no linear model structure will pass a 'standard' model validation procedure. Then the choice of model structure needs be done via other criteria. As a final point in this vein, we suggest that conventional model validation tests are appropriate as indicators of situations when significant bias errors do exist that may need to be quantified.

5 Identification from Time Domain Data

We begin our review of the literature by examining estimation methods that work from time domain data. We will subdivide our presentation under the headings of 'Worst Case Estimation Theory', 'Stochastic Embedding of Undermodelling' and 'ℓ_1 estimation'. The latter methods have links to the 'Estimation in H_∞' methods we will be discussing in section 8, and also to the worst case estimation theory we examine in the following section.

5.1 'Worst Case' Estimation Theory

The basic idea of 'Worst Case Estimation Theory' is to take observations of data, a parameterised model, and magnitude bounds on errors in the data and then find a region in parameter space which is compatible with these three pieces of information.

In other words, the techniques of worst case estimation theory can be viewed a methods of re-expressing the combination of assumptions (error bounds, model structure) and data (observations) in a more convenient and succinct format; this format happens to be as a region in a parameter space. The theory was developed independently of the interest in error quantification for control system design [94, 126, 125], but has been adapted to this latter purpose [139, 71].

The work of worst case estimation theory sometimes also appears under the names of 'set estimation' theory, 'bounded error' estimation theory, and 'robust' identification. However, since the appearance of several general works [94, 19, 126] unifying these and other ideas in an information based complexity setting [130, 131, 106], the term 'worst case estimation' theory seems to have emerged as the most generally accepted term and so we employ it.

There is by now a very large body of work on this area. See for example the seminal and survey papers [120, 30, 31, 91, 97, 92, 94, 141, 101, 100, 102, 22, 19]. The main algorithmic difficulty is that the parameter space region (Membership set) consistent with observations and assumptions can be a complicated shape which is difficult to characterise. In practice it is necessary to find a simpler outerbounding set for this shape. Such a set is, of course, non-unique and it is desirable to find the smallest one possible so as to closely approximate the complicated set.

However, for many cases of practical significance the algorithms involved are simple. Furthermore, the assumptions on disturbances are minimal. No probability density functions need be specified; only a bound on the magnitude of the disturbance need be known. Finally, the methods are eminently suitable for providing models complete with error bounds that are suitable for robust control system design. This is so since the manifestation of a stable undermodelling can readily be formulated as a deterministic bounded disturbance. Reflection on the rapproachment between these new methods and pre-existing estimation methods based on stochastic disturbance models can be found in [70, 128, 3, 57, 90].

Given these preliminary comments, the ideas we want to survey can be explained as

follows. Consider the original estimation problem (1) we posed in section 2

$$y_k = G_T(q)u_k + \nu_k \tag{14}$$

where instead of using a stochastic model for the disturbance sequence $\{\nu_k\}$ we use the much simpler magnitude bounding model of

$$|\nu_k| \leq \delta_k \tag{15}$$

and also postulate, as we did in section 2, a description $G(q,\theta)$ of $G_T(q)$ that is parameterised by d variables collected in the vector θ. Sensible values of θ must be consistent with the assumption (15) on disturbances, the assumption (14) on data generation and on the model structure $G(q,\theta)$. This means that θ should satisfy

$$|y_k - G(q,\theta)u_k| \leq \delta_k,$$

but this must apply for every sample we observe in an N point data record, so feasible θ must belong to the 'membership set' (aka 'feasible parameter set') Θ:

$$\theta \in \Theta = \bigcap_{k=0}^{N-1} \{\theta : |y_k - G(q,\theta)u_k| \leq \delta_k\}. \tag{16}$$

In mathematical terms, Θ is the pre-image of the map

$$\begin{aligned} G(q,\theta) : \mathbf{R}^d &\rightarrow \mathbf{R}^N \\ \theta &\mapsto [y_0 \pm \delta, \cdots, y_{N-1} \pm \delta]. \end{aligned}$$

and viewed this way, depending on the model structure $G(q,\theta)$, the map may be linear or non-linear, and hence the pre-image or membership set Θ may be easy or difficult to compute and characterise. In fact, if the map is non-linear, then the only methods available for finding Θ are essentially Monte-Carlo on choices of θ, perhaps with modifications on how test θ should be chosen so that that once boundaries of Θ are found, they may be tracked [102, 93].

Because of this difficulty, the earliest work in the area [120, 121] was concerned with estimating the state vector in a linear system. In our parameter estimation setting this corresponds to the model structure $G(q,\theta)$ being linear in the parameters

$$G(q,\theta)u_k = \phi_k^T\theta \tag{17}$$

which we first considered in (3) of section 2. In turn, this leads to a linear map that we want to find the pre-image Θ of. In this case Θ is at least easy to formulate since the equations

$$y_k - \phi_k^T\theta = \pm\delta_k$$

specify a parallel pair of straight lines in parameter space that are both orthogonal to ϕ_k and are separated by a distance $2\delta_k$. The region $\{\theta : |y_k - \phi_k^T\theta| \leq \delta_k\}$ is located

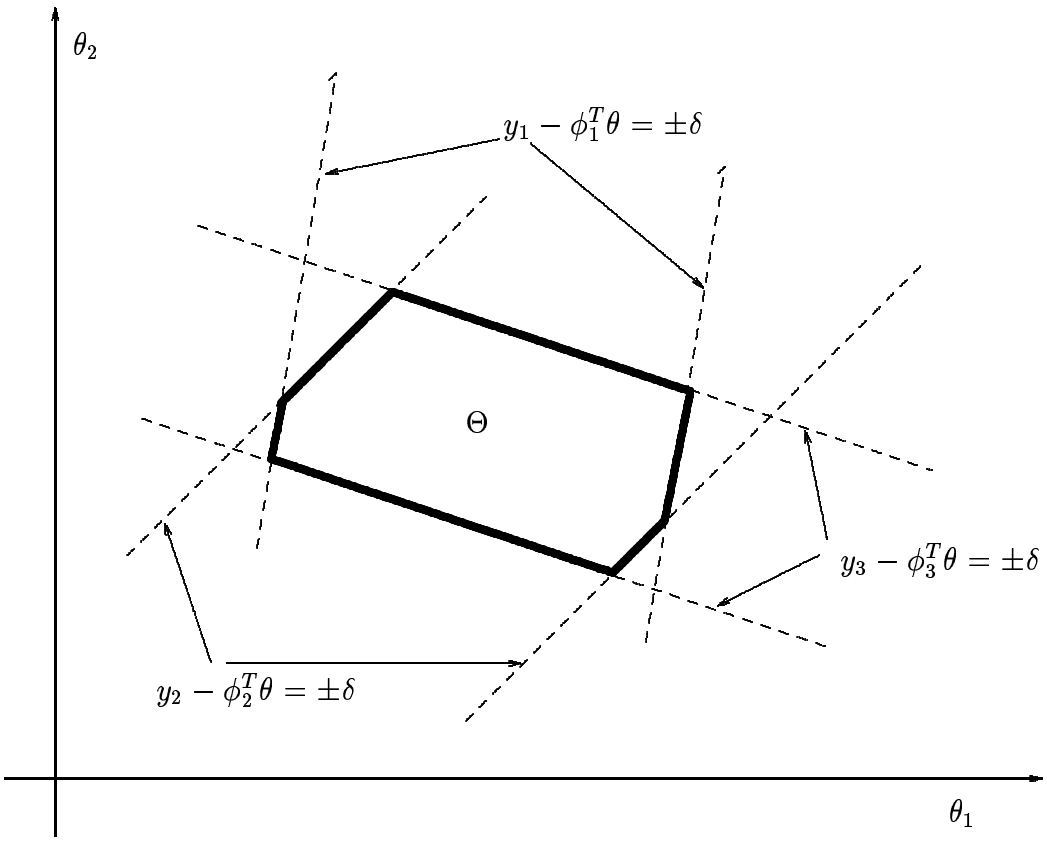

Figure 4: Calculation of Feasible Parameter set Θ for model structure $G(q, \theta)$ that is linear in the parameters.

between these lines. The calculation (16) then becomes the intersection of simple straight line bounded regions so that Θ is a convex polytope. This is illustrated in figure 4. The calculation of this polytopal Θ may be done numerically via linear programming, but since this will involve $2N$ constraints it could be computationally intensive. In response to this, a number of authors have developed efficient recursive algorithms for computing polytopal Θ [97, 140, 124]. Since the polytope Θ is unique, these methods differ only on the computer memory versus computational load tradeoff implied by the data structures used in representing the polytopes.

Indeed, this decision of data structure for representing the polytope raises an important point; it is difficult to concisely represent polytopal Θ in an easily manageable form. Some sort of complicated data structure is needed.

A popular solution to this representation problem is to use an ellipsoidal overbounding region to approximate Θ since such a shape can be economically represented by a positive definite matrix. To be more explicit, any θ consistent with the linear in the parameters model (17) and the disturbance assumption (15) must, for any choice of positive definite

weightings $\{\rho_k\}$, satisfy the following inequality

$$\sum_{k=0}^{N-1} \frac{\rho_k}{\delta_k^2}(y_k - \phi_k^T\theta)^2 \leq \sum_{k=0}^{N-1} \rho_k \qquad ; \rho_k > 0. \tag{18}$$

This summation operation immediately coarsens our discrimination of Θ. Some θ that satisfy the above inequality will not be in the intersection (16). The point of the weighting sequence is that by appropriate choice of $\{\rho_k\}$ this phenomenon can be minimised. If we express (18) in vectorised form using the notation in (6) we have

$$(Y - \Phi\theta)^T M(Y - \Phi\theta) \leq \sum_{k=0}^{N-1} \rho_k, \qquad M = \operatorname*{diag}_{0\leq k\leq N-1}\left\{\frac{\rho_k}{\delta_k^2}\right\}. \tag{19}$$

Simple linear algebra allows us to write the left hand side of this expression as

$$(Y - \Phi\theta)^T M(Y - \Phi\theta) = (\widehat{\theta}_N - \theta)^T P_N^{-1}(\widehat{\theta}_N - \theta) + \varepsilon^T M\varepsilon \tag{20}$$

where $\varepsilon = Y - \Phi\widehat{\theta}_N$ and

$$\widehat{\theta}_N = P_N\Phi^T MY, \qquad P_N^{-1} = \Phi^T M\Phi. \tag{21}$$

These last two equations are precisely the same as the Markov estimate (7) if we associate the positive semi-definite weighting matrix M with the inverse covariance matrix C_ν^{-1} in (7). Furthermore, substituting (20) into (19) gives us an overbounding characterisation of Θ as the set of θ lying inside the ellipse

$$(\theta - \widehat{\theta}_N)^T P_N(\theta - \widehat{\theta}_N) \leq \sum_{k=0}^{N-1} \rho_k - \varepsilon^T M\varepsilon \tag{22}$$

which is the same as the formulation of confidence regions for the Markov estimate. The only problem with this convenient ellipsoidal overbounding approximation for Θ is that it is difficult to solve for the optimum choice of the weighting sequence $\{\rho_k\}$. In a now famous paper [31] a recursive solution to this was proposed.

To explain it, suppose that we have chosen some weighting sequence of ρ's so that we have an ellipsoidal description (22) using observations of data up to sample $k-1$. That is, we have a $\widehat{\theta}_{k-1}$ and a P_{k-1}^{-1} (that has been normalised by dividing it by $\sum_{\ell=0}^{N-1}\rho_\ell - \varepsilon^T M\varepsilon$) given by (21) such that an ellipsoidal approximation to Θ is

$$(\theta - \widehat{\theta}_{k-1})^T P_{k-1}^{-1}(\theta - \widehat{\theta}_{k-1}) \leq 1.$$

This region is shown as the large ellipse in figure 5. Now suppose we observe new data at sample number k. As already discussed and illustrated in figure 4, this refines our feasible set Θ to lie between the straight lines $y_k - \phi_k^T\theta = \pm\delta_k$. The problem is to intersect this straight line bounded region with the ellipsoidal region defined by P_{k-1} and find an

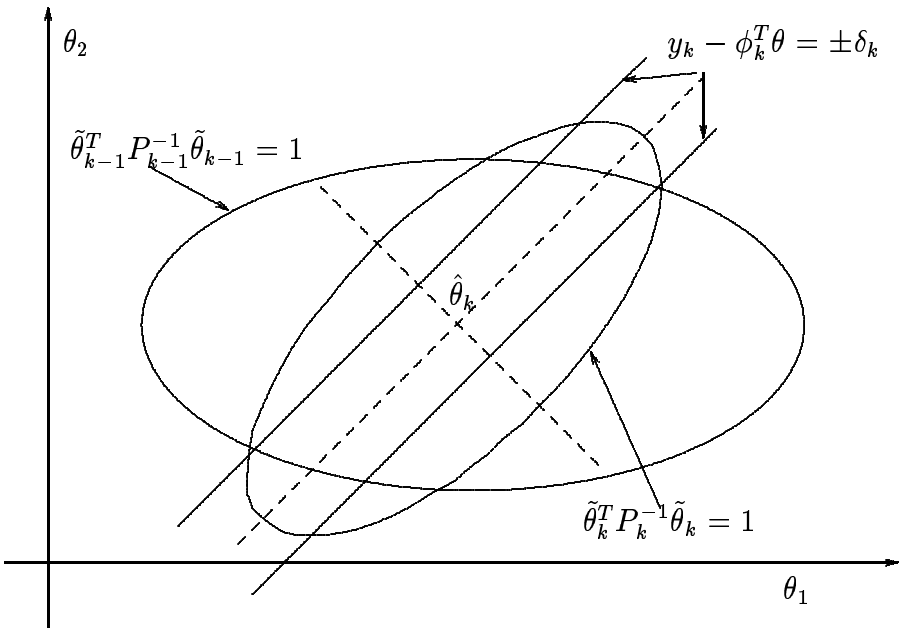

Figure 5: Illustration of how the weighting sequence $\{\rho_k\}$ can be found iteratively using the Fogel and Huang algorithm.

ellipsoidal overbound to this approximation. This can be done be forming a new ellipsoidal region as

$$(\theta - \widehat{\theta}_{k-1})^T P_{k-1}^{-1}(\theta - \widehat{\theta}_{k-1}) + \frac{\rho_k}{\delta_k^2}(y_k - \phi_k^T\theta)^2 \leq 1 + \rho_k$$

which is shown as the smaller ellipse in figure 5. The contribution of Fogel and Huang [31] was to suggest that at each iteration this ellipsoidal overbounding approximation should be made tight by minimising its volume, and that this can be achieved by choosing ρ_k as the largest positive root of the quadratic equation

$$(d-1)g_k^2\rho_k^2 - ((2d-1)\delta_k - g_k - \varepsilon_k^2)g_k\rho_k + \delta_k(d(\delta_k - \varepsilon_k^2) - g_k) = 0 \tag{23}$$

where $d = \dim\{\theta\}$, $g_k = \phi_{k-1}^T P_{k-1}\phi_{k-1}$ and $\varepsilon_k = y_k - \phi_k^T\widehat{\theta}_{k-1}$. If no positive root exists, then this indicates that the lines $y_k - \phi_k^T\theta = \pm\delta_k$ do not intersect the old ellipsoidal region $(\theta - \widehat{\theta}_{k-1})^T P_{k-1}^{-1}(\theta - \widehat{\theta}_{k-1})$. In this case, no updating of regions should be done, which is achieved by setting $\rho_k = 0$. Otherwise, the centre $\widehat{\theta}_k$ of the new ellipsoidal region becomes

$$\widehat{\theta}_k = \widehat{\theta}_{k-1} + \frac{P_{k-1}\phi_k\varepsilon_k}{\delta_k^2/\rho_k + \phi_k^T P_{k-1}\phi_k} \tag{24}$$

and the matrix P_k defining the size and orientation is

$$P_k = \beta_k\left(P_{k-1} - \frac{P_{k-1}\phi_k\phi_k^T P_{k-1}}{\delta_k^2/\rho_k + \phi_k^T P_{k-1}\phi_k}\right) \tag{25}$$

with gain term β_k given as

$$\beta_k = 1 + \rho_k - \frac{\rho_k \varepsilon_k^2}{\delta_k^2 + \rho_k \phi_k^T P_{k-1} \phi_k}.$$

Notice that with $\beta_k = 1$ the above recursions are precisely that of the Kalman Filter for the linear model (3) when $\{\nu_k\}$ is white Gaussian noise with variance δ_k^2/ρ_k.

However, in contrast to the Kalman filter, the recursions (24),(25) depend non-linearly on the data because of the ρ_k term. The upshot is that the recursions can be run more than once on the same set of data (with the re-initialisation $P_0 = P_N$ from the previous run) to try to find the smallest possible ellipsoidal bound on Θ. It has also been suggested that the recursions (24),(25) can be used as a pre-processing procedure to whittle the number of active constraints for the linear programming problem that can solve for the exact polytopal Θ.

For our purposes of survey in the context of quantifying undermodelling induced errors, this brief review has covered the essentials of worst case estimation theory. Aficionados of the field will know that we have neglected many interesting developments such as how orthotopal (box shaped) overbounding regions for Θ can be developed, how inner bounding regions for Θ may be found, how errors in u_k as well as y_k can be dealt with, details of how Θ may be found for non-linear in θ model structures, and how all these issues link with 'Information Based Complexity' ideas in Computer Science. Some of this work, and further references to this area of research can be found in [141, 102, 91, 94, 95, 126, 90, 113, 111] and the review work [89] also in this volume.

5.2 Bayesian Interpretation

In the preceding discussion, some links were shown between ellipsoidal outer bounding algorithms for Θ and estimation algorithms predicated on stochastic assumptions. The off-line equations (21) for ellipsoidal Θ are the Markov estimate (7) and the recursive equations (24) and (25) are very similar to the Kalman filter. Furthermore, both these estimates in a stochastic setting have a Bayesian interpretation. This is intriguing, and in fact the exact feasible parameter set Θ, even for model structures $G(q, \theta)$ much more general than the linear ones we have been considering, has a Bayesian interpretation. Readers may find this observation useful in trying to arrive at some sort of unified system identification perspective.

The idea with Bayesian estimation is to assume that the parameters θ to be estimated are in fact random variables, about which we have some prior knowledge which we express as a probability distribution $\mathbf{P}(\theta)$. Then, once we have observed some data $\{y_0, \cdots, y_{N-1}\}$, we incorporate this new information into our knowledge of θ by calculating the posterior distribution $\mathbf{P}(\theta \mid y_0, \cdots, y_{N-1})$ using Bayes rule as (we are neglecting a normalising constant $\mathbf{P}(y_0, \cdots, y_{N-1})$)

$$\mathbf{P}(\theta \mid y_0, \cdots, y_{N-1}) = \mathbf{P}(\theta)\mathbf{P}(y_0, \cdots, y_{N-1} \mid \theta). \tag{26}$$

It is here that we have to consider the data generation mechanism. In our worst case estimation setting we have been using

$$y_k = G(q,\theta)u_k + \nu_k.$$

Now suppose that $G(q,\theta)$ is any model structure which is strictly causal, it can be non-linear or not in θ. Suppose also that we assume that the $\{\nu_k\}$ process is independent (white noise). Then this allows us to calculate $\mathbf{P}(y_0,\cdots,y_{N-1} \mid \theta)$, again using Bayes rule, as

$$\mathbf{P}(y_0,\cdots,y_{N-1} \mid \theta) = \prod_{k=0}^{N-1} \mathbf{P}(y_k \mid \theta). \tag{27}$$

But by the 'worst case' assumption that $|\nu_k| \leq \delta_k$ it must be that $\mathbf{P}(y_k \mid \theta) = 0$ on the set $\{\theta : |y_k - G(q,\theta)u_k| > \delta_k\}$. Using this observation in (27) after it has been substituted into (26) tells us that the posterior distribution for θ is only non-zero on the set

$$\bigcap_{k=0}^{N-1} \{\theta : |y_k - G(q,\theta)u_k| \leq \delta_k\}$$

which is identical to the feasible parameter set Θ that we first formulated in (16). What we mean to say is that the set Θ arising in worst case estimation theory is precisely the same as the support for the posterior distribution of θ under the special stochastic assumptions of the disturbance sequence $\{\nu_k\}$ being white noise. In the case of model structures linear in θ, this same observation has been made in [135].

Of course, with Bayesian estimation one normally calculates the posterior distribution itself rather than just characterising its support, but this is only really possible for Gaussian distributions where a Kalman Filter can be used to perform the calculations. Worst case estimation theory can then be regarded as an attempt to do the next best thing of at least trying to calculate the region of θ where the distribution is non-zero.

Given these observations, it is possible to unify the new ideas of worst case estimation theory with pre-existing ideas. We arrive at the same conclusion Θ whether we make the special assumption of white measurement noise and find the support of the posterior distribution of θ, or whether we assert that we have made a paradigm shift by avoiding stochastic assumptions and simply finding θ consistent with observations, model structure and disturbance bounds. The choice of interpretation we leave to the readers taste.

5.3 Sensitivity analysis

There is no doubt that because worst case estimation theory eschews the sophisticated stochastic disturbance models in favour of much simpler magnitude bounding models it is very appealing. Unfortunately, as one would suspect, in some cases there is is price to be paid for reducing the required knowledge of disturbances to just a magnitude bound. The cost is that the set Θ of feasible parameters is sensitive to the veracity of this bound.

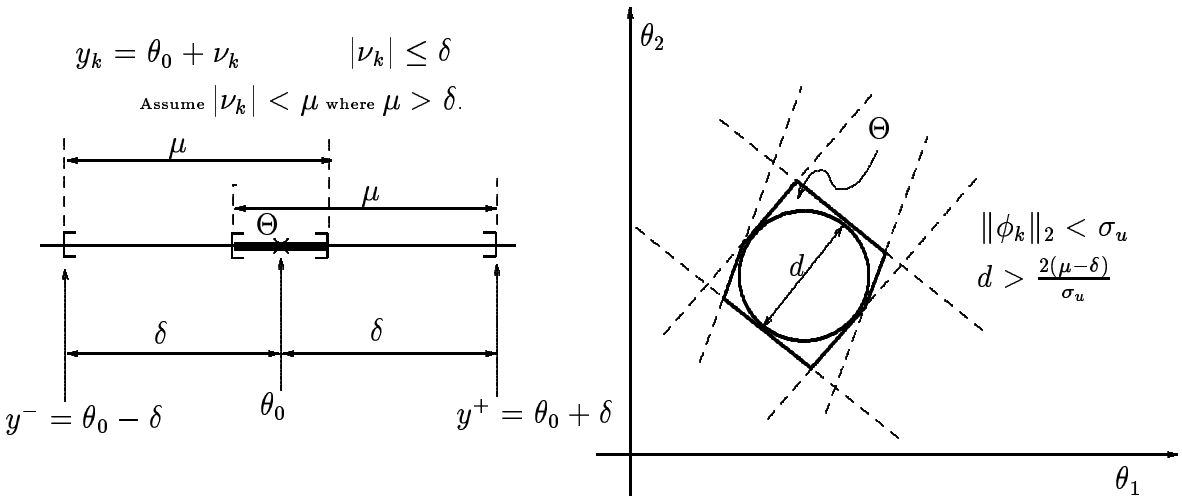

Figure 6: Illustrations showing how the feasible parameter sets Θ can be sensitive to the bound δ on disturbances.

In order to illustrate this we begin with a simple example. Suppose we want to estimate a fixed scalar value θ_0 in the presence of bounded disturbances

$$y_k = \theta_0 + \nu_k, \qquad |\nu_k| \leq \delta. \tag{28}$$

In most cases we will not a-priori know the bound δ exactly, and we will only be able to approximate the bound by some $\mu > \delta$. The sensitivity issue we refer to manifests itself in this simple case as the fact that with imprecise knowledge μ of δ we will never be able to find θ_0 to an accuracy greater than $\pm|\mu - \delta|$.

To explain what we mean, referring to the left hand diagram of figure 6 it is obvious that for the purposes of worst case estimation the most informative observations are $y^+ = \theta_0 + \delta$ and $y^- = \theta_0 - \delta$. If we are lucky enough to get these observations then we can at best conclude

$$\theta - \theta_0 \in [\delta - \mu, \mu - \delta]. \tag{29}$$

Furthermore, after observing the data points $y^+ = \theta_0 + \delta$, $y^- = \theta_0 - \delta$ we can never reduce the size of Θ, no matter how much more data we observe.

The extension of this sensitivity phenomenon to the dynamic system setting is illustrated in the right hand diagram of figure 6. What we are trying to convey in that diagram is that if the bound δ is only known as $\mu > \delta$ then we can always fit a $d = \dim\{\theta\}$ dimensional ball of radius $2(\mu - \delta)/\sigma_u$ inside the smallest possible feasible parameter set Θ [99]. The term σ_u is the largest value the Euclidean size of the vector ϕ_k attains. The general principle is that the size of Θ is sensitive in a proportional fashion to the prior assumption δ.

At this stage we should point out that worst case estimation theory is applicable to a wide variety of applications where this sensitivity phenomenon may not be a difficulty, and indeed may even be an advantageous feature. For example, in problems where the bound

δ is accurately known (such as, for example, when it arises from quantisation error [127]), then the sensitivity issue manifests itself as an extremely efficient use of data, since the flip side of the trivial example in (47) is that if δ is perfectly known, then using worst case estimation ideas, θ_0 can be perfectly known after only two observations of data if they are outliers. Analysis of this feature can be found in [135].

However, in the focus of this survey of quantifying errors for use in robust control system design we do believe that there is still work to be done with the application of ideas from worst case estimation theory. This is so since sensitivity of Θ will manifest itself as sensitivity of a controller design to prior information and this seems to contradict the original ambit of robust design which is to achieve insensitivity to the accuracy of prior knowledge. The simulation example to follow may help to make this reservation clearer.

5.4 Worst Case Estimation for Control Design

So far we have not indicated how these worst case estimation methods may be applied to the focus of this paper - error quantification for the purposes of subsequent robust controller design. In fact, to our knowledge, there are only two main schools of thought as to how this might be directly achieved. In case the reader believes this to be a paucity that should lead to disqualification from survey, we should point out that worst case estimation ideas have value beyond their direct application since they have to some extent inspired other contributions to the area, such as ℓ_1 estimation which we will presently review. Outside the context of this paper, the ideas also form a vigorous and independent area of study in its own right, with many successful applications [94].

The first school of application is due to Kosut and co-authors [66, 65, 67, 68] who were among the first to make the connection between robust control and worst case estimation. Indeed, as a co-worker Kosut is also one of the few to also consider controller designs that are married to the format of information provided by worst case estimation [74], although there are other perspectives available [20].

The central ideas in this first school of work are to use the the 'block' formulation (22), and then concentrate on examining the nature of Θ and the question of how bounds on $|G_\Delta(q)u_k|$ may be derived for various model structures $G(q,\theta)$.

The second school is espoused by Wahlberg and Ljung [139] who show how the Fogel and Huang recursions (24),(25) may be used. One of their main concerns is to examine how realistic prior information about $G_\Delta(q)$ may be translated into the required bounds on $|G_\Delta(q)u_k|$. Apart from the efforts of these two groups of authors, there have been other works in the area, for example [69], but the essential ideas available are captured by these two main groups.

Perhaps the best way to explain the application ideas of Kosut, Wahlberg and their colleagues is via pursuit of the simulation example we began in section 3 and continued in section 8.2. Recall that this involved the true plant (13) being excited with a 0.02Hz unit amplitude square wave and being observed for 50 samples, the output observations being corrupted by zero mean white Gaussian distributed noise of variance 0.005. Recall also that the point of considering this example with so little data was to examine a problem

where the bias/variance tradeoff prevented 'classical' stochastic estimation ideas being successfully being applied; otherwise it is hard to justify the incorporation of new theory. The demand for new ideas occurs because with a bias/variance imposed meager model structure of only 2 Laguerre basis functions the component $G_\Delta(q)u_k$ in

$$y_k = G(q,\theta)u_k + G_\Delta(q)u_k + \nu_k$$

cannot be well described as a realisation of a stationary stochastic process and yet is significant enough to be unable to be neglected. Worst case estimation methods would then seem a natural tool to use in attacking this problem.

From our survey we see that an essential step in using such techniques is to specify the bound δ. This is difficult. Wahlberg and Ljung [139] seem to have provided the most complete insight into how it may be done without using unrealistic assumptions. For example, following their suggestions we may perform the following calculation

$$|G_\Delta(q)u_k| = \left|\sum_{m=0}^{\infty} \eta_m u_{k-m}\right| \leq \|u\|_\infty \sum_{m=0}^{\infty} |\eta_m|$$

and it may be realistic to assume we are be able to formulate an exponential bound $|\eta_m| \leq M\rho^{-m}$ on the impulse response of G_Δ if the latter is stable. In this case, our bound for $G_\Delta(q)u_k$ is

$$|G_\Delta(q)u_k| \leq \frac{M\rho\|u\|_\infty}{\rho - 1} \tag{30}$$

and this may be incorporated into any bound available on ν_k in order to find δ via the triangle inequality. Wahlberg and Ljung [139] examine other formats for prior knowledge of $G_\Delta(q)$ such as Lipschitz smoothness constraints on $G_\Delta(e^{j2\pi f})$, but as they point out these alternatives are essentially equivalent to the above calculation.

It is here that our previous section on the sensitivity of Θ to δ becomes relevant. Let alone whether we know an accurate bound on $|\nu_k|$, the calculations leading to (30) are quite conservative and so any approximation μ we make to δ which bounds $|G_\Delta(q)u_k + \nu_k|$ will be correspondingly conservative. The size of our membership set Θ is proportional to this conservatism, and so our conclusions could be rather less precise than we would like.

To illustrate this difficulty, that we have chosen a bias/variance imposed parsimonious model structure of

$$G(q,\theta) = \theta_0\mathcal{B}_0(q,\xi) + \theta_1\mathcal{B}_1(q,\xi)$$

where the $\mathcal{B}_n(q,\xi)$ are Laguerre basis functions with the choice $\xi = e^{-0.2}$. Now $G_\Delta(q) = G_T(q) - G(q,\theta)$ so that G_Δ depends upon the choice of θ. This leads to an almost tautological problem since our estimate of θ will be predicated by our specifications of G_Δ which we have just noticed actually depends upon θ. What we mean to say is that there is no reason a priori for us to place any greater importance on one value of θ over another, and yet we have to if we are to somehow inject some prior knowledge about G_Δ. Given this difficulty in even defining G_Δ the issue of the sensitivity of Θ to a δ which depends upon G_Δ is even more acute.

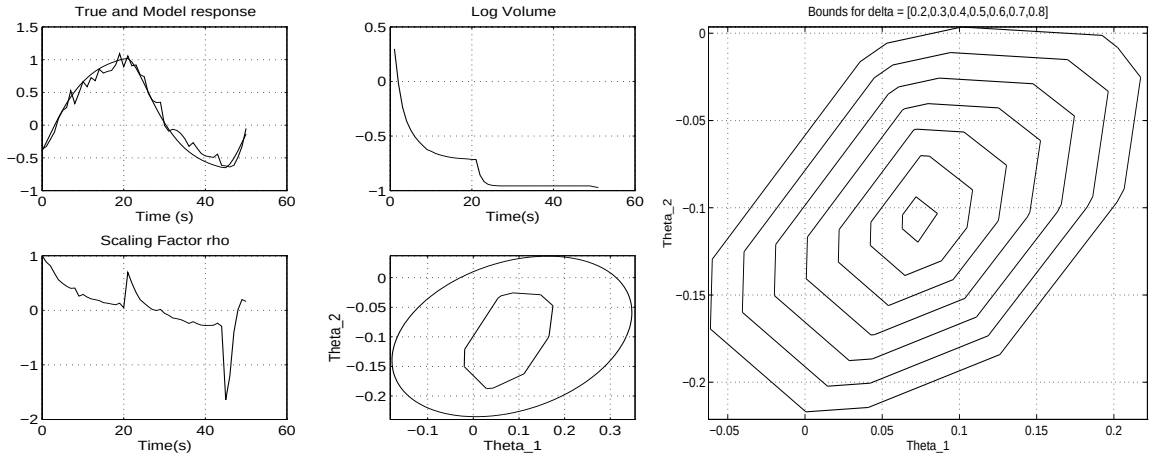

Figure 7: Using worst case estimation ideas in our running simulation example. In the left group of four figures, the results of using Fogel and Huangs recursions. The final ellipsoidal overbound and the true polytopal Θ are shown. On the right the sensitivity of this polytopal Θ to the choice of δ is shown.

Hjalmarsson [57, 58] has considered this problem. He traces the tautological difficulty to the fact that treating $G_\Delta(q)u_k$ as a disturbance is in contradiction to the usual notion of a disturbance as being something that cannot be explained by correlation with the input signal $\{u_k\}$.

Nevertheless, we can adopt the viewpoint in [139] and attempt to be unprejudiced by setting $\theta_\star^T \triangleq [\theta_0, \theta_1]$ as the first two co-efficients in the Fourier expansion of G_T with respect to the basis vectors $\mathcal{B}_0(q,\xi), \mathcal{B}_1(q,\xi)$:

$$\theta_\star = [\langle G_T, \mathcal{B}_0\rangle, \langle G_T, \mathcal{B}_1\rangle] = [0.1195, 0.1724].$$

Since we have the unrealistic luxury in this simulation example of knowing G_T we can actually calculate what $G_\Delta(q) = G_T(q) - G(q,\theta_\star)$ is, and then tightly bound its impulse response with $M = 0.16$, $\rho = 1.67$. Then, since $\|u\|_\infty = 1$ and since again through simulation luxury we know that $|\nu_k| \leq 0.2$ we can use (30) to find a bound $\delta = 0.16 \times 1.67/0.67 + 0.2 \approx 0.6$.

Using this bound, the Fogel and Huang recursions provide the results shown in the left hand diagram of figure 7. Proceeding clockwise from the top left we have the true and model response, the evolution of the volume of the ellipsoidal overbound to Θ, the polytopal feasible parameter set Θ together with the Fogel and Huang derived ellipsoidal overbound, and finally the sequence $\{\rho_k\}$ of scaling factors.

To illustrate the sensitivity issue we discussed in the previous section we have plotted the exact polytopal Θ corresponding to a range of choices of δ linearly spaced in the range $[0.2, 0.8]$ in the right hand diagram of figure 7. The direct dependence between the quality of our prior knowledge for δ and the conclusions we make about θ and hence G_T are obvious. Given the difficulty of specifying G_Δ and hence δ this would seem to be an important issue if the aim is to use the inferences about G_T for controller design.

6 Estimation in ℓ_1

This work is closely aligned with worst case estimation studies. The essentials of it are as follows. Assume data is generated as in (1) where $G_T(q)$ is described as an infinite impulse response

$$y_k = g \circledast u + \nu = \sum_{m=0}^{\infty} g_m u_{k-m} + \nu_k$$

and where, as in all the previous two sections, we assume the errors $\{\nu_k\}$ can be bounded in magnitude as $|\nu_k| \leq \delta$. Assume also that the impulse response vector g of $G_T(q)$ is known to be contained in a feasible parameter set

$$g \in \mathcal{S} = \left\{ g : |g_m| \leq M\rho^{-m} \right\}$$

for some exponential bounding parameters M and ρ. Once this prior information is defined, the issue is to find an estimate $\{\widehat{g}_m\}$ of the impulse response such that the worst case ℓ_1 norm error over all possible disturbances

$$\operatorname*{Sup}_{\substack{g \in \mathcal{S} \\ \|\nu\|_\infty \leq \delta}} \|\widehat{g} - g\|_1 \tag{31}$$

can be quantified, and hopefully minimised. It should be plain that this is a special case of worst case estimation theory where the model structure $G(q, \theta)$ is chosen as FIR and the format for describing the feasible parameter set is a special type of orthotope that can be conveniently described by ℓ_1 norm bounding. This is analogous to the previous section where weighted (by $\{\rho_k\}$) quadratic norm bounding was a convenient way to specify Θ in ellipsoidal form.

Perhaps the most important feature distinguishing ℓ_1 estimation theory from worst case estimation ideas is that in the latter no attempt is made to select any one element in the feasible parameter set Θ as being more important than any other whereas with ℓ_1 estimation, most of the attention is on doing precisely that.

Even though this 'Estimation in ℓ_1' theory is very closely aligned to worst case estimation ideas, originally it seems to have begun with [61, 62] where the motivated came from the 'Estimation in H_∞' work we discussed in section 8. This is because those authors were focused on providing estimation algorithms directly compatible with H_∞ robust control synthesis tools, and the ℓ_1 norm of the impulse response of a system is an overbound on the H_∞ norm of its frequency response.

In latter days the 'Estimation in ℓ_1' school [132, 86, 85, 126, 8, 64, 96] has drawn inspiration from 'Information based complexity' (IBC) ideas from the Computer Science [130, 131, 106] literature. The work [89] also in this volume provides an interesting overview of these IBC ideas and how they relate to and unify worst case identification methods.

The aim now seems to be to attack questions about the fundamental limits of estimation theory along the lines of Kolmogorov's ideas of approximation theory as made popular by Zames [46]; see for example [79]. The tools are mainly those of functional analysis,

operator theory and analytic topology which may be intimidating for a practitioner, but for the theoretician allows such things as the elegant unification of time and frequency domain results [112].

Perhaps the main feature of the area, which sets it apart from the previous areas we surveyed is that a lot of attention is given to the nature of the input signal [85, 84, 132]. This is so since, in the interests of minimising the worst case error, one often seeks to take $\widehat{g}$ as the Chebychev centre (element whose norm distance to any boundary of a set is minimised) of the set Θ of plants that are indistinguishable given our assumptions

$$\Theta = \{g \in \mathcal{S} : y = g \circledast u + \nu; \|\nu\|_\infty \leq \delta\} = \bigcap_{k=0}^{N-1} \{g : |y_k - g \circledast_k u| \leq \delta\} \cap \mathcal{S}.$$

Since this by now familiar set Θ depends on the input $\{u_k\}$, then so does its Chebychev centre, which for general $\{u_k\}$ can be quite difficult to compute [62].

Because of this difficulty, Jacobson and co-authors [62] and others [17] have proposed that $\{u_k\}$ should be chosen as an impulse, or a step with the observed data being differentiated before being used. In this special case they observe that the Chebychev central estimate $\{\widehat{g}_k\}$ can be simply calculated as

$$\widehat{g}_k = \frac{1}{2}(\overline{g_k} + \underline{g_k}) \tag{32}$$

where

$$\begin{aligned} \overline{g_k} &= \min(y_k + \delta, M\rho^{-k}) \\ \underline{g_k} &= \max(y_k - \delta, -M\rho^{-k}). \end{aligned}$$

and the estimation error can be bounded as

$$\sup_{\substack{g \in \mathcal{S} \\ \|\nu\|_\infty \leq \delta}} \|\widehat{g} - g\|_1 \leq \frac{M\rho}{\rho^n(\rho - 1)} + \sum_{k=0}^{N-1} \min(\delta, M\rho^{-k}).$$

If we return to our running simulation example (13) with the modifications that $\{u_k\}$ is a unit impulse, the 2 second time delay is removed, and the measurement noise is **decreased** by a factor of 50 to $\sigma_\nu^2 = 0.0001$ then the results of using the ℓ_1 estimation algorithm (32) are shown in the left hand diagram of figure (8). The results when the noise variance is set to the level we have used in the previous simulations of $\sigma_\nu^2 = 0.005$ is shown in the right hand diagram. In both diagrams we show, clockwise from top left, the true impulse response and the assumed bound $M\rho^{-k}$, the observed output data after impulse response input $\{u_k\}$ together with the true impulse response, and finally the estimate $\{\widehat{g}_k\}$ together with the bounds $\overline{g_k}, \underline{g_k}$.

At this point the reader may be questioning why we have changed the simulation example slightly from that we began with in section 3. In answer, the input has been changed since the particular ℓ_1 identification method we are currently profiling involves this experimental procedure for best performance. Remember that we consider the experiment

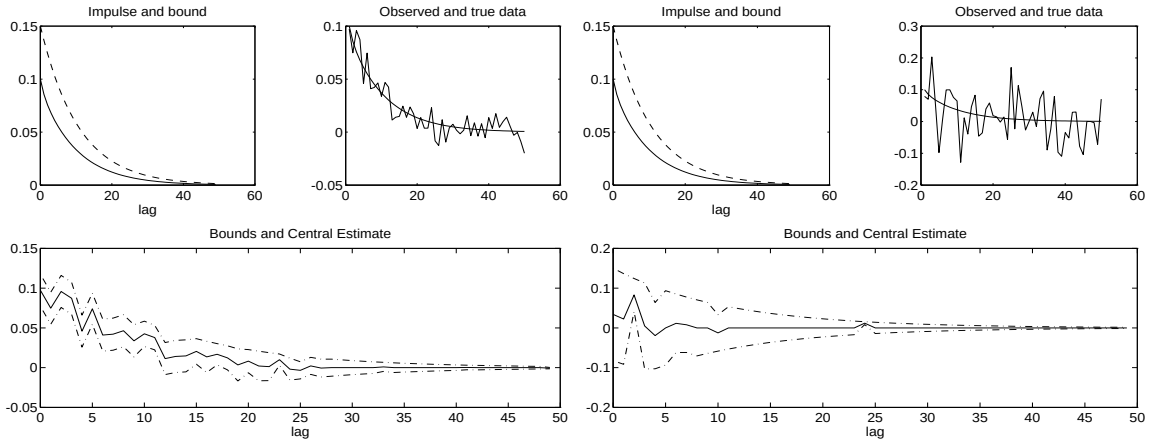

Figure 8: ℓ_1 Optimal estimates. On the left is the case of noise variance $\sigma_\nu^2 = 0.0001$, on the right is $\sigma_\nu^2 = 0.005$.

design itself to be one of features that distinguishes between the methods we are reviewing. The common denominator in our running simulation is thus not meant to be this experiment selection, rather it is meant to be the intrinsic problem of deciding how to go about extracting appropriate information about a particular plant in a particular noisy environment.

Given this, the reader will also notice that we have changed the true plant slightly by removing the time delay, and we have also performed one simulation with greatly reduced noise corruption. This was done in the interests of improving the clarity of our exposition, since the estimates we obtained were somewhat poor with the time delay and full noise in place, and we felt it was difficult for the reader to gain an impression of the nature of the method.

Poolla and Tikku [115] and others [21] have examined the genesis of the poor quality of estimates such as shown in the right hand diagram of figure 8 and have come to the discouraging conclusion that the number of data N required to ensure that the worst case ℓ_1 estimation error (31) is less that ε is exponential in $1/\varepsilon$. Poolla and Tikku attribute their result to the fact that a deterministic bounded magnitude description of noise provides a rather pessimistic viewpoint; nature is perhaps assumed to be more malicious than necessary.

7 Stochastic Embedding Methods

A motif of the work we have surveyed to this point is the premise that stochastic assumptions on disturbances are avoided in favour of deterministic bounded magnitude ones. One of the rationales for this choice is that by its nature a deterministic description encompasses stochastic assumptions and also describes the component of the residual that is produced by an undermodelling term G_Δ. Conventional stochastic formulations fail with this last

component.

The area of survey in this section, the so-called 'Stochastic Embedding' methods, are distinguished by the fact that stochastic descriptions for disturbances are not discarded. In spite of the descriptive title, neither is a stochastic description of the undermodelling induced residual $\{G_\Delta(q)u_k\}$ used. Rather, the basic idea is to use the mathematical formulation of what a random variable is in order to draw some inspiration as to how G_Δ should be described. Most succinctly, G_Δ itself is described in the same way as a random variable is, and this provides the mathematical machinery required to quantify the estimation error due to G_Δ.

To be more explicit, recall that in section 5.4 a key difficulty with using worst case estimation ideas in our control design setting is the tautological problem that the information Θ depends upon a disturbance bound δ which depends upon G_Δ which in turn depends upon Θ. This indicates that some sort of description of a class of G_Δ that is specified **relative** to some specific θ, say θ_0, needs to be provided.

This class description is the key to the 'Stochastic Embedding' work under review. To explain it we remember that the main idea in characterising ν_k as a random variable is that ν_k is a function not only of time k, but also of the state of nature. In this case the disturbance is more accurately written as $\nu_k(\lambda)$, the dependence on λ capturing the dependence on nature. Since the latter is such a nebulous thing, the dependence is not stipulated as being smooth or linear or anything so restrictive. Instead one merely assumes that the size of sets in the domain Λ of $\nu_k(\lambda)$ can be measured with some function $\mathbf{P}_\nu(\lambda)$ (called a probability measure). In this way, an event λ in nature Λ arises and gives rise to a particular disturbance realisation $\{\nu_k(\lambda)\}$.

The essential idea in stochastic estimation theory is to then relate averages of functions of $\nu_k(\lambda)$ over λ (Means and variances) with averages of a particular realisation sequence $\{\nu_k\}$ over time k. The Theorems giving these relations are known as the Law of Large Numbers and the Central Limit Theorem [23] and form the basis for stochastic estimation theory.

The core of the 'Stochastic embedding' approach to quantifying estimation errors is to use the same mathematical framework for describing $G_\Delta(q)$ as is used to describe the random variable ν_k. A range of $G_\Delta(q)$ needs to be considered, so one indexes them by writing $G_\Delta(q,\lambda)$. The nature of this dependence of $G_\Delta(q)$ on λ is not stipulated other than to say that the index parameter λ belongs to some set Λ. We then allow ourselves to concentrate attention on some G_Δ more than others by assigning how important some regions of Λ are. This is done by the specification of a measure $\mathbf{P}_\Delta(\lambda)$. The last step is to replace the more natural error quantification $|G_T(e^{j2\pi f}) - G(e^{j2\pi f}, \widehat{\theta}_N)|^2$ with an average with respect to this measure of importance

$$\int_\Lambda |G_T(e^{j2\pi f}) - G(e^{j2\pi f}, \widehat{\theta}_N)|^2 \, \mathrm{d}\mathbf{P}_\Delta(\lambda). \tag{33}$$

Now if the measure $\mathbf{P}_\Delta$ is chosen so that the size $\mathbf{P}_\Delta(\Lambda)$ of the whole index space is finite, then this formulation is equivalent to describing G_Δ as a random variable, and (33) can

be interpreted as an expected value $\mathbf{E}\left\{|G_T - G(\widehat{\theta}_N)|^2\right\}$. Hence the description 'Stochastic Embedding'.

This probabilistic interpretation is almost co-incidental, but once recognised it becomes apparent that the usual tools of stochastic analysis (Law of Large Numbers and Central Limit Theorem) can be applied to estimate (33) from averages over k of observed data sequences. Note that this formulation does not imply that the manifestation $\{G_\Delta(q)u_k\}$ of G_Δ is a stochastic process. On the data set we have at hand λ and hence $G_\Delta(q,\lambda)$ is fixed.

In a sense, the 'Stochastic Embedding' approach can be regarded as an attempt to extend stochastic theory to handle undermodelling rather than rejecting it and replacing it with a deterministic framework. It was first formulated in [38, 39, 36, 35] and later extended in [34, 98, 99]. An obvious disadvantage of the approach is that it does not provide 'hard' bounds on model uncertainty. Rather it aims at providing regions of model confidence.

To use these 'Stochastic Embedding' ideas, the first thing we need to do is specify $\mathbf{P}_\Delta$. This is most easily done by using an FIR description for $G_\Delta(q,\lambda)$

$$G_\Delta(q,\lambda) = \sum_{m=0}^{L-1} \eta_m(\lambda) q^{-m}$$

so that $\mathbf{P}_\Delta$ may be specified by assuming it corresponds to a Gaussian density function such that the impulse response of an 'average' G_Δ is captured in an exponential envelope:

$$\begin{aligned}\mathbf{E}\{\eta_m\} = \int_\Lambda \eta_k(\lambda)\,\mathrm{d}\mathbf{P}_\Delta(\lambda) &= 0\\ \mathbf{E}\left\{\eta_m^2\right\} = \int_\Lambda \eta_k^2(\lambda)\,\mathrm{d}\mathbf{P}_\Delta(\lambda) &= M\rho^{-k}.\end{aligned}$$

The translation of this prior information about G_Δ into an error quantification of the form (33) is most easily done by once again using the vectorised notation (6) with the extensions

$$\eta^T \triangleq [\eta_0, \cdots, \eta_{L-1}], \qquad [\Psi]_{m,n} \triangleq u_{m-n}$$

so that the least squares estimate $\widehat{\theta}_N$ in (5) can be written as

$$\widehat{\theta}_N = \theta_0 + P_N \Psi \eta + P_N V.$$

If $\{\nu_k\}$ can also be modelled as a Gaussian distributed process then this gives the error in $\widehat{\theta}_N$ averaged over important (as measured by $\mathbf{P}_\Delta$) undermodellings and averaged over feasible disturbance realisations as

$$\mathbf{E}\left\{(\widehat{\theta}_N - \theta_0),(\widehat{\theta}_N - \theta_0)^T\right\} = P_N \Phi^T (\Psi C_\eta \Psi^T + C_\nu)\Phi P_N, \qquad C_\eta \triangleq M \operatorname*{diag}_{0\le k\le L-1}\left\{\rho^{-k}\right\}. \quad (34)$$

In the specification of C_η we have assumed $\{\eta_k\}$ to be an uncorrelated process. This implies an assumption that averaged over possible undermodellings the magnitude frequency response of G_Δ is constant

$$\int_\Lambda |G_\Delta(e^{j2\pi f},\lambda)|^2\,\mathrm{d}\mathbf{P}_\Delta(\lambda) = \frac{M\rho}{\rho - 1}.$$

If instead one assumes a correlated structure for $\{\eta_k\}$ such as

$$\mathbf{E}\{\eta_m\eta_n\} = \int_\Lambda \eta_m(\lambda)\eta_n(\lambda)\,\mathrm{d}\mathbf{P}_\Delta(\lambda) = \begin{cases} \dfrac{M}{(\rho\beta^2-1)}\left[\dfrac{\beta^{n+m}}{\beta^2} - \dfrac{\rho\beta^n}{(\rho\beta)^m}\right] & ; m \le n \\ \dfrac{M}{(\rho\beta^2-1)}\left[\dfrac{\beta^{n+m}}{\beta^2} - \dfrac{\rho\beta^m}{(\rho\beta)^n}\right] & ; m > n \end{cases}$$

then this corresponds to

$$\int_\Omega \left|G_\Delta(e^{j2\pi f},\omega)\right|^2 \mathrm{d}\mathbf{P}_\Delta(\omega) = \left(\frac{M\rho}{\rho-1}\right)\frac{1}{|1-\beta e^{-j2\pi f}|^2}$$

which specifies more about the high frequency behavior of G_Δ at the expense of having to supply a value for the extra parameter β. The final ingredient, once C_η is settled on, is to transfer the parameter space error quantification (34) into the frequency domain quantification (33). Again, this is best done with a vectorised notation. Since G_Δ is specified relative to some θ_0, then

$$G_T(e^{j2\pi f}) = G(e^{j2\pi f},\theta_0) + G_\Delta(e^{j2\pi f},\lambda) = \Gamma(f)\theta_0 + \Pi(f)\eta$$

where

$$\Pi(f) \triangleq [1, e^{-j2\pi f}, \cdots, e^{-j(L-1)2\pi f}].$$

If we then re-define $\Omega(f)$ as

$$\Omega(f) \triangleq \begin{bmatrix} \mathrm{Re}\{\Gamma(f)\}, & \mathrm{Re}\{\Pi(f)\} \\ \mathrm{Im}\{\Gamma(f)\}, & \mathrm{Im}\{\Pi(f)\} \end{bmatrix}$$

then with the final assumption that $\{\eta_k\}$ is not correlated with $\{\nu_k\}$ we can calculate the average frequency response error (we use (10) to define g) as

$$\mathbf{E}\left\{g(f)g(f)^T\right\} = \Omega(f)\Upsilon\Omega^T(f), \qquad \Upsilon \triangleq \begin{bmatrix} P_N\Phi^T(\Psi C_\eta\Psi^T + C_\nu)\Phi P_N & -P_N\Phi^T\Psi C_\eta \\ -C_\eta\Psi^T\Phi^T P_N & C_\eta \end{bmatrix}$$

in which case (33) becomes Trace$\{\Omega(f)\Upsilon\Omega^T(f)\}$. However, if the measures $\mathbf{P}_\nu$ and $\mathbf{P}_\Delta$ are chosen to correspond to Gaussian density functions then

$$g(f)^T[\Omega(f)\Upsilon\Omega(f)^T]^{-1}g(f) \tag{35}$$

will be χ^2 distributed with two degrees of freedom. This allows us to draw confidence ellipsoids on Nyquist plots which give an indication of phase as well as magnitude error. The interior of these ellipsoids represent possible responses for $G_T(e^{j2\pi f})$ that are valid save for a few unimportant (with respect to the measure $\mathbf{P}_\Delta$) possibilities for $G_\Delta(e^{j2\pi f})$.

An important criticism of the stochastic embedding methods is that it is difficult to provide a physical interpretation of the meaning of the bounds beyond that we have just

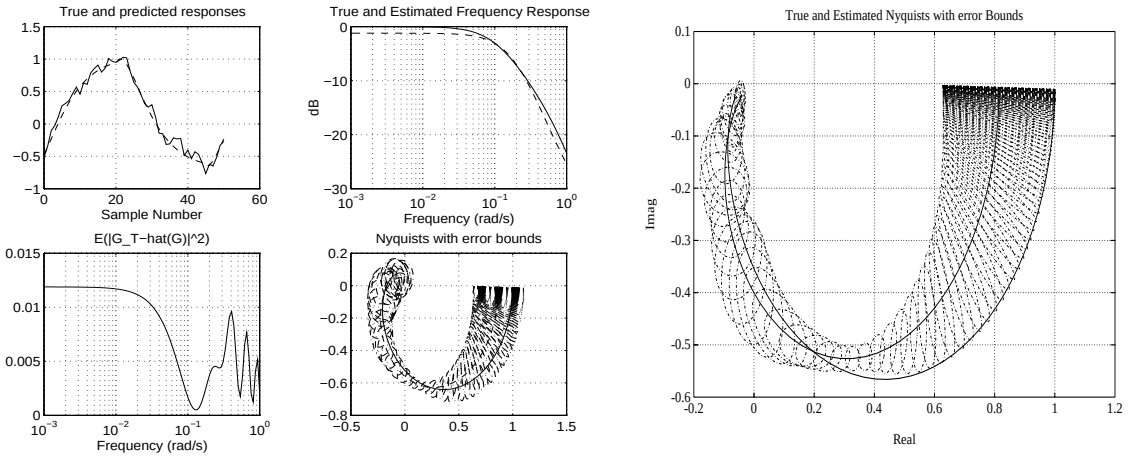

Figure 9: The use of the 'Stochastic embedding' method on our running simulation example. On the left, we use M and ρ derived from prior information. On the right we estimate M and ρ from the available data.

given. To many workers, this is far less appealing that the unequivocal idea of having a hard bound on errors. Perhaps due to this, the stochastic embedding ideas have not been subject to widespread acceptance, although authors other than those originating the ideas have taken them up [5, 129].

The use of this stochastic embedding method on our running simulation example is illustrated in figure 9. Again, we use the short data record/high noise case where we believe the bias/variance tradeoff precludes us from using 'standard' solutions. Clockwise from the top left of the left hand diagram in figure 9 we show the true and estimated system response, the true and estimated frequency responses, the true and estimated Nyquists with error ellipsoids calculated by assuming (35) will have a χ^2 distribution and finally, the estimated magnitude of the modelling error using expression (33).

In these latter two error quantification diagrams we used the values $M = 0.16$ and $\rho = 1.67$ that we previously used in the worst case simulation example. On the Nyquist diagram we have plotted 90% 'confidence' regions, which corresponds to plotting the locus of (35) equal to 4.61. Although these bounds capture the nature of the error, they are not particularly tight. Furthermore, the choice of 90% confidence regions rather than say 80% or 95% regions seems rather arbitrary. This is due to the rather nebulous issue of trying to provide a physical interpretation for these stochastically embedded derived bounds. They are best understood as being more of a guide to the errors that can expected. This sort of information is not in keeping with the demands of robust control theorists.

One aspect of the 'Stochastic Embedding' approach that we have not commented on is that it is possible to estimate the parameters M, ρ and the noise variance σ_ν^2 from the observed data rather than try to specify them from prior knowledge. This is achieved by calculating the likelihood function for $\{y_k\}$ conditional upon M,ρ and σ_ν^2 and then taking estimates $\widehat{M}$,$\widehat{\rho}$ and $\widehat{\sigma}_\nu^2$ as the values maximising the likelihood [34, 99]. For this simulation

example we obtain the estimates

$$\widehat{M} = 0.217, \quad \hat{\rho} = 2.05, \quad \hat{\sigma}_\nu^2 = 0.0061.$$

Using these estimated values we obtain the error quantification shown in the right hand diagram of figure 9. It is possible to show that under certain conditions these maximum likelihood estimates converge to the true values, although balanced against this is that they do so slowly; $\widehat{M}$, $\hat{\rho}$ and $\hat{\sigma}_\nu^2$ converge like $1/\log N$,$1/\log^3 N$ and $1/N$ respectively [99]. In some applications, this may imply that the estimates have such a large variance as to be unusable.

8 Identification from Frequency Response Data

The final area of work we wish to survey involve schemes that start from frequency domain measurements and collectively have become known as 'H_∞ identification methods'. This name derives from an aim of finding a model whose frequency response is in H_∞ together with a non-probabilistic bound on the error in this model. The line of research was begun by Parker and Bitmead [104, 103] in 1987. By and large, the multitude of papers [51, 52, 53, 54, 56, 55, 61, 108, 87, 107, 41, 42, 48] following have concentrated mainly on improving the asymptotic in model order properties of Parker and Bitmead's original ideas through the use of windowing functions and two step algorithms. However, in recent times fundamentally different schemes have started to arise [16, 15, 25, 33, 44, 48, 88, 110, 143, 146] that do not follow this prescription.

The original problem attacked by Parker and Bitmead has been stated in canonical form by Helmicki, Jacobson and Nett in [51]:

- Assume the true system transfer function $G_T(z)$, but evaluated at $z = z^{-1}$ is analytic on the domain $\mathbf{D}_\rho = \{z \in \mathbf{C} : |z| < \rho\}$,$\rho > 1$ and is bounded in magnitude by M on $\mathbf{D}_\rho$. A common shorthand notation for this is $G_T \in H_\infty(\mathbf{D}_\rho, M) \triangleq \mathcal{S}$

- Assume we have available to us n measurements of the frequency response of G_T at the n roots of unity. These measurements $\{f_0, \cdots, f_{n-1}\}$ are also corrupted by uncertainty as follows:

$$f_k = G_T(e^{-j2\pi k/n}) + \nu_k \tag{36}$$

  where

$$|\nu_k| \leq \varepsilon. \tag{37}$$

  That is, the corruption is by the components of some element $\nu \in \ell_\infty$.

Under these conditions, the goal is to derive an estimate $\widehat{G}(z)$ such that $\widehat{G} \in H_\infty(\mathbf{D}_1)$ and

$$\lim_{\substack{\varepsilon \to 0 \\ n \to \infty}} \operatorname*{Sup}_{\substack{G_T \in \mathcal{S} \\ \|\nu\|_\infty \leq \varepsilon}} \left\| G_T - \widehat{G} \right\|_\infty = 0. \tag{38}$$

The common approach to this problem in the H_∞ identification literature, see for example [51, 52, 53, 61, 108, 87, 41, 42] is as follows. Select the model to be FIR of the form

$$\widehat{G}(z) = \sum_{k=0}^{d-1} \widehat{g}_k z^k \tag{39}$$

and try to make its frequency response the same as the measurements:

$$\sum_{k=0}^{d-1} \widehat{g}_k e^{-j2\pi mk/n} = f_m \quad ; m = 0, \cdots, n-1. \tag{40}$$

Now originally Parker and Bitmead only considered the case of $d = n$. However, if the choice of the model order is $d < n$ then the equations (40) are overdetermined and a natural choice for $\widehat{g}_k$ would be one that minimised the total squared error in (40). This solution is most easily seen by vectorising:

$$\widehat{\theta}_n^T = [\widehat{g}_0, \cdots, \widehat{g}_{d-1}], \tag{41}$$

$$\Omega_n = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & e^{-j2\pi/n} & e^{-j4\pi/n} & \cdots & e^{-j(d-1)2\pi/n} \\ 1 & e^{-j4\pi/n} & e^{-j8\pi/n} & \cdots & e^{-j2(d-1)2\pi/n} \\ \vdots & & & & \vdots \\ 1 & e^{-j(n-1)2\pi/n} & e^{-j2(n-1)2\pi/n} & \cdots & e^{-j(n-1)(d-1)2\pi/n} \end{bmatrix}, \tag{42}$$

$$F^T = [f_0, \cdots, f_{n-1}], \tag{43}$$

so that (40) becomes

$$\Omega_n \widehat{\theta}_n = F \tag{44}$$

with solution minimising the squared error as:

$$\widehat{\theta}_n = (\Omega_n^\star \Omega_n)^{-1} \Omega_n^\star = \frac{1}{n} \Omega_n^\star F. \tag{45}$$

to give

$$\widehat{g}_m = \frac{1}{n} \sum_{k=0}^{n-1} f_k e^{j2\pi mk/n} \quad ; m = 0, \cdots, d-1. \tag{46}$$

That is, the model co-efficients are found by taking the inverse DFT of the frequency measurements. Note that with this solution there is no guarantee that the $\widehat{g}_m$ are real, as we would like. However, as Gu and Khargonekar suggest [41], realness of $\widehat{g}_m$ can be ensured by ensuring f_0 is real (as it should be since it represents a d.c. response) and then setting

$$f_k = f_{n-k}^\star \tag{47}$$

which means we should in fact conduct experiments at only $n/2$ frequencies and form the other $n/2$ by conjugation.

Substituting (46) back into (39) gives the relationship between the measurements $\{f_0, \cdots, f_{n-1}\}$ and the final model when $d = n$:

$$\widehat{G}(z) = \frac{1}{n}\sum_{k=0}^{n-1}\sum_{r=0}^{n-1} f_r e^{j2\pi kr/n} z^k = \frac{1}{n}\sum_{r=0}^{n-1} f_r \sum_{k=0}^{n-1}\left[e^{j2\pi r/n} z\right]^k = \frac{1}{n}\sum_{r=0}^{n-1} f_r \left(\frac{z^n - 1}{z e^{j2\pi r/n} - 1}\right). \quad (48)$$

Parker and Bitmead note that there is a parallel here with signal processing literature where the filters

$$\frac{1}{n}\left(\frac{z^n - 1}{z e^{jr\omega_s} - 1}\right) \quad (49)$$

are known as 'frequency sampling' filters because of their property

$$\frac{1}{n}\left(\frac{z^n - 1}{z e^{jr\omega_s} - 1}\right) = \begin{cases} 1 & ; z = e^{-jr\omega_s} \\ 0 & ; z = e^{-j\ell\omega_s} \quad \ell \neq r. \end{cases} \quad (50)$$

Therefore, the final estimate $\widehat{G}(z)$ matches the measured data at the n roots of unity and then interpolates between them using an nth order polynomial which is the filter (49).

The final ingredient originally examined by Parker and Bitmead was the fact that the measurements we have available are corrupted according to (36). They arrive at the following bound on the estimation error

$$e_n(\varepsilon) \triangleq \sup_{\substack{G_T \in \mathcal{S} \\ \|\nu\|_\infty \leq \varepsilon}} \left\| G_T - \widehat{G} \right\|_\infty \leq \underbrace{\frac{2M\rho}{\rho^n(\rho - 1)}}_{\text{Bias Error}} + \underbrace{\varepsilon\left(1 + \frac{2}{\pi}\ln n\right)}_{\text{Variance Error}} \quad (51)$$

which is only proved true by Parker in [103] when n is some power of 2. The first term in the bound (51) is due to the undermodelling in representing $G_T(z)$ with an nth order model. The second term in the bound (51) is due to the corruption in the measurements via ν. The reason this term grows with n is that there is no Lipschitz smoothness guarantee on the observed frequencies. That is, they can differ in magnitude by ε no matter how closely together the measuring frequencies are. Therefore, any smooth function interpolating this arbitrarily non-smooth one must have 'overshoot' between the interpolating points.

A simpler explanation of the divergence phenomenon is that, from (48), the estimate is formed as a convolution between the measurements $\{f_r\}$ and an interpolating function. The magnitude of this interpolating function is $|(\sin \omega n/2)^{-1} \sin \omega/2|$, and the sum of the magnitudes of the maxima of this latter function grows with n at the rate $\ln n$. It is therefore possible to find a particular bounded disturbance which gives an estimation error that grows like $\ln n$.

A great deal of the work following Parker and Bitmead's original contributions has placed great importance on dealing with the noise induced $O(\ln n)$ divergence in (51). For example, Helmicki, Jacobson and Nett in [51, 53] rederive the bound (51) but their $\ln n$ bound applies only asymptotically in n and has an unspecified constant term [3]. In [42]

[3] Unfortunately their asymptotic bound is incorrect in a minor way since, as pointed out by Brutman [13], the result of Gronwall [40] that Helmicki et al. use is in error due to a mistake in Gronwall's proof. According to Brutman, the correct asymptotic bound for the noise term is $\frac{2}{\pi}\left(\ln n + \gamma + \ln \frac{8}{\pi}\right) + o(1)$ as $n \to \infty$ where γ is Euler's constant ≈ 0.57722

Gu and Khargonekar apply a triangular window to the $\widehat{g}_m$ calculated by (46). That is, the estimated model they propose is[4]

$$\widehat{G}(z) = \sum_{k=0}^{n-1} \tau_k \widehat{g}_k z^k \tag{52}$$

$$\tau_k = 1 - \frac{k}{n} \tag{53}$$

The noise bounding term they subsequently obtain depends on an experimentally determined constant that makes their bound slightly tighter than Parker and Bitmead's, but still $O(\ln n)$ as $n \to \infty$.

The reason for the interest in the $\ln n$ growth in error is that the goal (38) is not achieved unless one can know and decrease ε at a rate greater than $\ln n$. On the other hand, two step algorithms that operate non-linearly on the available data and were first found by Helmicki, Jacobson and Nett [51, 56] exist that achieve (38) without prior knowledge of ρ or M. Such algorithms were defined in [51] to be 'robustly' convergent. If knowledge of ε is also not required then the algorithm is defined in [51] to be 'untuned'.

Of course, the fact that overbounds such as (51) diverge does not imply that the true estimation error diverges. However, the problem of divergence of Lagrange interpolants, of which the Parker and Bitmead scheme is an example, does exist and has a long history in the mathematics literature dating back to the 1910's with Fejér [28] showing that divergence does in fact occur at the rate of $\ln n$ even if a continuous function is being interpolated. This divergence was commented on by Erdös [27] as being '*in contrary to everything what was expected since NEWTON (sic)*'. Erdös went on to conjecture that it was optimal with respect to sup norm to interpolate at evenly spaced points [29]. This conjecture was finally proved true for the case of interest to us by Brutman et.al. [12, 13, 14]. Unfortunately, as pointed out by Akçay et.al. [2, 109] it may, in practice, be difficult to obtain uniformly spaced frequency response measurements.

Of course if $d < n$ in (39) then we no longer have an interpolatory scheme which raises the hope that the noise divergence may be avoided. However, it is still an algorithm that is linear in the data. Furthermore in [51] it was conjectured that no linear in the data robustly convergent algorithm existed. One year later [107, 108] Partington proved this was so.

The first workers to derive a robustly convergent algorithm where Helmicki, Jacobson and Nett in [52, 56]. They avoided divergent behavior by using a low order (namely linear), interpolating polynomial between the frequency response measurement in order to impose a smoothness constraint on the estimate. They then noticed, that if one wanted to equate the impulse response of an Nth order FIR model to the Fourier co-efficients of the spline interpolant, then these impulse response co-efficients could be obtained in a computationally efficient way by performing an inverse DFT on the point frequency

[4]Note that this scheme is motivated by classical works [60, 145] on recovering a function from its Fourier co-efficients via Cèsaro means in order to guarantee uniform convergence on $C([-\pi, \pi])$

responses and then windowing as in (52) with the sequence $\{\tau_k\}$ given by

$$\tau_k = \left(\frac{n}{\pi k}\right)^2 \left(\sin\frac{\pi k}{n}\right)^2 . \tag{54}$$

Although this spline based smoothing opeation avoids noise induced divergence, a concomitant problem is that the spline interpolant usually will not have a frequency response that corresponds to the evaluation on the unit circle of some $\widehat{G} \in H_\infty$. In this case, the FIR model obtained is non-causal.

Helmicki, Jacobson and Nett overcame this latter problem in [52, 56] by obtaining the final estimate $\widehat{G}(z)$ as the element in H_∞ closest in L_∞ norm to the L_∞ system

$$\widetilde{G_N}(z) = \sum_{k=-N}^{N} \tau_k \widehat{g}_k z^k \tag{55}$$

with $\widehat{g}_k$ given by (46) and with $d = n$. In [52, 56] it was shown that this does result in a robustly convergent algorithm provided

$$\lim_{n\to\infty} \frac{n^2}{N} = 0 \tag{56}$$

which suggests a model order of $N = n^3$. Later in [41, 42] Gu and Khargonekar show that Helmicki, Jacobson and Nett were very conservative in their calculations. In fact $N \approx n/2$ is sufficient to guarantee robust convergence.

This two stage non-linear in the data scheme that Helmicki, Jacobson and Nett invented can be represented as

$$\underbrace{\text{Data} \longrightarrow L_\infty}_{\substack{\text{Reconstruction of} \\ \text{function as Fourier Series}}} \qquad \underbrace{\longrightarrow H_\infty.}_{\substack{\text{Nonlinear Step} \\ \text{Nehari extension}}} \tag{57}$$

Until recently, the majority of the 'robustly convergent' schemes that have been developed since this invention can be represented in this two stage way and differ only in the choice of the windowing sequence $\{\tau_k\}$ used in (55) which is the first step of (57) . Gu and Khargonekar have examined both time domain [41] and frequency domain [43] criteria that the sequence $\{\tau_k\}$ should satisfy in order for robust convergence to hold. They also find that for any window sequence there is a tradeoff between the size of error term due to noise and the size of the error term due to truncated model order; the triangular window sequence (53) achieves the smallest noise induced error component and that the cosine sequence

$$\tau_k = \cos\frac{k\pi}{2n+1} \tag{58}$$

is favourable at minimising the error component due to truncation errors. This is the bias/variance tradeoff manifesting itself again. Rubin and Limebeer [116] have conducted a study of how the choice of weighting sequence $\{\tau_k\}$ affects the bias/variance tradeoff.

In [41] Gu and Khargonekar note that the choice of $\{\tau_k\}$ in (53) when used in (55) becomes $\tau_k = 1 - |k|/N$ and corresponds to approximation by so called 'Jackson' polynomials that arise in the classical Fourier series analysis invented by Fejér of using Césaro mean reconstruction [147]. In [107] Partington studies approximations which are in turn derived from Jackson polynomials in such a way that

$$\widetilde{G_N}(z) = \frac{2n+1}{n} \sum_{k=-N}^{N} \tau_k \bar{g}_k z^k - \frac{n+1}{n} \sum_{k=-N}^{N} \tau_k \hat{g}_k z^k \tag{59}$$

with $\hat{g}_k$ as in (46) and $\bar{g}_k$ as in (46) with the substitution $n = 2n$. This corresponds to approximation by so called De La Vallée Poussin polynomials that again originally arose in classical Fourier Series analysis at the beginning of this century. It is complicated to write out the choice of $\{\tau_k\}$ that corresponds to this De La Vallée Poussin choice, but Partington [107] shows that it enjoys the utility of providing the fastest possible (exponential) rate of convergence.

This completes a review of the so-called 'two step' methods that form the bulk of the work in the H_∞ identification literature. However, we should point out that recently some new work in the area has appeared which departs significantly from the more common two-step, FIR model structure approach. For example, the theory presented in [110, 88, 25] pertains to far more general model structures than FIR. As well, the work in [110] suggests estimation methods based on linear programming rather than the DFT/Nehari extension method we have just reviewed. The work in [143] suggests a convex programming approach, and the work in [16] suggests finding an interpolatory FIR model by relating to the classical Carathéodory-Fejér problem [4] which is better known in the spectrum estimation and time series prediction field, where well known methods such as Schur recursions are available to solve it [50].

Closely aligned with this is the recent work in [44, 18, 15] which involves interpolating a rational model, rather than a polynomial FIR model to the observed data via. This is achieved by constructing the model using Blaschke products according to the Nevanlinna-Pick theory [4]. Finally, the efforts in [33, 146, 48] show how these H_∞ methods may be extended to work directly from time domain, rather than frequency domain data. We refer the reader to [89] in this volume for expert comment on these newer ideas.

8.1 Choice of Assumed Prior Knowledge

In almost all of the two-step H_∞ identification work we have just surveyed the bounds on the estimation errors are derived from prior information of the form $G_T \in H_\infty(\mathbf{D}_\rho, M) \triangleq \mathcal{S}$. This is achieved by using ρ and M parameterising $\mathcal{S}$ to derive estimation error bounds on the magnitude of the impulse response $\{g_k\}$ of $G_T(z) = \sum_{k=0}^{\infty} g_k z^k$ via Cauchy's estimate [26].

It is important to note that Cauchy's estimate is **very** loose, and hence the estimation error bounds can be very conservative. To see this note that the bound on $|g_k|$ is derived

by appeal to Cauchy's estimate as follows. By Parseval's Theorem $\forall r : |r| < \rho$ and $\forall n \geq 0$

$$|g_n|^2 r^{2n} \leq \sum_{k=0}^{\infty} |g_k|^2 r^{2k} = \frac{1}{2\pi} \int_{-\pi}^{\pi} |G_T(re^{j\theta})|^2 \, \mathrm{d}\theta \leq M^2. \tag{60}$$

This leads to

$$|g_n| \leq \frac{M}{\rho^n}. \tag{61}$$

However, the first inequality in (60) is obviously grossly conservative since we have underbounded an infinite sum of positive terms by just **one** term in the sum. There is also conservatism in the bound on the right hand side of (60). We should therefore expect (61) to be correspondingly highly conservative.

These considerations lead us to suggest that one should try to bound the impulse response directly as in (61) rather than try to infer it from the frequency domain assumption of $G_T \in H_\infty(\mathbf{D}_\rho, M)$.

8.2 Simulation Example

To finish our survey of the H_∞ methods we continue with our running simulation example. As the preceding sections have surveyed, there are many possibilities for the specific choice of algorithm to use in this profile by simulation exercise. However, the majority of them suggest an experiment design involving sine wave inputs. We accommodate this by conducting a simulation on our example plant of section 3 in which $n = 20$ sine wave experiments are conducted, each of a 500 sample duration, and where the measurements are noise corrupted as in section 3. This involves the collection of a total of 10 000 data points. After each sine wave experiment, we accept the suggestion of Helmicki, Jacobson and Nett in [56] and estimate the system frequency response by taking the DFT of the observed output signal $\{y_k\}$. That is, we take f_k in (36) as

$$f_k = \frac{1}{500} \sum_{m=0}^{499} y_k e^{-j2\pi mk/n}$$

According to Theorem 3.3 of [56], since the noise corruption in our simulations can be bounded in magnitude by 0.1, this leads to a bound ε for the error in our frequency response estimate of (see below for how M and ρ are chosen)

$$\begin{aligned} |\nu_k| &\leq \frac{M\rho}{(\rho-1)^2}\left(\frac{1-\rho^{-500}}{500}\right) + 0.1 \qquad ; \rho = 1.1, M = 0.1 \\ &= 0.144. \end{aligned}$$

We point out that this is a very different experiment set up to previous sections where various other estimation schemes have been profiled. The explanation for this seeming incongruity is that our profiling exercise is meant to permit a concrete comparison at the level of solving an intrinsic problem of learning about a particular plant from noise

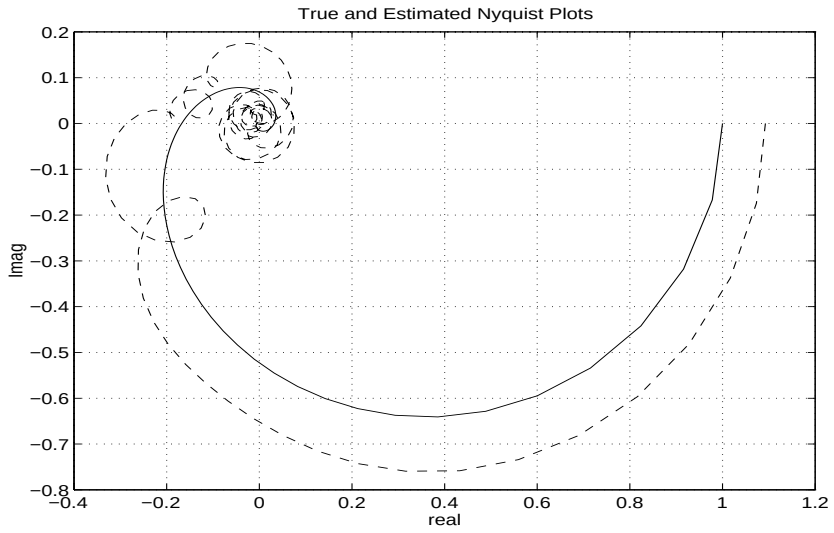

Figure 10: True and estimated Nyquist plots. Solid line is true system. Dashed line is estimate. The left the estimate is formed by finding a 40th order interpolant to 20 frequency domain measurements which required 10000 time domain measurements to form.

corrupted observations, and this involves an experiment design choice in addition to an algorithm design choice.

The simulation studies we present are not meant to profile how various schemes perform on identical data sets. If for no other reason we adopt this approach since it would be impossible due to the richness and diversity of the ideas in the area we are surveying leading to schemes that *ab-initio* assume different experiment designs.

With these comments in mind, we proceed by choosing one among the many H_∞ identification schemes possible. Our choice is the simplest possible, the early linear interpolation approach of Parker and Bitmead [104, 103] that we gave details of in (39)-(46). The results for a particular noise realisation are shown in figure 10 where the Nyquist diagram for the true system is shown as a solid line, and the Nyquist plot of the Lagrange interpolant to the 40 frequency points (20 formed by conjugation as per (47) is shown as a dashed line. This estimate is not particularly accurate, but possesses the feature the when we use the prior information that the impulse response $\{g_k\}$ of $G_T(q)$ can be bounded as $|g_k| \leq 0.2/1.1^k$ then using (51) it is possible to provide a guaranteed error bound on this estimate of

$$\sup_{\substack{G_T \in \mathcal{S} \\ \|\nu\|_\infty \leq \varepsilon}} \left\| G_T - \widehat{G} \right\|_\infty \leq 0.5214.$$

Note that in line with the previous section, we make this bound as tight as possible by noting that the M and ρ that define $\mathcal{S} = H_\infty(\mathbf{D}_\rho, M)$ provide an error bound purely through bounding the impulse response, and if this impulse response can be bounded directly, rather then through Cauchy's estimate, then more accurate answers are obtained. If we do not do this, and instead elect to use Cauchy's estimate, then since $G_T(z^{-1})$ is

analytic in $\mathbf{D}_\rho$ for the choice $\rho = 1.1$ we arrive at a bound $|g_k| \leq M/1.1^k$ with

$$M = \operatorname*{Sup}_{\omega}(|G_T(z)|_{z=1.1e^{j\omega}}) \approx 23.$$

It is possible to improve the estimate shown in figure 10 by progressing to the more sophisticated two-step schemes we have surveyed, but we do not present the results here since we believe the simple example as it stands adequately conveys the flavour of the methods for our review purposes.

9 Conclusions

Robust control synthesis requires not only a nominal model, but also a quantification of the uncertainty in that model. System identification from a 'classical' stochastic viewpoint provides a nominal model, but has difficulty in also providing an estimate of the error in that model. This gap has been the motivation for the new system identification ideas we have surveyed. Mostly, these new ideas have rejected stochastic disturbance descriptions in favour of deterministic ones since this results in an error quantification that applies with probability one, and this is in keeping with the operator theoretic flavour of robust control design. Unfortunately, such features as conservatism of bounds, sensitivity of bounds to accuracy of prior information, and difficulty in providing prior information are consequences of this approach.

The 'stochastic embedding' method we surveyed does not follow the deterministic analysis path, and does not suffer from the same problems. However it doesn't adequately answer the error quantification question either. Firstly, it is difficult to attach a physical significance to the bounds it provides since they are averages over a hypothetical ensemble of systems. Many engineers may find this as nebulous and objectionable as the random variable idea of looking at averages over imaginary ensembles of disturbance sequences. Secondly, although the stochastic embedding idea allows information about undermodelling (M and ρ) to be estimated from available data, these estimates can have high variability.

Although significant progress has been made by the many workers addressing this error quantification question, there are still aspects of it that remain open and challenging. In pursuing it further, we would suggest that a key idea must be to focus on solving problems that cannot already be solved straightforwardly by 'classical' techniques. Problems in which the bias/variance tradeoff forces us to choose a model structure so parsimonious that the stochastic process model for residuals is not appropriate are one such example.

References

[1] *Special issue on system identification for robust control design*, IEEE Transactions on Automatic Control, 37 (1992).

[2] H. Akçay, G.Gu, and P.P.Khargonekar, *Identification in H_∞ with non-uniformly spaced frequency response measurements*, in Proceedings of the American Control Conference, IEEE, 1992, pp. 246–250.

[3] H. Akçay and P. Khargonekar, *The lease squares algorithms, parametric system identification and bounded noise*, Automatica, 29 (1993), pp. 1535–1540.

[4] N. Akhiezer, *The Classical Moment Problem*, University Mathematical Monographs, Oliver and Boyd, Edinburgh, 1965.

[5] P. Andersen, S. Tøffner-Clausen, and T. Pedersen, *Estimation of frequency domain model uncertainties with application to robust controller design*, in Proceedings of the 10th IFAC Symposium on System Identification, 1994, pp. 603–608.

[6] E. Bai, *Adaptive quantification of model uncertainties by rational approximation*, IEEE Transactions on Automatic Control, AC-36 No.4 (1991), pp. 441–453.

[7] E. Bai and S. Raman, *A linearly robustly convergent interpolatory algorithm for system identification.*, in Proceedings of the American Control Conference., IEEE, 1992, pp. 3165–3169.

[8] G. Belforte and T. Tay, *Optimal input design for worst-case system identification in $\ell_1/\ell_2/\ell_\infty$*, Systems and Control Letters, 20 (1993), pp. 273–278.

[9] R. Bitmead, *Iterative control design approaches*, in Preprints of the 12th IFAC World Congress, Sydney, 1993, pp. 381–384, volume 9.

[10] R. Bitmead, M. Gevers, and V. Werz, *Adaptive Optimal Control, The Thinking Man's GPC.*, Prentice Hall, Englewood Cliffs, New Jersey, 1990.

[11] R. Bitmead and Z. Zang, *An iterative identification and control strategy*, in Proceedings of the European Control Conference, 1991, pp. 1395–1400.

[12] C. D. Boor and A.Pinkus, *Proof of the conjectures of Berstein and Erdös concerning the optimal nodes for polynomial interplation*, Journal of Approximation Theory, 24 (1978), pp. 289–303.

[13] L. Brutman, *On the polynomial and rational projections in the complex plane*, SIAM Journal of Numerical Analysis, 17 (1980), pp. 366–372.

[14] L. Brutman and A. Pinkus, *On the Erdös conjecture concerning minimal norm interpolation on the unit circle*, SIAM Journal on Numerical Analysis, 17 (1980), pp. 373–375.

[15] J. Chen, G. Gu, and C. Nett, *Worst case identification of continuous time systems via interpolation*, in Proceedings of the American control conference, San Francisco, 1993, pp. 1544–1548.

[16] J. Chen and C. Nett, *The Carathéodory-Fejér problem and H_∞ identification: A time domain approach*, in Proceedings of the 32nd Conference on Decision and Control, San Antonio Texas, 1993.

[17] J. Chen, C. Nett, and M. Fan, *Optimal non-parametric system identification from arbitrary corrupt finite time series: A control-oriented approach*, in Proceedings of the American Control Conference, Chicago, 1992, pp. 279–285.

[18] J. Chen, C. Nett, and M. Fan, *Worst case system identification in H_∞:Validation of apriori information, essentially optimal algorithms, and error bounds*, Proceedings of the American Control Conference, (1992), pp. 251–257.

[19] P. Combettes, *The foundations of set theoretic estimation*, Proceedings of the IEEE, 81 (1993), pp. 182–208.

[20] M. Dahleh and M. Khammash, *Controller design for plants with structured uncertainty*, Automatica, 29 (1993), pp. 37–56.

[21] M. Dahleh, T. Theodosopoulos, and J. Tsitsiklis, *The sample complexity of worst-case identification of fir linear systems*, Systems and Control Letters, 20 (1993), pp. 157–166.

[22] J. Deller, *Set membership identification in digital signal processing*, Acoustics Speech and Signal and Signal Processing Magazine, 6 (1990), pp. 4–20.

[23] J. Doob, *Stochastic Processes*, John Wiley and Sons, London, 1953.

[24] J. Doyle, B. Francis, and A. Tannenbaum, *Feedback Control Theory*, Macmillan Publishing Company, New York, 1992.

[25] N. Dudley and J. Partington, *Robust identification in the disc algebra using rational wavelets and orthonormal basis functions*, preprint, University of Leeds, 1994.

[26] E.C.Titchmarsh, *The Theory of Functions*, Oxford University Press, second ed., 1932.

[27] P. Erdös, *An extremal problem in the theory of interpolation*, Acta Mathematica Hungarica, 12 (1961), pp. 222–234.

[28] ——, *Problems and results on the theory of interpolation II*, Acta Mathematica Hungarica, 12 (1961), pp. 235–244.

[29] ——, *Problems and results on the convergence and divergence properties of the Lagrange interpolations polynomials and some extremal problems*, Mathematica(Cluj.), 10 (1968), pp. 65–73.

[30] E. Fogel, *System identification via membership set constraints with energy constrained noise*, IEEE Transactions on Automatic Control, AC-24, No 5 (1979), pp. 615–622.

[31] E. Fogel and Y. Huang, *On the value of information in system identification-bounded noise case*, Automatica, 18 (1982), pp. 229–238.

[32] M. Gevers, *Essays on Control: Perspectives in the Theory and its Applications*, Birkhäuser, Boston, 1993, ch. Towards a joint design of identification and Control?, pp. 111–151.

[33] L. Giarrè and M. Milanese, *H_∞ identification with mixed parametric and non-parametric models*, in Proceedings of the 10th IFAC Symposium on System Identificaion, Copenhagen, 1994, pp. 255–259.

[34] G. 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 (1992), pp. 913–928.

[35] G. Goodwin and B. Ninness, *Model error quantification for robust control based on quasi-bayesian estimation in closed loop*, Proceedings of CDC, (1991).

[36] G. Goodwin, B. Ninness, and M. Salgado, *Quantification of uncertainty in estimation*, Proceedings of the American Control Conference, (1990), pp. 2400–2405.

[37] G. Goodwin and R. Payne, *Dynamic System Identification*, Academic Press, 1977.

[38] G. Goodwin and M. Salgado, *Quantification of uncertainty in estimation using an embedding principle*, Proceedings of ACC, Pittsburgh, (1989).

[39] ——, *A stochastic embedding approach for quantifying uncertainty in the estimation of restricted complexity models*, International Journal of Adaptive Control and Signal Processing, 3(4) (1989), pp. 333–356.

[40] T. Gronwall, *A sequence of polynomials connected with the nth roots of unity*, Bulletin of the American Mathematical Society, 27 (1921), pp. 275–279.

[41] G. Gu and P. Khargonekar, *A class of algorithms for identification in H_∞*, Automatica, 28 (1992), pp. 299–312.

[42] ——, *Linear and nonlinear algorithms for identification in H_∞ with error bounds*, IEEE Transactions on Automatic Control, 37 (July 1992), pp. 953–963.

[43] G. Gu, P. Khargonekar, and Y. Li, *Robust convergence of two-stage nonlinear algorithms for identification in H_∞*, Systems and Control Letters, 18 (1992), p. 253.

[44] G. Gu, D. Xiong, and K. Zhou, *Identification in H_∞ using Pick's interpolation*, Systems and Control Letters, 20 (1993), pp. 263–272.

[45] L. Guo and L. Ljung, *The role of model validation for assessing the size of unmodelled dynamics*, in Proceedings of the 33rd Conference on Decision and Control, IEEE, December 1994, pp. 3894–3899.

[46] G.Zames, *On the metric complexity of causal linear systems:ϵ entropy and ϵ dimension for continuous time*, IEEE Transactions on Automatic Control, AC-24 (1979), pp. 222–230.

[47] R. Hakvoort, *Frequency domain curve fitting with maximum amplitude criterion and guaranteed stability*, in Proceedings of the second European Control Conference, Groningen, Netherlands, 1993, pp. 452–257.

[48] ——, *Worst-case system identification in $\mathcal{H}_\infty$:Error bounds and Optimal Models*, in Preprints of the 12th IFAC World Congress,Sydney, 1993, pp. 161–164, Volume 8.

[49] ——, *Worst-case system identification in ℓ_1:Error bounds, optimal models and model reduction*, in Proceedings of 31st IEEE Conference on Decision and Control, Tucson Arizona, 1993, pp. 499–504.

[50] E. Hannan, *The asymptotic theory of linear time series*, Journal of Applied Probability, 10 (1973), pp. 130–145.

[51] A. Helmicki, C. Jacobson, and C. Nett, *H_∞ identification of stable LSI systems: A scheme with direct application to controller design*, Proceedings of the American Control Conference, (1989), pp. 1428–1434.

[52] ——, *Identification in H_∞: A robustly convergent, nonlinear algorithm*, Proceedings of the American Control Conference, (1990), pp. 386–408.

[53] ——, *Identification in H_∞:Linear algorithms*, Proceedings of the American Control Conference, (1990), pp. 2418–2423.

[54] ——, *Identification in H_∞:The continuous time case*, Proceedings of the American Control Conference, (1990), pp. 1893–1898.

[55] ——, *Fundamentals of control oriented system identification and their application for identification in H_∞*, Proceedings of the American Control Conference, (1991), pp. 89–99.

[56] ——, *Control oriented system identification:A worst case/deterministic approach in H_∞*, IEEE Transactions on Automatic Control, 36 (October 1991), pp. 1163–1176.

[57] H. Hjalmarson and L. Ljung, *A unifying view of disturbances in identification*, in Proceedings of the 10th IFAC Symposium on System Identification, Copenhagan, Volume 2, 1994, pp. 73–78.

[58] H. Hjalmarsson, *Aspects of Incomplete Modeling in System Identification*, PhD thesis, Linköping University, Sweden, 1993.

[59] H. Hjalmarsson, S. Gunnarsson, and M. Gevers, *A convergent iterative restricted complexity control design scheme*, Tech. Rep. LiTH-ISY-I-1653, Department of Electrical Engineering, Linköping University, Sweden, 1994.

[60] K. Hoffman, *Banach Spaces of Analytic Functions*, Prentice-Hall, Inc., New Jersey, 1962.

[61] C. Jacobson and C. Nett, *Worst case system identification in ℓ_1:Optimal algorithms and error bounds*, Proceedings of the American Control Conference, (1991), pp. 3152–3157.

[62] C. Jacobson, C. Nett, and J. Partington, *Worst case system identification in ℓ_1:Optimal algorithms and error bounds*, Systems and Control Letters, 19 (1992), pp. 419–424.

[63] C. Jacobson and G. Tadmor, *A note on H_∞ system identification with probabilistic a-priori information*, in Proceedings of the American Control Conference, San Francisco, IEEE, 1993, pp. 1539–1543.

[64] B. Kacewicz and M. Milanese, *On optimal experiment design in the worst-case ℓ_1 system identification*, in Proceedings of the 31st Conference on Decision and Control,Tucson Arizona, 1992, pp. 296–300.

[65] R. Kosut, *Adaptive control via parameter set estimation*, International Journal of Adaptive Control and Signal Processing, 2 No.4 (1988), pp. 371–400.

[66] ——, *Adaptive robust control via transfer function uncertainty estimation*, Proceedings ACC, Atlanta, (1988).

[67] R. Kosut, M. Lau, and S. Boyd, *Identification of systems with parametric and non-parametric uncertainty*, Proceedings of the American Control Conference, (1990), pp. 2412–2417.

[68] R. Koust, M. Lau, and S. Boyd, *Set-membership identification of systems with parametric and nonparametric uncertainty*, IEEE Transactions on Automatic Control, AC-37 (1992), pp. 929–941.

[69] J. Krause and P. Khargonekar, *Parameter identification in the presence of nonparametric dynamic uncertainty*, Automatica, 26 (1990), pp. 113–123.

[70] ——, *A comparison of classical stochastic estimation and deterministic robust estimation*, IEEE Transactions on Automatic Control, 37 (1992), pp. 994–1000.

[71] J. Krause, G. Stein, and P. Khargonekar, *Robust performance of adaptive controllers with general undertainty structure*, in Proceedings of the 29th Conference on Decision and Control, Honolulu, 1990.

[72] R. LaMaire, L. Valavani, M. Athans, and G. Stein, *A frequency domain estimator for use in adaptive control systems*, in Proceedings of the American Control Conference, IEEE, 1987, pp. 238–244.

[73] ——, *A frequency domain estimator for use in adaptive control systems*, Automatica, 27 (1991), pp. 23–38.

[74] M. Lau, S. Boyd, R. Kosut, and G. Franklin, *Robust control design for ellipsoidal plant set*, in Proceedings of the 30th Confreence on Decisions and Control, Brighton England, 1991, pp. 291–296.

[75] L. Lee and K. Poolla, *Statistical testability of uncertainty models*, in Proceedings of the 10th IFAC Symposium on System Identification, 1994, pp. 189–194.

[76] W. S. Lee, B. Anderson, R. Kosut, and I. Mareels, *On adaptive robust control and control relevant system identification*, in Proceedings of the Automatic Control Conference, IEEE, 1992, pp. 2834–2841.

[77] ——, *A new approach to adaptive robust control*, 7 (1993), pp. 183–211.

[78] ——, *On robust performance improvement through the windsurfer approach to adaptive robust control*, in Proceedings of 32nd IEEE Conference on Decision and Control, San Antonio, Texas, IEEE, 1993, pp. 2821–2827.

[79] L. Lin, L. Wang, and G. Zames, *Uncertainty principles and identification n-widths for lti and slowly varying systems*, in Proceedings of the American Control Conference, 1992, pp. 296–300.

[80] L. Ljung, *System Identification: Theory for the User*, Prentice-Hall, Inc., New Jersey, 1987.

[81] L.Ljung, *Asymptotic variance expressions for identified black-box transfer function models*, IEEE Transactions on Automatic Control, AC-30 (1985), pp. 834–844.

[82] L.Ljung, B. Wahlberg, and H.Hjalmarsson, *Model quality: The roles of prior knowledge and data information*, Proceedings of 30th Conference on Decision and Control, (1991), pp. 273–278.

[83] L.Ljung and Z.D.Yuan, *Asymptotic properties of black-box identification of transfer functions*, IEEE Transactions on Automatic Control, AC-30 (1985), pp. 514–530.

[84] P. Mäkila, *Laguerre series approximation of infinite dimensional systems*, Automatica, 26 (1990), pp. 985–995.

[85] P. MÄKILÄ, *Robust identification and Galois sequences*, International Journal of Control, 54 (1991), pp. 1189–1200.

[86] ——, *Worst-case input-output identification*, International Journal of Control, 56 (1992), pp. 673–689.

[87] P. MÄKILÄ AND J. PARTINGTON, *Robust approximation and identification in H_∞*, Proceedings of the American Control Conference, (1991), pp. 70–76.

[88] ——, *Robust identification of strongly stabilizable systems*, IEEE Transactions on Automatic Control, 37 (1992), pp. 1709–1716.

[89] P. MÄKILÄ, J. PARTINGTON, AND T. GUSTAFSSON, *Robust identification*, Automatica, (1995).

[90] M. MILANESE, *Properties of least squares estimates in set membership identification*, Automatica, 31 (1995), pp. 327–332.

[91] M. MILANESE AND G. BELFORTE, *Estimations theory and uncertainty intervals evaluation in the presence of unknown but bounded errors:Linear families of models and estimators*, IEEE Transactions on Automatic Control, AC-27 (1982), pp. 408–414.

[92] M. MILANESE AND R. TEMPO, *Optimal algorithms theory for robust estimation and prediction*, IEEE Transactions on Automatic Control, AC-30 (1985), pp. 730–738.

[93] M. MILANESE AND A. VICINO, *Estimation theory for nonlinear models and set membership uncertainty*, Automatica, 27 (1991), pp. 403–408.

[94] ——, *Optimal estimation theory for dynamic systems with set membership uncertainty:An overview*, Automatica, 27 (1991), pp. 997–1009.

[95] ——, *Optimal inner bounds of feasible parameter set in linear estimation with bounded noise*, IEEE Transactions on Automatic Control, 36 (1991), p. 759.

[96] ——, *Information based complexity and nonparametric worst-case system identification*, Journal of Complexity, 9 (1993), pp. 427–446.

[97] S. MO AND J. NORTON, *Recursive parameter bounding algorithms which compute polytope bounds*, Proceedings of 12th IMACS World Congress, Paris, 2 (1988), pp. 477–480.

[98] B. NINNESS AND G. GOODWIN, *Robust frequency response estimation accounting for noise and undermodelling*, in Proceedings of the American Control Conference, 1992.

[99] ——, *The Modelling of Uncertainty in Control Systems*, Springer Verlag, 1993, ch. Estimation for Robust Control.

[100] J. Norton, *Identification and application of bounded parameter models*, Automatica, 23 (1987), pp. 497–507.

[101] ——, *Identification of parameter bounds of armax models from records with bounded noises*, International Journal of Control, 42 (1987), pp. 375–390.

[102] J. Norton and S. Veres, *Topics in Stochastic Systems: Modelling, Estimation and Adaptive Control*, vol. 161 of Lecture Notes in Control and Information Sciences, Springer-Verlag, 1991, ch. Developments in Parameter Bounding, pp. 137–158.

[103] P. Parker, *Frequency Domain Descriptions of Linear Systems*, PhD thesis, Australian National University, 1988.

[104] P. Parker and R. Bitmead, *Adaptive frequency response estimation*, Proceedings of the Conference on Decision and Control, (1987), pp. 348–353.

[105] ——, *Adaptive frequency response identification*, Proceedings of 26th Conference on Decision and Control, (1987), pp. 348–353.

[106] B. Parlett, *Some basic information on information-based complexity theory*, Bulletin of the American Mathematical Society, 26 (1992), pp. 3–27.

[107] J. Partington, *Robust identification and interpolation in H_∞*, International Journal of Control, 54 (1991), pp. 1281–1290.

[108] ——, *Robust identification in H_∞*, Journal of Mathematical Analysis and Application, 166 (1991), pp. 428–441.

[109] ——, *Algorithms for identification in H_∞ with unequally spaced function measurements.*, International Journal of Control, 58 (1993), pp. 21–31.

[110] ——, *Interpolation in normed spaces from the values of linear functionals*, Bulletin of the London Mathematical Society, 26 (1994), pp. 165–170.

[111] ——, *Worst-case identification in ℓ_2: linear and nonlinear algorithms*, Systems and Control Letters, 22 (1994), pp. 93–98.

[112] J. R. Partington, *Worst-case identification in banach spaces*, Systems and Control Letters, 18 (1992), pp. 423–428.

[113] H. Piet-Lahanier and E. Walter, *Bounded error tracking of time varying parameters*, IEEE Transactions on Automatic Control, AC-39 (1994), pp. 1661–1664.

[114] K. Poolla, P. Khargonekar, A. Tikku, J. Krause, and K. Nagpal, *A time domain approach to model validation*, IEEE Transactions on Automatic Control, AC-39 (1994), pp. 951–959.

[115] K. Poolla and A. Tikku, *On the time complexity of worst-case system identification*, IEEE Transactions on Automatic Control, AC-39 (1994), pp. 944–950.

[116] N. Rubin and D. Limebeer, *System identification for H_∞ control*, Proceeding of the Conference on Decision and Control, (1992), pp. 1694–1695.

[117] M. Salgado, *Issues in Robust Identification*, PhD thesis, University of Newcastle, 1989.

[118] R. Schrama, *Accurate identification for control:The necessity of an iterative scheme*, Transactions on Automatic Control, AC-37 (1992), pp. 991–994.

[119] R. Schrama and P. V. den Hof, *An iterative scheme for identification and control design based on coprime factorizations*, in Proceedings of the American Control Conference, 1992, pp. 2842–2846.

[120] F. Schweppe, *Recursive state estimation-unknown but bounded errors and system inputs*, IEEE Transactions on Automatic Control, (1968), pp. 22–28.

[121] F. Schweppe, *Uncertain Dynamic Systems*, Prentice Hall, 1973.

[122] R. Smith and M. Dahleh, eds., *Proceedings of the 1992 Santa Barbara Workshop of 'The Modeling of Uncertainty in Control Systems'*, Springer Verlag, 1994.

[123] R. Smith and J. Doyle, *Model validation:A connection between robust control and identification*, IEEE Transactions on Automatic Control, AC-37 (1992), pp. 942–952.

[124] S.M.Veres, *Limited complexity and parallel implementation of polytope updating*, in Proceedings of the American Control Conference, Chicago, 1992, pp. 1061–1063.

[125] R. Tempo, *Robust estimation and filtering in the presence of bounded noise*, IEEE Transactions on Automatic Control, 33 (1988), pp. 864–867.

[126] R. Tempo, *IBC:A working tool for robust parameteric identification*, in Proceedings of the American Control Conference, Chicago, 1992, pp. 237–240.

[127] R. Tempo, *Robust and optimal algorithms for worst-case parametric system identification*, in Proceedings of the 10th IFAC Symposium on System Identification, 1994, pp. Volume 2, 261–265.

[128] R. Tempo and G. Wasilkowski, *Maximum likelihood estimator and worst case optimal algorithms for system identification*, Systems and Control Letters, 10 (1988), pp. 265–270.

[129] S. Tøffner-Clausen, P. Andersen, J. Stoustrup, and H. Niemann, *Estimated frequency domain model uncertainties used in robust controller design - μ approach*, in Proceedings of the 3rd IEEE Conference on Control Applications, Glasgow, 1994.

[130] J. Traub, G. Wasilkowski, and H. Woźniakowski, *Information-Based Complexity*, Academic Press, New York, 1988.

[131] J. Traub and H. Woźniakowski, *Perspectives on information-based complexity*, Bulletin of the American Mathematical Society, 26 (1992), pp. 29–52.

[132] D. Tse, M. Dahleh, and J. Tsitsiklis, *Optimal asymptotic identification under bounded disturbances*, IEEE Transactions on Automatic Control, AC-38 (1993), pp. 1176–1190.

[133] T.Söderström and P.Stoica, *System Identification*, Prentice Hall, New York, 1989.

[134] P. Van den Hof, P. Heuberger, and J. Bokor, *Identification with generalized orthonormal basis functions-statistical analysis and error bounds*, Selected Topics in Identification Modelling and Control, 6 (1993), pp. 39–48.

[135] S. Veres and J. Norton, *Structure selection for bounded-parameter models: Consistency conditions and selection criterion*, IEEE Transactions on Automatic Control, AC-36 (1991), pp. 474–481.

[136] D. D. Vries and P. V. den Hof, *Quantification of model uncertainty from data:Input design, interpolation, and connection with robust control design specifications*, Proceedings of the American Control Conference, Chicago, 4 (1992), pp. 3170–3175.

[137] D. D. Vries and P. V. den Hof, *Quantification of uncertainty in transfer function estimation:A mixed deterministic-probabilistic approach*, in Preprints of the 12th IFAC World Congress, Sydney, IFAC, 1993, pp. 157–160.

[138] B. Wahlberg, *System identification using Laguerre models*, IEEE Transactions on Automatic Control, AC-36 (1991), pp. 551–562.

[139] B. Wahlberg and L. Ljung, *Hard frequency-domain model error bounds from least-squares like identification techniques*, IEEE Transactions on Automatic Control, 37 (1992), pp. 900–912.

[140] E. Walter and H.Piet-Lahanier, *Exact recursive polyhedral description of the feasible parameter set for bounded-error models*, IEEE Transactions on Automatic Control, AC-34 (1989), pp. 911–914.

[141] E. Walter and H. Piet-Lahanier, *Estimation of parameter bounds from bounded error data:A survey*, Mathematics and Computers in Simulation, 32 (1990), pp. 449–468.

[142] A. Weinmann, *Uncertain Models and Robust Control*, Springer-Verlag, New York, 1991.

[143] D. Xiong, G. Gu, and K. Zhou, *Identification in H_∞ via Convex Programming*, in Proceedings of the American Control Conference, San Francisco, 1993, pp. 1537–1538.

[144] R. Younce and C. Rohrs, *Identification with parameteric and non-parametric uncertainty*, IEEE Transactions on Automatic Control, 37 (1992), pp. 715–728.

[145] N. Young, *An Introduction to Hilbert Space*, Cambridge University Press, Cambridge, 1988.

[146] T. Zhou and H. Kimura, *Time domain identification for robust control*, Systems and Control Letters, 20 (1993), pp. 167–178.

[147] A. Zygmund, *Trigonometric Series*, Cambridge University Press, 1959.