Rapproachment Between Bounded Error and Stochastic Estimation Theory

Brett M. Ninness [1] and Graham C. Goodwin

September 9, 1994

ABSTRACT

There has been a recent surge of interest in estimation theory based on very simple noise descriptions; for example, the absolute value of a noise sample is simply bounded. To date this line of work has not been critically compared to pre-existing work on stochastic estimation theory which uses more complicated noise descriptions. The present paper attempts to redress this gap by examining the rapproachment between the two schools of work. For example, we show that for many problems a bounded error estimation approach is precisely equivalent in terms of final result to the stochastic approach of Bayesian estimation. We also show that in spite of having the advantages of being simple and intuitive, bounded error estimation theory is demanding on the quantitative accuracy of prior information. In contrast, we discuss how the assumptions underlying stochastic estimation theory are more complex, but have the key feature that qualitative assumptions on the nature of a typical disturbance sequence can be made to reduce the importance on quantitative assumptions being correct. We also discuss how stochastic theory can be extended to deal with a problem at present tackled only by bounded error estimation methods: the quantification of estimation errors arising from the presence of undermodelling.

Technical Report EE9241

Department of Electrical and Computer Engineering, University of Newcastle, Callaghan 2308, AUSTRALIA

[1] The author is grateful to the Centre for Industrial Control Science at the University of Newcastle, Australia and the Australian Telecommunications and Electronics Research Board for Scholarship Assistance.

1 Introduction

The mainstream of estimation theory to the present date relies on modelling any disturbances corrupting observed data via random variable descriptions [12, 20, 35]. This necessitates specification of probability density functions and correlation information for the random variables making up a disturbance sequence. A variety of theoretically sound estimation schemes have been developed from this framework such as Maximum Likelihood, Least Squares and Bayesian techniques. The choice of scheme depends on the application. For example, if the desire is to find feasible regions for the parameters of a system, then it is possible to use Bayes' rule to find the probability density function for these parameters conditional upon observed data. For the case when the disturbances can be taken to have Gaussian distributions this approach is very popular since the relevant calculations are very easily performed by running a Kalman filter. Unfortunately, when the disturbances are not Gaussian distributed, then the calculations can be very complex. For example, in general it will not be sufficient to just keep track of the first two moments of the conditional distribution as the Kalman filter does.

In these cases the usual approach is to choose a simpler estimation scheme, such as least squares, and then use the Central Limit Theorem to obtain information on the distribution of the estimate. This leads to confidence regions for the parameters. A deficiency with this paradigm is that the confidence regions provide only soft, not hard bounds, even when the disturbance may satisfy a hard bound. Furthermore, these confidence regions only apply as the amount of observed data tends to infinity. It is difficult to determine how reliable they are for finite data.

Motivated by these and other issues a large body of work has emerged under the title of 'bounded error' estimation theory [33, 6, 7, 23, 24, 25, 19, 18, 16]. By comparing assumptions to observed data an ingenious solution to the parameter bounding problem is obtained by only calculating the support for the a-posteriori distribution of the parameters. 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. The attraction of these minimal assumptions has resulted in rapid progress in the bounded error estimation area with little reflection on the rapproachment between these new methods and pre-existing estimation methods based on stochastic disturbance models. The work in [17, 32] are examples of such reflection and this paper is intended for the same area. For example, we show that for problems expressible in linear regressor form, the feasible parameter set obtained by bounded error estimation methods is precisely the support of the posterior probability density for the parameters calculated by Bayes rule in the special case of the random variables modelling the disturbance sequence being independent. That is, for a large class of problems, bounded error estimation methods can be considered as a special case of stochastic estimation - Bayesian estimation when the disturbance is an independent process.

The motivating idea of bounded error estimation theory is the minimal assumptions it makes on disturbances. As we shall discuss, a consequence of these minimal assumptions is that the resultant parameter bounds can be sensitive to the disturbance bound regardless of how much data is observed. In contrast, when stochastic models for disturbances are used then confidence intervals tend to be insensitive to prior quantitative specifications on disturbances, and the sensitivity decreases with the amount of observed data. For example the quantitative specifications can, if necessary, be estimated with increasing accuracy as the amount of available data grows. This is intriguing and we trace the phenomenon to the fact that the more complex stochastic model is able to account for the 'on-average' nature of the disturbance. We conclude that if one can make qualitative assumptions on the average properties of a disturbance sequence, then a stochastic model for the disturbance can be employed to allow quantitative properties of the disturbance sequence, such as bounds, to be estimated from the available data. We also examine, via example, how the complex density function for least squares estimates from finite data can be approximately calculated. This leads to the observation that the convergence in the central limit theorem can in some cases be very rapid so that confidence intervals can be accurate even for short data lengths.

These results suggest that depending on the application, both stochastic and bounded error estimation methods have their place; the choice of paradigm depending chiefly on a tradeoff

between the required accuracy of qualitative information and quantitative information. If one can make qualitative assumptions about the on-average properties of a disturbance realisation, then a stochastic framework seems appropriate. If such assumptions cannot be made then a bounded error approach may be appropriate.

An important example of where this decision of framework must be made is the problem of providing parameter bounds that take into account undermodelling for the subsequent purpose of robust controller design. The mainstream approaches to this problem [37, 30, 29, 31, 21, 14, 15, 28] draw on bounded error estimation ideas in their solution to the problem. However, as discussed, the cost of the simple disturbance models used is sensitivity to the quantitative accuracy of the bounds assumed for disturbances and undermodelling. Consequently, the robust controller for a plant arrived at via these mainstream approaches will be sensitive to prior information about the plant. This contradicts the original ambit of robust design which is to achieve insensitivity to prior information. To obtain the required insensitivity we detail an appropriate random variable description of the undermodelling which allows a stochastic approach to the problem. These ideas have been discussed, but not compared to bounded error methods, in [22, 10, 11, 13, 8]. The key idea in this new approach is to find bounds that apply not only over a chosen class of disturbance sequences, but also over a chosen class of undermodellings. In the current paper we go on to explain, as in [9], how parameters describing the undermodelling need not be specified a-priori; they can be estimated from the data if certain qualitative assumptions on the nature of the undermodelling are made. We conclude by reporting some of the results of [1] that examine the success of such a scheme.

2 A Motivating Example

We begin with a simple scalar motivating example; namely estimation of a fixed scalar value θ_0 in the presence of bounded disturbances:

$$y_k = \theta_0 + \nu_k \tag{1}$$

Here $\{y_k\}$ is our observation sequence and $\{\nu_k\}$ is the disturbance sequence. Suppose that each realization of ν_k is bounded:

$$|\nu_k| \leq \delta \qquad \delta \in \mathbb{R}^+ \tag{2}$$

In most cases we will not a-priori know the bound δ exactly, but we will be able to find an overbound $\mu > \delta$. The key idea of bounded error estimation theory is to identify feasible parameter sets that are consistent with observed data $\{y_k\}$ and the assumed disturbance bound μ. Using this principle we will never be able to find θ_0 to an accuracy greater than $\pm|\mu - \delta|$. Specifically, the most informative observations for the purposes of bounded error estimation will be $y^+ = \theta_0 + \delta$ and $y^- = \theta_0 - \delta$. If we are lucky enough to get these observations then we can conclude (at best)

$$\theta - \theta_0 \in \mathcal{D} = [\delta - \mu, \mu - \delta] \tag{3}$$

Furthermore, after observing $y^+ = \theta_0 + \delta$, $y^- = \theta_0 - \delta$ we can never reduce $\mathcal{D}$, no matter how much more data we observe. Additionally, the size of $\mathcal{D}$ is linearly related to the assumption μ.

This is in contrast to the case where a zero mean random variable description is appropriate for $\{\nu_k\}$. Within this framework a confidence interval is generated which shrinks around the true value θ_0 as the amount of observed data grows. This necessitates more complicated assumptions on disturbances. A probability space $\{\Omega, \mathcal{F}, \mathbb{P}\}$ is assumed to model the effects of nature. The disturbance model is that an event ω from nature Ω arises to result in a particular disturbance realisation via the random variable mapping $\nu_k(\omega)$. The likelihood of whole sets of events $A \in \mathcal{F} \subseteq \Omega$ is determined as $\mathbb{P}(A) = \int_X \nu(x) f_\nu(x) dx$ where $\nu_k(A) \mapsto X \in \mathbb{R}$ and f_ν is the probability density function of ν. An appropriate f_ν reflecting the assumption (2) is the uniform density

$$f_\nu(x) = \begin{cases} \frac{1}{2\delta} & ; x \in [0, \delta] \\ 0 & ; \text{otherwise} \end{cases} \tag{4}$$

Under this more complicated stochastic model we can adopt a Bayesian approach to the problem by calculating the probability of θ given observed data $\{y_1, \cdots, y_N\}$. This is given, modulo a constant normalizing factor as

$$\mathbb{P}(\theta \mid y_1, \cdots, y_N) = \mathbb{P}(y_1, \cdots, y_N \mid \theta)\mathbb{P}(\theta) \tag{5}$$

and by the definition of conditional probability

$$\begin{aligned} \mathbb{P}(y_1, \cdots, y_N \mid \theta) &= \prod_{k=1}^{N} \mathbb{P}(y_k \mid y_1, \cdots, y_{k-1}, \theta) && (6) \\ &= \prod_{k=1}^{N} \mathbb{P}(y_k \mid \theta) && (7) \end{aligned}$$

where the last line follows by assuming independence of the $\{\nu_k\}$ process to give

$$\mathbb{P}(\theta \mid y_1, \cdots, y_N) = \mathbb{P}(\theta) \prod_{k=1}^{N} \mathbb{P}(y_k \mid \theta) \tag{8}$$

where $\mathbb{P}(y_k \mid \theta)$ has corresponding density function

$$f_{y_k|\theta}(x) = \begin{cases} \frac{1}{2\delta} & ; |x - \theta| < \delta \\ 0 & ; \text{otherwise} \end{cases} \tag{9}$$

Therefore, due to the product nature of (8), the a-posteriori density for $\theta, f_{\theta|Y}$ will only be non-zero on the set Θ given by

$$\Theta = \bigcap_{k=1}^{N} \Theta_k \qquad \Theta_k \triangleq \{\theta : |y_k - \theta| \leq \delta\} \tag{10}$$

This set Θ is precisely the set $\mathcal{D}$ calculated by the bounded error estimation method we previously discussed. This latter method is thus subsumed by the stochastic approach of Bayesian estimation under the special assumption of disturbances being independent.

Of course, under a stochastic disturbance model many other estimation schemes are possible such as maximum likelihood and least squares. For this scalar example the least squares estimator $\hat{\theta} = \frac{1}{N}\sum_{k=1}^{N} y_k$ for θ_0 results in the estimation error $\tilde{\theta} \triangleq \hat{\theta} - \theta_0$ being a random variable given by

$$\tilde{\theta}_N(\omega) = \frac{1}{N}\sum_{k=1} \nu_k(\omega) \tag{11}$$

It is well known that if the ν_k are assumed independent we can find the density function $f_{\tilde{\theta}}(\theta)$ by convolution [38]:

$$f_{\tilde{\theta}}(\theta) = \left[\bigotimes_{k=1}^{N} f_\nu(x)\right](\theta) \tag{12}$$

where

$$[f_\nu(x) \otimes f_\nu(x)](\theta) = \int_{-\infty}^{\infty} f_\nu(\theta - x) f_\nu(x) dx \tag{13}$$

If we know δ exactly as $\delta = 0.5$ and look at data lengths of $N = 1, \cdots, 4$ then the corresponding densities $f_{\tilde{\theta}}$ are shown in figure 1. As can be seen, after only 4 terms the true density function obtained by convolving the uniform density function (9) is almost indistinguishable from a Gaussian. This illustrates that extremely quick convergence in central limit theorems is possible if the true variance is known. The feasible parameter set is then taken to be the set in θ space centered on $\hat{\theta}_N$ where $f_{\tilde{\theta}}(\theta)$ is greater than some tolerance ϵ. The smaller ϵ is made, the less stress is placed

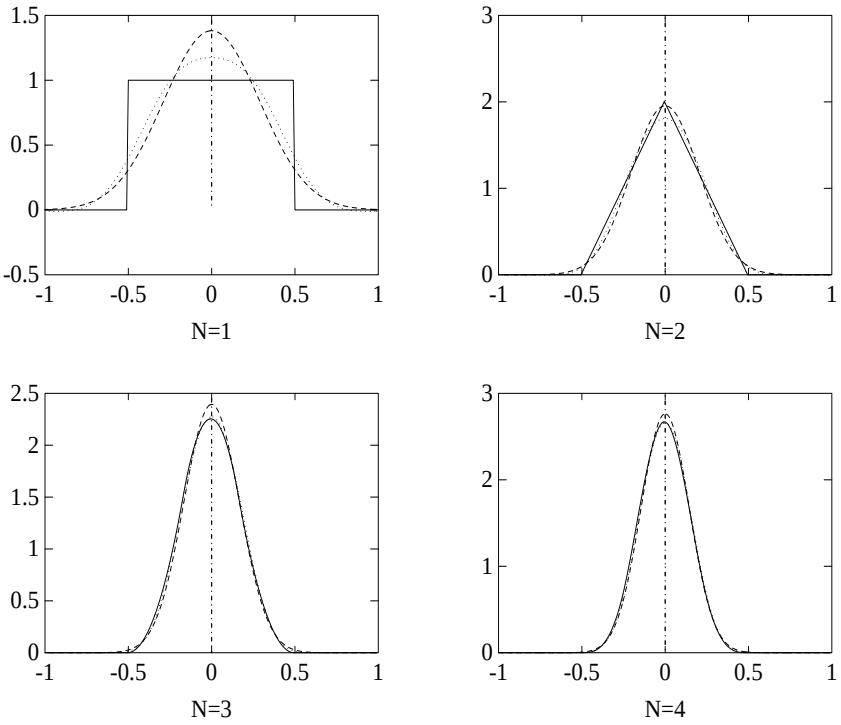

Figure 1: Illustration of how quickly the density function for the average of independent random variables converges to the Gaussian density function. The solid line is the true density function, the dashed line is a Gaussian distribution with the same variance. The dotted line is a closed form approximation to the true density.

upon quantitative assumptions being correct, but the larger the consequently more conservative parameter set $\mathcal{D}$ will be.

If we consider the more realistic case of the bound $\delta = 0.5$ not being known exactly so that a conservative overbound $\mu = 0.8$ is assumed then the estimation results are shown in figure 2 for data lengths of $N = 1, 2, 3, 30$. The vertical dash-dot lines indicate the optimum feasible parameter set $\mathcal{D}$ that can be provided by bounded error estimation theory (see (3)). However, we see that after we have observed only 30 data points the density function shown as the curved dashed line, will allow us to conclude a set $\mathcal{D}$ smaller than that provided by bounded error estimation theory. The size of the feasible parameter set $\mathcal{D}$ is not sensitive to precise knowledge of δ and the sensitivity decreases as more data is observed. The averaging of the least squares estimator provides robustness under a random variable model that is able to exploit this averaging.

Furthermore, this confidence interval set will be obtained every time we do such an experiment, whereas to get a minimal size for the set $\mathcal{D}$ from bounded error estimation theory we need a very lucky set of observations. This latter set will not decrease in size as the number of observations grows, whereas for fixed ϵ the set $\mathcal{D}$ obtained from the density function will shrink. Finally, and perhaps most importantly, we note that the density function $f_{\hat{\theta}}(\theta)$ tells us that some values of θ within $\mathcal{D}$ are far more worthy of our consideration than others.

Given this example, we move on to the general class of problems we would like to consider, namely data generated as:

$$y_k = \phi_k^T \theta_0 + \nu_k \tag{14}$$

where ϕ_k is a vector of known regressors, $\{y_k\}$ is an observed output realization, θ_0 is a vector of parameters to be estimated, and $\{\nu_k\}$ is a bounded disturbance sequence. We assume ϕ_k to be of

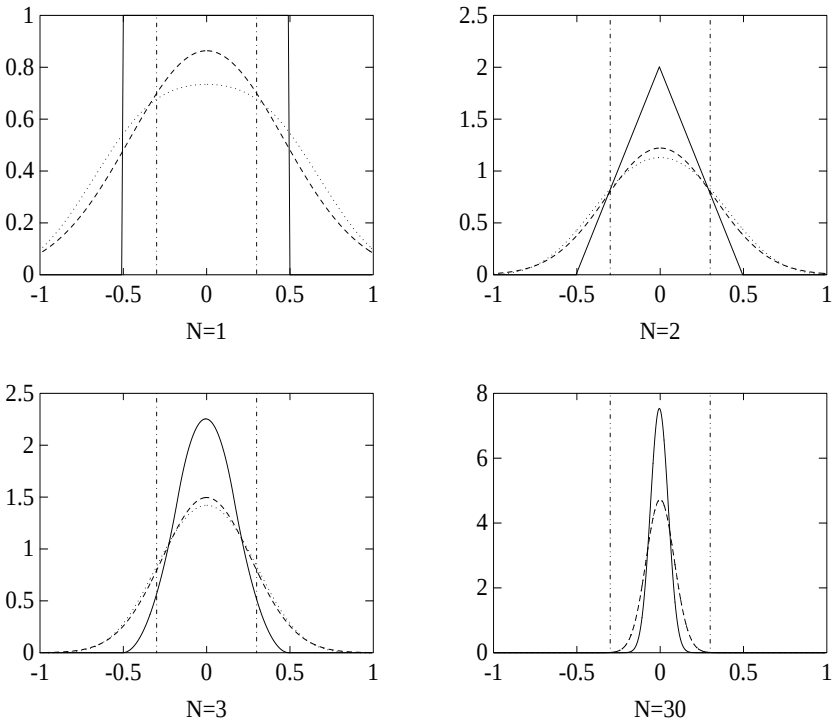

Figure 2: Illustration of the utility of using distribution functions for finding bounds on $\mathcal{D}$. The solid line is the true density function, the dashed line is a Gaussian distribution with the same variance. The vertical dash-dot lines give that smallest $\mathcal{D}$ that bounded error estimation theory can provide under conservative assumed disturbance bound $\delta = 0.8$ when the true value is $\delta = 0.5$.

the form:

$$\phi_k^T = \left[\mathcal{B}_1(q^{-1})u_k, \cdots, \mathcal{B}_p(q^{-1})u_k\right] \tag{15}$$

where the $\mathcal{B}_k(q^{-1})$ are stable functions in the backward shift operator q^{-1} and $\{u_k\}$ is a known input sequence to the system. Note that for simplicity we are not considering the case of filtered versions of $\{y_k\}$ appearing in ϕ_k since this leads to awkward feasible sets $\mathcal{D}$ for θ_0 [19]. In this case the generation of feasible parameter sets $\mathcal{D}$ consistent with bounded noise assumptions and observations is a well studied problem [7, 16, 24, 23] which can be most accurately solved as a linear programming problem that finds $\mathcal{D}$ as

$$\mathcal{D} = \bigcap_{k=1}^{N} \mathcal{D}_k \tag{16}$$

where $\mathcal{D}_k$ is the area in θ space between two hyperplanes:

$$\mathcal{D}_k = \left\{\theta \in \mathbb{R}^p : -\delta \leq y_k - \phi_k^T\theta \leq \delta\right\} \tag{17}$$

The resulting polytope $\mathcal{D}$ can be found off-line by linear programming, can be calculated on-line recursively [25] or, as is more usual, overbounded by an ellipsoid [23]. Whatever the method, if we assume the bound $|\nu_k| \leq \mu$ when in fact the true bound is $|\nu_k| \leq \delta$ where $\delta < \mu$ then we have the following lower bound on how small $\mathcal{D}$ can be made:

Theorem 1. *Define d as the diameter of the largest ball that can be placed inside $\mathcal{D}$. Assume the energy in the regressors is bounded as $\|\phi_k\|_2 \leq \sigma_\phi$. Then*

$$d > \frac{2(\mu - \delta)}{\sigma_\phi} > 0 \tag{18}$$

Proof. See Appendix A □□□

As in the preceding scalar example, if we employ an independent random variable modelling for the disturbance ν_k with density given by (4), then a Bayesian approach to the estimation problem involves using (8) together with

$$f_{y_k|\theta}(y) = \begin{cases} \frac{1}{2\delta} & ; |y - \phi_k^T \theta| \leq \delta \\ 0 & ; \text{otherwise} \end{cases} \tag{19}$$

so that again, the product in the expression (8) means that the posterior density for θ is only non-zero on the domain $\mathcal{D}$ given by (16) and (17). Therefore, for the linear regression model (14) and (15) we see that bounded error estimation is subsumed by a stochastic estimation method: Bayesian estimation when the random variables modelling the disturbance sequence are independent.

As in the preceding scalar example we examine the properties of another estimator that has been well studied in a stochastic framework: the least squares estimate of θ_0 given by

$$\hat{\theta}_N = P_N^{-1} \frac{1}{N} \sum_{k=0}^{N-1} \phi_k y_k \tag{20}$$

$$P_N = \frac{1}{N} \sum_{k=0}^{N-1} \phi_k \phi_k^T \tag{21}$$

Combining (20) and (21) with the model (14) gives the estimation error as

$$\tilde{\theta}_N = P_N^{-1} \frac{1}{N} \sum_{k=0}^{N-1} \phi_k \nu_k \tag{22}$$

In the case where the disturbance is an independent sequence then use of the Central Limit Theorem [20, 35] gives[2].

$$\frac{\sqrt{3N}}{\delta} P_N^{1/2} (\hat{\theta}_N - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I) \quad \text{as} \quad N \to \infty \tag{23}$$

This allows us to generate ellipsoidal confidence regions for $\tilde{\theta}$ by recognising that quadratic forms of Gaussian random variables are χ^2 distributed:

$$\frac{3N}{\delta^2} (\hat{\theta}_N - \theta_0)^T P_N (\hat{\theta}_N - \theta_0) \xrightarrow{\mathcal{D}} \chi_p^2 \quad \text{as} \quad N \to \infty \tag{24}$$

where $p = \dim \theta_0$. In the case of fixed confidence level the size of the ellipsoidal confidence region for θ_0 may be measured by V_N, the sum of the lengths of the principle axes and will obey

$$V_N \propto \frac{\delta^2}{3N} \operatorname{Tr}\{P_N\} \tag{25}$$

If the input is of finite power then P_N will tend to a constant $< \infty$ [20] so we have that the sensitivity of the size measure V_N to knowledge of δ is

$$\frac{1}{\delta} \frac{\partial V_N}{\partial \delta} \propto \frac{2}{3N} \tag{26}$$

That is, the sensitivity decreases linearly with increasing data. This is in contrast to the result of Theorem 1 where no sensitivity decrease occurs. Additionally, it is well known that the variance

[2] Note with f_ν as defined in (4) we have $\operatorname{Var}\{\nu_k\} = \delta^2/3$

of measurement noise can be estimated from prediction error residuals. This leads to an estimate $\hat{\delta}_N$ of δ which can be estimated from N observed data points as

$$\hat{\delta}_N = \sqrt{\frac{3}{N-p}\sum_{k=0}^{N-1}(y_k - \phi_k^T\hat{\theta}_N)^2} \tag{27}$$

This estimation ability for δ further decreases the stress on prior assumptions on δ being correct. All these results on least squares estimation do, however, require a qualitative assumption to be correct. The average over a typical sample path for a disturbance sequence must be near zero. If this is not the case then perhaps one can achieve it by adding an offset term in the parameter vector θ_0 to be estimated. If this still doesn't seem a reasonable model for an expected disturbance sequence then perhaps a bounded error estimation method is a better approach to the problem. The point is to recognize that if one can make a qualitative assumption that disturbances wander evenly around zero, then the stress on the quantitative assumption of δ being correct can be lessened by estimating δ and/or using an estimator that averages the data in an effort to reduce the effect of the disturbance.

As mentioned in the introduction, these results on least squares estimation only provide soft bounds and these soft bounds only apply as the amount of data is allowed to increase to infinity. If one requires bounds that apply for finite data under a stochastic disturbance model then the probability density of $\tilde{\theta}_N$ is required. In the case of independent disturbance sequences this can be accurately approximated via a Fourier expansion in Hermite polynomials [5]:

Theorem 2. *For scalar θ and $f_\nu(x)$ given as in (4), the density function $f_{\tilde{\theta}}(\theta)$ for the estimation error $\tilde{\theta}_N$ for finite N can be approximated by a Gaussian density multiplied by a correction term $\varepsilon(\theta)$.*

$$f_{\tilde{\theta}}(\theta) \approx \frac{1}{\sigma}\Phi\left(\frac{\theta}{\sigma}\right)\varepsilon(\theta) \tag{28}$$

where

$$\varepsilon(\theta) \triangleq \left[1 + \frac{1}{24}\left(\frac{m_4}{\sigma^4} - 3\right)\left(\left(\frac{\theta}{\sigma}\right)^4 - 6\left(\frac{\theta}{\sigma}\right)^2 + 3\right)\right] \tag{29}$$

$$m_4 \triangleq \left(\frac{\delta}{P_N}\right)^4\left[\frac{1}{5}\sum_{k=1}^{N}\phi_k^4 + \frac{2}{3}\sum_{k=1}^{N}\sum_{j>k}^{N}\phi_k^2\phi_j^2(1 - 3\ln\phi_k^2\phi_j^2)\right] \tag{30}$$

and

$$\sigma^2 = \frac{\delta^2}{3P_N} \qquad P_N = \sum_{k=1}^{N}\phi_k^2 \tag{31}$$

where $\Phi(x)$ is the standard normal density function and N is the number of observed data points.

Proof. See Appendix B □□□

If we apply this result to the scalar motivating example where $\phi_k = 1$ we obtain for $N = 10$

$$f_{\tilde{\theta}}(\theta) \approx \frac{1}{\sigma}\Phi\left(\frac{\theta}{\sigma}\right)\left[0.985 + 0.03\left(\frac{\theta}{\sigma}\right)^2 - 0.0005\left(\frac{\theta}{\sigma}\right)^4\right] \tag{32}$$

Therefore, for $|\theta/\sigma| < 1$ the true density function is very close to the Gaussian suggested by the Central Limit Theorem even though $N = 10$ is so few observations that one may not expect an asymptotic result to hold with accuracy. For $|\theta/\sigma| > 1$, $\Phi(\theta/\sigma)$ becomes very small so the multiplicative correction term is not important for the calculation of confidence regions. To illustrate the accuracy of the approximation (32) it is plotted as a dotted line in figure 1 and should be compared to the solid line, which is the true density. The agreement is quite close indicating that in the case of an independent disturbance process it is possible to accurately calculate hard bounds for the least squares estimation error and to calculate the probability density within these bounds.

3 Multivariable Simulation Example

We now wish to introduce a simulation example which will be carried through the remainder of the paper. The following continuous time system, sampled with period 1 second, was simulated :

$$G(s) = \frac{1}{(10s+1)(s+1)}$$

The test input sequence $\{u_k\}$ was a 0.02 Hz fundamental square wave. The output of this system was corrupted with a noise sequence $\{\nu_k\}$ distributed uniformly as in (4) with $\delta = 0.1$. One hundred and fifty samples of data were collected, the first one hundred were used to get rid of initial condition effects in the simulated plant and regressor filters, and the last fifty were used for least squares model fitting. A 2nd order model of the form:

$$y_k = \left[\frac{\theta_1 q^{-1}}{(1+\xi q^{-1})} + \frac{\theta_2 q^{-1}(1-(2+\xi)q^{-1})}{(1+\xi q^{-1})^2} \right] u_k \tag{33}$$

was fitted to the data using least squares. Here $\xi = -0.8$ was chosen (between the true system poles). Note that the unusual regressors are motivated by Laguerre polynomials [36, 26]. The resulting least squares estimates were:

$$\hat{\theta}_1 = 0.5104 \quad \hat{\theta}_2 = 0.3471 \tag{34}$$

The Bounded error estimates were found recursively via the ellipsoidal overbounding methods obtained by Fogel and Huang [7] with an assumed disturbance bound of $\mu = 0.1$. This algorithm appears to be one of the more popular methods of obtaining simple ellipsoidal feasible parameter sets in bounded error estimation theory. The results are shown in figure 3. The top left plot shows the observations $\{y_k\}$ together with the response of the identified plant parameterized by (34). The top right plot shows the evolution of $\pi\sqrt{\det P_t}$ which indicates the volume of the overbounding set $\mathcal{D}_t$ given in Fogel and Huang's algorithm. The bottom left plot shows the data weighting sequence ρ_k involved in the algorithms. In this simulation it indicates that the data sequence is highly informative, $\mathcal{D}_k$ being refined on almost every sample. Finally, the lower right plot shows the prediction error residuals demonstrating that $\mu = 0.1$ was a valid bound to assume. Feasible parameter sets for this simulation example are shown in figure 4. The true feasible parameter set $\mathcal{D}_{50} = \bigcap_{k=1}^{50} \mathcal{D}_k$ is the small polytope formed by the intersection of the straight lines. The largest ellipse is the ellipsoidal overbounding region for this polygon obtained by Fogel and Huang's algorithm. The smallest ellipse is the set motivated by the central limit theorem result (24):

$$\frac{3N}{\delta^2}\tilde{\theta}^T P_N \tilde{\theta} \leq 50 \tag{35}$$

where P_N is given in (21). That is, the smallest ellipsoid is found by assuming $\tilde{\theta}$ has a density function $f_{\tilde{\theta}}(\theta)$ approximately that of the Gaussian density shown in figure 5. The boundary of $\mathcal{D}$ is taken to be the values of $\tilde{\theta}$ where $\int_{\mathcal{D}} f_{\tilde{\theta}}(\theta) > 0.99995$ and $f_{\tilde{\theta}}$ is taken to be a χ^2 density with 2 degrees of freedom. We chose a value very close to 1.0 in order to approximate a hard bounded set. As can be seen, this set is a good indication of the true feasible parameter polytope. This supports our claim that it is possible to obtain useful hard bound information in a stochastic framework by choosing the confidence region to be one of probability almost 1.0. Note that the confidence region is chosen on the assumption that the disturbance sequence is zero mean and independent. In the present simulation example this will not be the case due to the influence of undermodelling - the restricted fixed denominator model set we chose does not contain the 'true' system we simulated. The fact that the the confidence regions are informative even though the stochastic assumptions are not exactly satisfied indicates a robustness property of stochastic analysis that we believe has in large part been responsible for its popularity and acceptance. It is also interesting to note the conservatism of the Fogel and Huang overbound which indicates the utility of approaches [25] to calculate the true polytopal feasible parameter set.

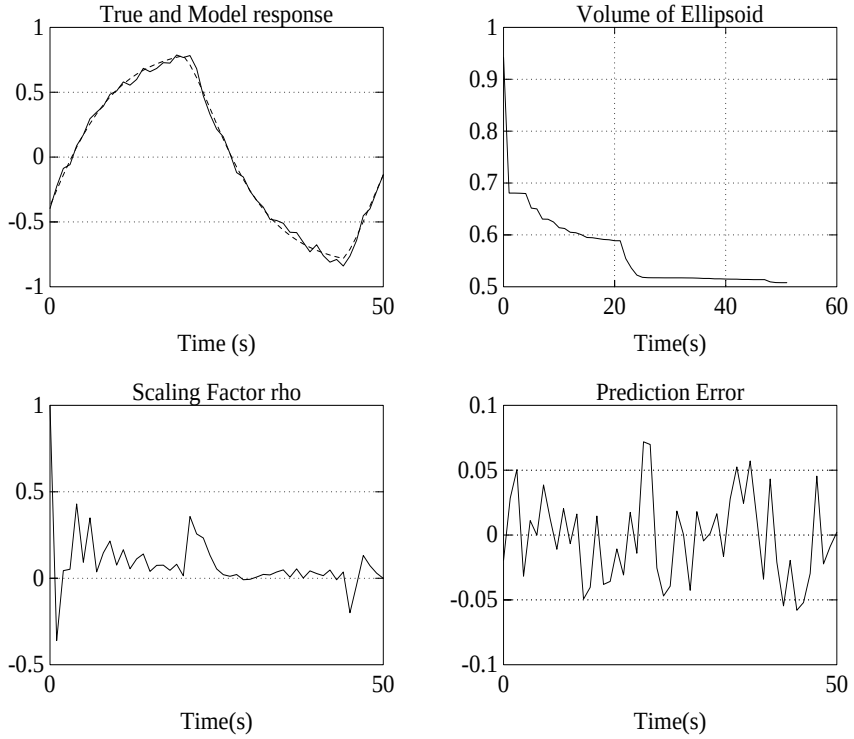

Figure 3: The top left plot shows the true plant response compared with that of its least squares identified model. The top right plot shows the evolution of the volume of the recursively calculated bounding ellipsoid using the algorithm of Fogel and Huang, while the bottom left plot shows the evolution of the weighting parameters ρ_k. The bottom right plot shows the prediction residuals of the least squares identified system.

4 Undermodelling, Bounded Error Estimation, and Stochastic Embedding

An important problem that has been addressed by application of bounded error estimation theory is the problem of quantifying the errors in estimated frequency responses that arise due to undermodelling. The motivation is the desire to employ robust methods in the design of control systems. This robust theory requires estimated frequency responses to be provided with bounds on any errors in their specification. The mainstream of thought on this problem is to suppose a-priori restrictions on quantities such as the smoothness or magnitude of any unmodelled dynamics (usually in the frequency domain) and then translate this information into pointwise bounds on any output mismatch that will occur due to the use of a restriced complexity model. This becomes a bounded disturbance. Thus the problem seems perfectly set up for bounded error estimation theory to be applied in order to find bounds on the undermodelling induced bias errors in θ. There are many examples of this approach covering a gamut of different model structures, undermodelling descriptions and methods of translating time domain to frequency domain bounds. See [28, 29, 31, 37, 2] for examples.

However, as Theorem 1 indicates, the bounds arrived at will be sensitive to prior knowledge of the undermodelling induced bounded disturbance, and hence will be sensitive to prior knowledge of the undermodelling itself. Moreover, these bounds are intended for use in robust controler synthesis whose mandate is to provide closed loop performance that is insensitive to the accuracy of prior information on the system to be controlled. There appears to be a contradiction here.

In an effort to overcome this contradiction while still providing useful results, the current

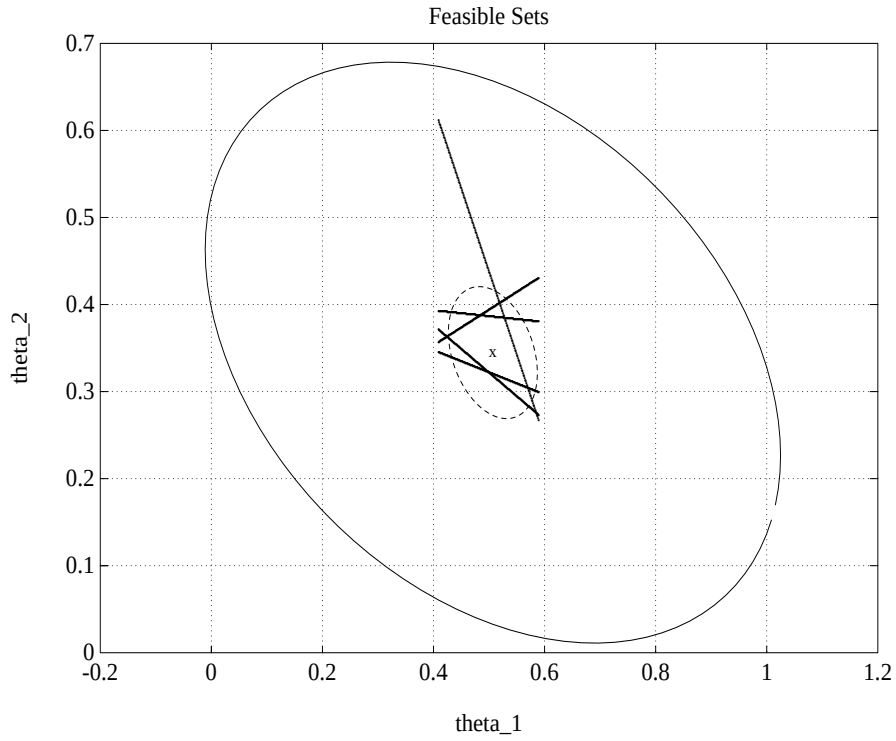

Figure 4: Feasible parameter sets for the previous simulation. Largest ellipse is obtained via Fogel and Huang's algorithm. Polygon is the true feasible parameter set that the largest ellipse is attempting to approximate. Smallest ellipse is the 99.995% confidence region obtained by the Central Limit Theorem. The cross marks the least squares estimate of the parameters.

authors and others [11, 8, 9] have explored the possibility of embedding the description of the undermodelling in a stochastic framework in an effort to reduce the sensitivity to prior information. At first this may seem inappropriate since the effects of undermodelling are manifestly different from those we normally associate with noise which is modelled via a random variable description. However, the association of probability theory with notions of randomness is largely historical. Indeed, as Catlin [4] points out, 'a random variable $[\nu_k(\omega)]$ is like the Holy Roman Empire- it wasn't holy, it wasn't Roman and it wasn't an Empire. A random variable is neither random nor variable, it is simply a function',

The key issue is that we need to provide error bounds that apply over an ensemble of different undermodellings. The solutions using bounded error estimation theory is to use hard bounds valid for all undermodellings. The stochastic embedding approach we are about to describe amounts to working with weighted averages over the ensemble of possible undermodellings. That we choose to do this averaging over a probability space and with a weighting that corresponds to a probability measure is essentially co-incidental.

To be specific, we assume that for some $\theta_0 \in \mathbb{R}^p$ and for the chosen model set we can write the true frequency response as

$$G_T(e^{-j2\pi f}) = G(e^{-j2\pi f}, \theta_0) + G_\Delta(e^{-j2\pi f}) \tag{36}$$

where $G(e^{-j2\pi f}, \theta_0)$ is the frequency response of a p dimensional model parameterised by θ_0 and $G_\Delta(e^{-j2\pi f})$ is the frequency response of any 'unmodelled dynamics' not explained by $G(e^{-j2\pi f}, \theta_0)$. Here we have used the symbol f for cyclic frequency normalised with respect to the sampling frequency since we don't want confusion with $\omega \in \Omega$ the probability space.

The key idea of the stochastic embedding approach is that we now assume that $G_\Delta(e^{-j2\pi f})$ can

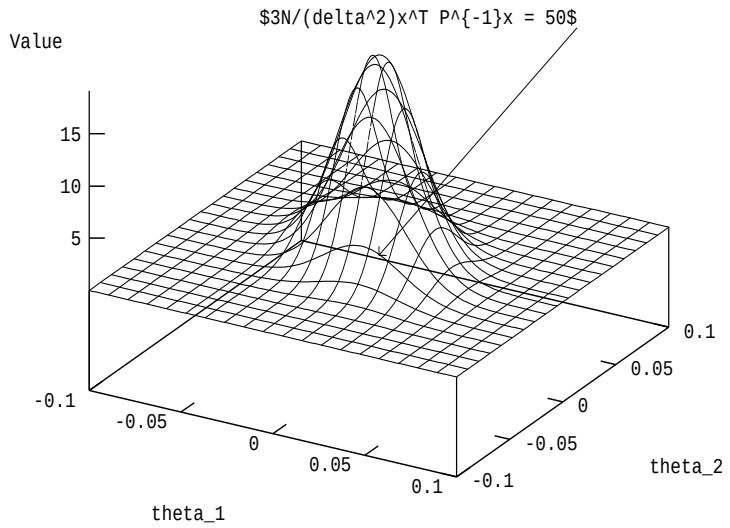

Figure 5: The multivariable density function used in finding the feasible parameter set in the stochastic case. It was formed by intersecting a plane at the level of the arrow $3N/\delta^2 x^T P^{-1} x = 50$ with the Gaussian surface.

be modelled as a function on the measure space $\{\Omega, \mathcal{F}_\Delta, \mathbb{P}_\Delta\}$. This does not imply that $G_\Delta(e^{-j2\pi f})$ is random. It is merely convenient to index the ensemble of possible undermodellings by $\omega \in \Omega$, and to concentrate iterest on some classes of undermodellings more than others by the choice of $\mathbb{P}_\Delta$. For example, we may assume that averaged over all undermodellings, the contribution of the undermodelling $G_\Delta(e^{-j2\pi f}, \omega)$ is zero:

$$\int_\Omega G_\Delta(e^{-j2\pi f}, \omega) d\mathbb{P}_\Delta = \int_{-\infty}^{\infty} G_\Delta(e^{-j2\pi f}, x) f_\Delta(x) dx = 0 \tag{37}$$

We may make assumptions also on how variations in f and ω interact. For example, we may assume:

$$\int_\Omega G_\Delta(e^{-j2\pi f_1}, \omega) G_\Delta(e^{j2\pi f_2}, \omega) d\mathbb{P}_\Delta = \frac{\alpha\lambda e^{-j2\pi f}}{1 - \lambda e^{-j2\pi f}} \quad \alpha, \lambda \in \mathbb{R}^+, \quad f = f_1 - f_2 \tag{38}$$

If we go further to assume that $G_\Delta(e^{-j2\pi f}, \omega)$ is the frequency response of a linear system with impulse response sequence $\{\eta_k(\omega)\}$ then we have

$$G_\Delta(e^{-j2\pi f}, \omega) = \sum_{k=1}^{L} \eta_k(\omega) e^{-j2\pi f k} \tag{39}$$

That is

$$\eta_k(\omega) = \int_0^1 G_\Delta(e^{-j2\pi f}, \omega) e^{j2\pi f k} df \tag{40}$$

so that the assumption (38) becomes:

$$\int_\Omega \eta_k(\omega)\eta_\ell(\omega) d\mathbb{P}_\Delta = \begin{cases} \alpha\lambda^k & k = 1 \\ 0 & k \neq \ell \end{cases} \tag{41}$$

Assumptions (38) and (41) are thus equivalent. In the sequel, assumptions (41) will be more convenient from an implementational perspective, but (38) is more useful for intuition purposes since from it we get the smoothness constraint

$$\int_\Omega \left|G_\Delta(e^{-j2\pi f_1},\omega)-G_\Delta(e^{-j2\pi f_2},\omega)\right|^2 d\mathbb{P}_\Delta = \frac{2\alpha\lambda(1+\lambda)(1-\cos 2\pi f)}{(1-\lambda)(1+\lambda^2-2\lambda\cos 2\pi f)} \leq \frac{4\pi^2\alpha\lambda(1+\lambda)}{(1-\lambda)^3}(f_1-f_2)^2 \tag{42}$$

Given this method of describing the undermodelling, we next examine how it impacts on the estimation error $\hat{\theta}_N-\theta_0$. That is, in the previous sections, we considered the errors to be produced only by a disturbance sequence $\{\nu_k(\omega)\}$:

$$\begin{aligned} y_k &= G(q^{-1},\theta_0)u_k+\nu_k(\omega) && (43)\\ &= \phi_k^T\theta_0+\nu_k(\omega) && (44)\end{aligned}$$

Now we have an additional error inducing factor due to the undermodelling:

$$\begin{aligned} y_k &= G(q^{-1},\theta_0)u_k+G_\Delta(q^{-1},\omega)u_k+\nu_k(\omega) && (45)\\ &= \phi_k^T\theta_0+\psi_k^T\eta(\omega)+\nu_k(\omega) && (46)\end{aligned}$$

where

$$\psi_k^T \triangleq [u_{k-1},u_{k-1},\cdots,u_{k-L}] \tag{47}$$

since (39) requires an FIR description for the undermodelling component. If we now adopt a convenient vectorized notation:

$$\begin{aligned} \Phi^T &\triangleq [\phi_1,\cdots,\phi_N] && (48)\\ \Psi^T &\triangleq [\psi_1,\cdots,\psi_N] && (49)\\ Y^T &\triangleq [y_1,\cdots,y_N] && (50)\\ V^T(\omega) &\triangleq [\nu_1(\omega),\cdots,\nu_N(\omega)] && (51)\\ \eta^T(\omega) &\triangleq [\eta_1(\omega),\cdots,\eta_L(\omega)] && (52)\end{aligned}$$

then for an observed N point data sequence (47) leads us to

$$Y=\Phi\theta_0+\Psi\eta(\omega)+V(\omega) \tag{53}$$

with the least squares estimate given as:

$$\hat{\theta}_N=\left(\Phi^T\Phi\right)^{-1}\Phi^T Y \tag{54}$$

so

$$\left(\hat{\theta}_N-\theta_0\right)(\omega)=\left(\Phi^T\Phi\right)^{-1}\Phi^T\Psi\eta(\omega)+\left(\Phi^T\Phi\right)^{-1}\Phi^T V(\omega) \tag{55}$$

If we assume the function $\eta(\omega)$ representing the undermodelling to be independent of that representing the disturbance $V(\omega)$ then it is straightforward to conclude that

$$\mathrm{Cov}\{\tilde{\theta}\}=\int_\Omega \tilde{\theta}(\omega)\tilde{\theta}^T(\omega)d\mathbb{P}_{\tilde{\theta}}=\left(\Phi^T\Phi\right)^{-1}\Phi^T\left(\Psi C_\eta\Psi^T+C_\nu\right)\Phi\left(\Phi^T\Phi\right)^{-1} \tag{56}$$

where if we adopt the example assumptions given in (37) together with its consequence (41) we have

$$C_\eta=\int_\Omega \eta(\omega)\eta^T(\omega)d\mathbb{P}_\Delta=\begin{bmatrix}\alpha\lambda & & \\ & \ddots & \\ & & \alpha\lambda^L\end{bmatrix} \tag{57}$$

and

$$C_\nu=\int_\Omega V(\omega)V^T(\omega)d\mathbb{P}_\nu \tag{58}$$

This allows us to develop, as in the previous section, a feasible parameter set $\mathcal{D}$ that now is valid no only over an ensemble of possible bounded disturbances modelled by $\nu_k(\omega)$, but also over an ensemble of possible undermodellings indexed by $\omega \in \Omega$. This information can be translated into estimated frequency responses with frequency domain error bounds . This development is straighforward since by (39):

$$G_\Delta(e^{-j2\pi f}, \omega) = \Pi(e^{-j2\pi f})\eta(\omega) \tag{59}$$

where

$$\Pi(e^{-j2\pi f}) = \left[e^{-j2\pi f}, \cdots, e^{-j2\pi fL}\right] \tag{60}$$

and

$$\begin{aligned} G(e^{-j2\pi f}, \theta) &= \Lambda(e^{-j2\pi f})\theta & (61)\\ \Lambda(e^{-j2\pi f}) &= \left[\mathcal{B}_1(e^{-j2\pi f}), \cdots, \mathcal{B}_p(e^{-j2\pi f})\right] & (62) \end{aligned}$$

so that

$$\left(G_T(e^{-j2\pi f}) - G(e^{-j2\pi f}, \hat{\theta}_N)\right)(\omega) = \Lambda(e^{-j2\pi f})(\theta_0 - \hat{\theta}_N) + \Pi(e^{-j2\pi f})\eta(\omega) \tag{63}$$

The final step of deriving error bounds on the left hand side of (63) is provided by the following theorem.

Theorem 3. *Define*

$$\begin{aligned} \tilde{g}(e^{-j\omega}) &\triangleq \begin{bmatrix} Re\left\{G_T(e^{-j2\pi f}) - G(e^{-j2\pi f}, \hat{\theta}_N)\right\} \\ Im\left\{G_T(e^{-j2\pi f}) - G(e^{-j2\pi f}, \hat{\theta}_N)\right\} \end{bmatrix} & (64)\\ Q &\triangleq \left(\Phi^T\Phi\right)^{-1}\Phi^T & (65)\\ \Upsilon &\triangleq \begin{bmatrix} Q(\Psi C_\eta \Psi^T + C_\nu)Q^T & -Q\Psi C_\eta \\ -C_\eta\Psi^T Q^T & C_\eta \end{bmatrix} & (66)\\ \Gamma(e^{-j2\pi f}) &\triangleq \begin{bmatrix} Re\{\Lambda(e^{-j2\pi f}), & \Pi(e^{-j2\pi f})\} \\ Im\{\Lambda(e^{-j2\pi f}), & \Pi(e^{-j2\pi f})\} \end{bmatrix} & (67) \end{aligned}$$

Then

$$\int_\Omega \tilde{g}(e^{-j2\pi f})\tilde{g}(e^{-j2\pi f})^T d\mathbb{P}_{\tilde{g}} \triangleq P_{\tilde{g}} = \Gamma(e^{-j2\pi f})\Upsilon\Gamma^T(e^{-j2\pi f}) \tag{68}$$

Proof. See appendix C. □□□

Note that, with this formulation,

$$\int_\Omega \left|G(e^{-j2\pi f}, \hat{\theta}_N(\omega)) - G_T(e^{-j2\pi f})\right|^2 (\omega) d\mathbb{P} = \mathrm{Tr}\left\{P_{\tilde{g}}\right\} \tag{69}$$

but that the provision of $P_{\tilde{g}}$ allows phase error information to be provided as well. Specifically, motivated by Theorem 2 and the earlier simulation examples we may assume the density function for $\hat{\theta}_N - \theta_0$ to be approximately Gaussian. If we also specify the density $f_\Delta(\eta)$ to be of Gaussian shape then the density function for $\tilde{g}$ will be Gaussian with covariance $P_{\tilde{g}}$. It is then well known that the quantity

$$\tilde{g}(e^{-j2\pi f})^T P_{\tilde{g}}^{-1}\tilde{g}(e^{-j2\pi f}) \tag{70}$$

will have a χ_2^2 density function and this may be used to calculate feasible frequency response sets that will be ellipsoidal and carry not only the magnitude information of (69) but also phase information as well. We will illustrate this discussion with a continuation of our multivariable simulation example.

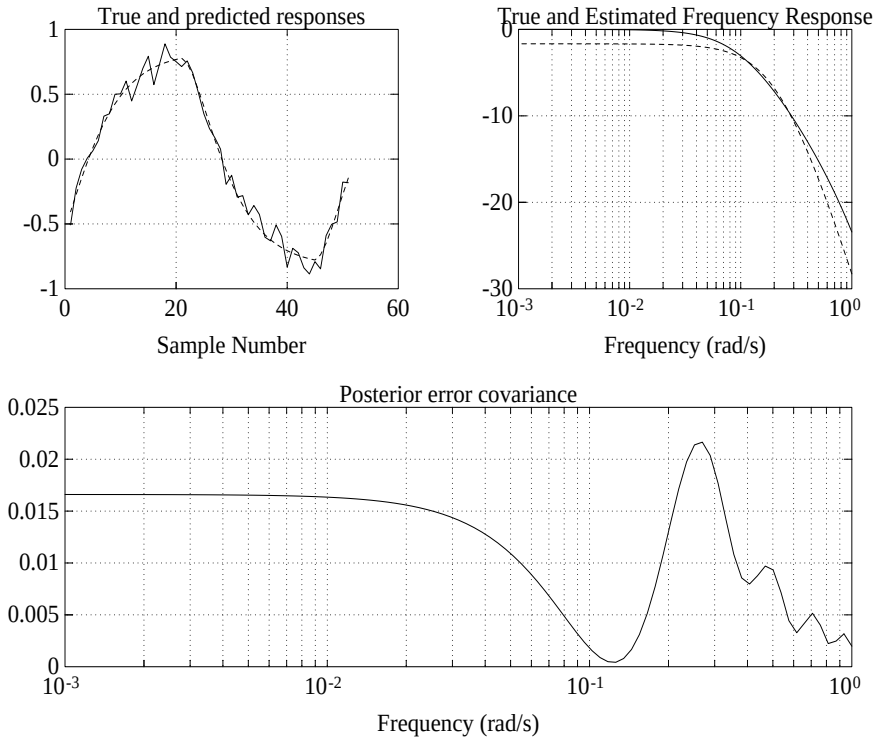

Figure 6: Multivariable simulation results presented in the frequency domain. Top right plot is true response and that of the model, as was shown in figure 3. Top left plot is the true frequency reponse (solid) versus that of the model (dashed). The discrepency indicates the presence of undermodelling. The bottom left plot shows the estimate of the frequency repsonse magnitude error calculated by the stochastic embedding method.

5 Multivariable Simulation Revisited

We return to the multivariable simulation example of the previous section. Now however we view the estimation results in the frequency domain as shown in figure 6. The top left plot shows the observed noise corrupted data and the response of the least squares estimated model parameterised by (34). The top right plot shows a comparison of the true and estimated frequency responses as Bode Magnitude plots. Note that these latter plots indicate that there is some undermodelling involved in our estimate, as is to be expected given the restricted complexity of the fixed denominator model structure (33). If we embed our description of this undermodelling in a description (37),(38) where

$$\alpha = 1.0 \qquad \lambda = 0.2 \tag{71}$$

then this corresponds to a magnitude error assumption at d.c of

$$\int_{\Omega} |G_{\Delta}(1,\omega)|^2 \, d\mathbb{P}_{\Delta} = \frac{\alpha\lambda}{1-\lambda} = 0.25 \tag{72}$$

and a Lipschitz smoothness assumption over the ensemble of possible undermodellings as

$$\int_{\Omega} \left| G_{\Delta}(e^{-j2\pi f_1},\omega) - G_{\Delta}(e^{-j2\pi f_2},\omega) \right|^2 d\mathbb{P}_{\Delta} \leq 0.31 \times 4\pi^2 (f_1 - f_2)^2 \tag{73}$$

The resultant quantification of magnitude estimation error given by the previous theorem and (69) is shown in the bottom plot of figure 6. Note that it shows highly discriminatory information

about frequency response errors. Note in particular, as predicted in [3], the estimation error is minimised at the fundamental ($0.02 * 2\pi$) frequency of the input signal spectrum and then at the harmonics (odd multiples of the fundamental). That this quantification of error is meaningful, in that it truly does overbound the error is shown in figure 7 where the true and estimated Nyquist

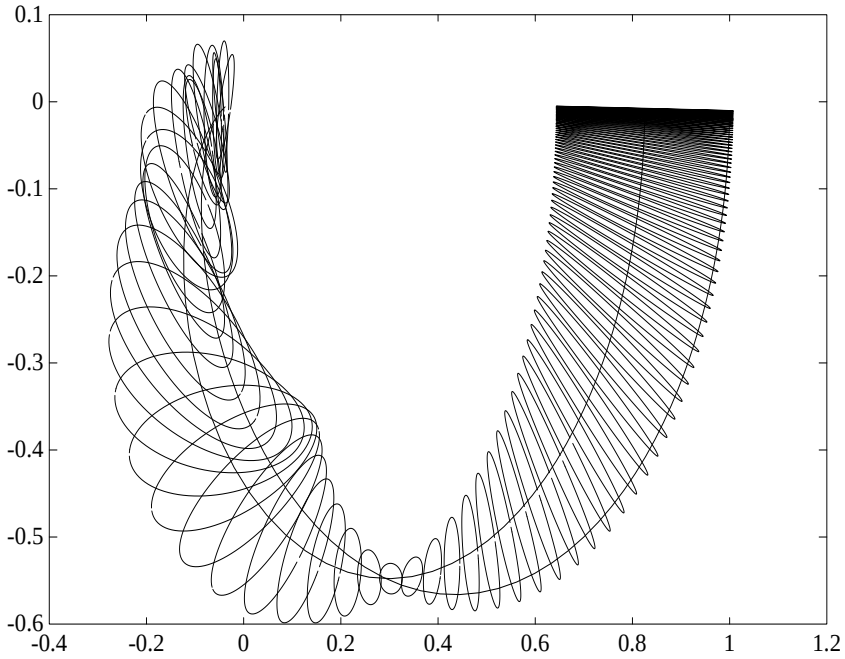

Figure 7: Stochastic Embedding overbounds together with true and estimated Nyquist Plots

plots are shown together with the error bounding ellipses

$$\tilde{g}(e^{-j2\pi f})^T P_{\tilde{g}}^{-1} \tilde{g}(e^{-j2\pi f}) \leq 4 \tag{74}$$

This plot shows that our error bounds are valid whilst being highly informative. If more conservative bounds are required the right hand side of (74) may be increased to some value greater than 4.

To evaluate the quality of the stochastic embedding paradigm for providing error bounds, we will compare it to a bounded error approach. There are many available in the literature, [29, 28] give examples and references. We chose to use the method proposed in [37, 2] since it will tie in nicely with the multivariable bounded error simulation we have already run. The idea in [37, 2] is to use bounded error estimation theory to estimate both θ **and** η. However, prior information is used to bound η as closely as possible before the estimation experiment. Such prior information discussed in [37, 2] may be of many forms. The frequency domain ones fit into the magnitude and Lipschitz smoothness ones we have discussed for our stochastic embedding approach although of course in the context of bounded error estimation the assumptions are not taken over the ensembles of undermodelling as we suggest. This information, together with prior information about θ_0 is used to form the initial matrix P_0 used in the algorithm developed by Fogel and Huang in [7]. The bound σ_t is specified, for example, by the induced norm arguments discussed in [28, 27]. The algorithm of Fogel and Huang is then used to recursively update a feasible parameter set for

$$\Theta^T \triangleq \left[\theta^T, \eta^T\right] \tag{75}$$

That is, the algorithm provides after N iterations a matrix P_N such that the true parameter lies in the set

$$(\Theta - \Theta_0)^T P_N^{-1} (\Theta - \Theta_0) < 1 \tag{76}$$

To use these parameter space bounds to provide frequency domain bounds the following result presented in [37] is required

Lemma 1. *Let $x \in \mathbb{R}^n, P \in \mathbb{R}^{n \times n}$ be positive definite, and assume that*

$$x^T P^{-1} x \leq 1 \tag{77}$$

Consider $y = Ax$, where $y \in \mathbb{R}^p, p \leq n$, $A \in \mathbb{R}^{p \times n}$ and A has full rank. Then

$$y^T (APA^T)^{-1} y \leq 1 \tag{78}$$

Proof. See [37]. □□□

The result is applied in our context by noting that

$$\Delta G(e^{-j2\pi f}) \triangleq \begin{bmatrix} \mathrm{Re}\left\{ G_T(e^{-j2\pi f}) - G(e^{-j2\pi f}, \hat{\Theta}_N) \right\} \\ \mathrm{Im}\left\{ G_T(e^{-j2\pi f}) - G(e^{-j2\pi f}, \hat{\Theta}_N) \right\} \end{bmatrix} = \Gamma(e^{-j2\pi f})(\Theta_0 - \hat{\Theta}_N) \tag{79}$$

so combining this with the bound on $\hat{\Theta}_N - \Theta_0$, (76) and using the lemma we get the frequency domain bound

$$\Delta G^T \left[\Gamma(e^{-j2\pi f}) P_N \Gamma(e^{-j2\pi f})^* \right]^{-1} \Delta G \leq 1 \tag{80}$$

We employed this method of Wahlberg and Ljung on the simulation example used to illustrate the stochastic embedding bounds. We made the following choices which correspond roughly to earlier assumptions about the noise and undermodelling:

$$\sigma_t = 0.1 \qquad P_0 = \begin{bmatrix} I_{2\times 2} & 0 \\ 0 & 0.05 I_{15\times 15} \end{bmatrix} \tag{81}$$

The results we obtained are shown in figure 8. The top left and right plots show the evolution of the volume of the bounding error ellipsoid, and the weighting sequence $\{\rho_t\}$. We see that the data was informative for updating the feasible parameter set. The quantified magnitude response error obtained via (80) is shown in the bottom of the figure. As can be seen it does not show the fine discriminatory structure that was obtained via stochastic embedding in figure (6). The information obtained from (80) is shown in the Nyquist diagram form in figure 9. This plot is very similar to the results displayed in [37, 2]. Again the fine discriminatory structure displayed by the stochastic embedding approach is lost. The results are very conservative.

6 Estimation of the Parameterization of the Noise and Undermodelling

We conclude the paper what we believe to be one of the most compelling reasons why the stochastic embedding approach to bounded error estimation should be considered over the set estimation approaches; namely the possibility of estimating the parameters α, λ, δ from the data. Estimating δ would amount to estimating the variance σ_ν^2 of the measurement 'noise' and then noting that for a uniform distribution $\sigma_\nu^2 = \delta^2/3$. This idea is well known in stochastic estimation theory and is usual accomplished by considering the sample second moment of the prediction residuals. In this section we discuss the more unusual issue of estimating parameters of the undermodelling α, λ from the prediction residuals as well. We give only a brief presentation here. For a fuller discussion see [9, 1]. To begin with, we define the N-vector of residuals

$$\varepsilon \triangleq Y - \Phi\hat{\theta} \tag{82}$$

$$= [I - \Phi(\Phi^T\Phi)^{-1}\Phi^T]Y \triangleq PY. \tag{83}$$

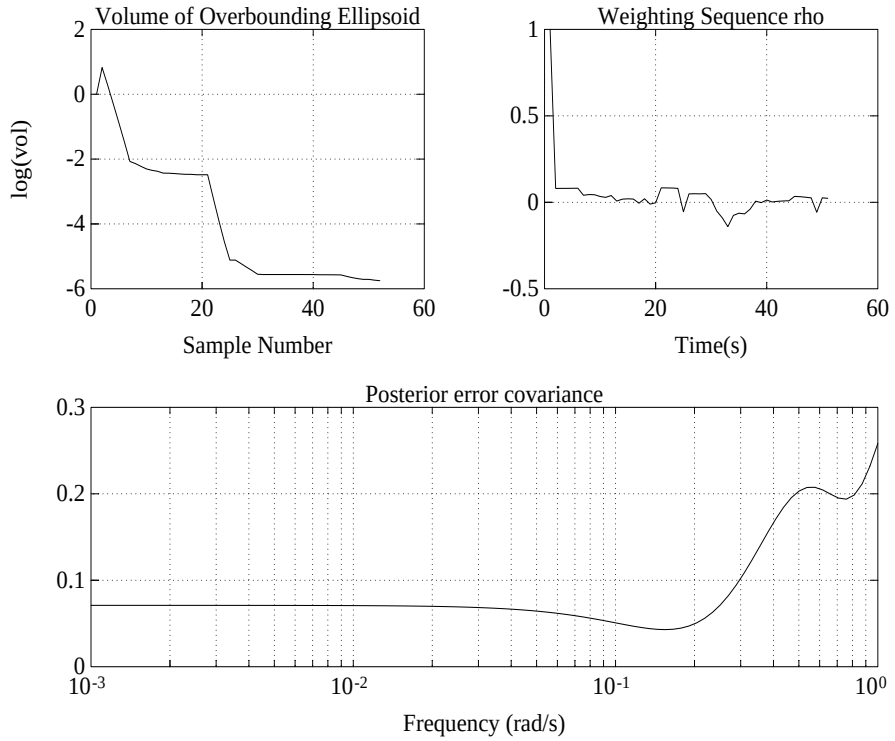

Figure 8: Quantifying the error in estimated frequency responses using bounded error estimation theory.

The matrix in (83) has rank N-p. Therefore ε has a singular distribution of rank N-p. To obtain a new full rank data vector, we represent ε in a new coordinate system that forms a basis for the space orthogonal to the columns of Φ. Let R be any matrix whose columns span the subspace orthogonal to the columns of Φ. One way of constructing such an R is to take any $N-p$ independent linear combinations of the columns of P. Now define $W \in \mathbb{R}^{N-p}$ as follows :

$$W \triangleq R^T \varepsilon. \tag{84}$$

Then W has a nonsingular distribution and, by the construction of R,

$$W = R^T Y = R^T \Psi \eta + R^T V. \tag{85}$$

Since R^T and Ψ depend on the input signal only, we observe that W is the sum of two independent random vectors whose probability density functions are computable functions of the unknown parameters $\alpha, \lambda, \sigma_\nu^2$. We can therefore compute the probability density function of W, conditioned on the observed input data vector U, and on $\xi^T \triangleq (\alpha, \lambda, \sigma_\nu^2)$. We denote the corresponding likelihood function by $\mathcal{L}(W \mid U, \xi)$. Maximizing this likelihood function yields the desired estimate for the unknown parameters :

$$\hat{\xi} = \arg\max_{\xi} \{\mathcal{L}(W \mid U, \xi)\} \tag{86}$$

We investigate the properties of $\hat{\xi}$ for Gaussian assumptions on f_Δ and f_ν as follows. Consider the special case of the stochastic embedding assumptions being as in (41)

$$\eta \sim \mathcal{N}(0, C_\eta(\alpha, \lambda)) \tag{87}$$

$$C_\eta(\alpha, \lambda) = \operatorname*{diag}_{1 \le k \le L} \left\{\alpha \lambda^k\right\} \tag{88}$$

In addition, assume that the noise $\{\nu_k\}$ is an independent sequence which is also independent of η and with $\nu_k \sim \mathcal{N}(0, \sigma_\nu^2)$. The Gaussian assumptions on the distributions f_Δ and f_ν give the log

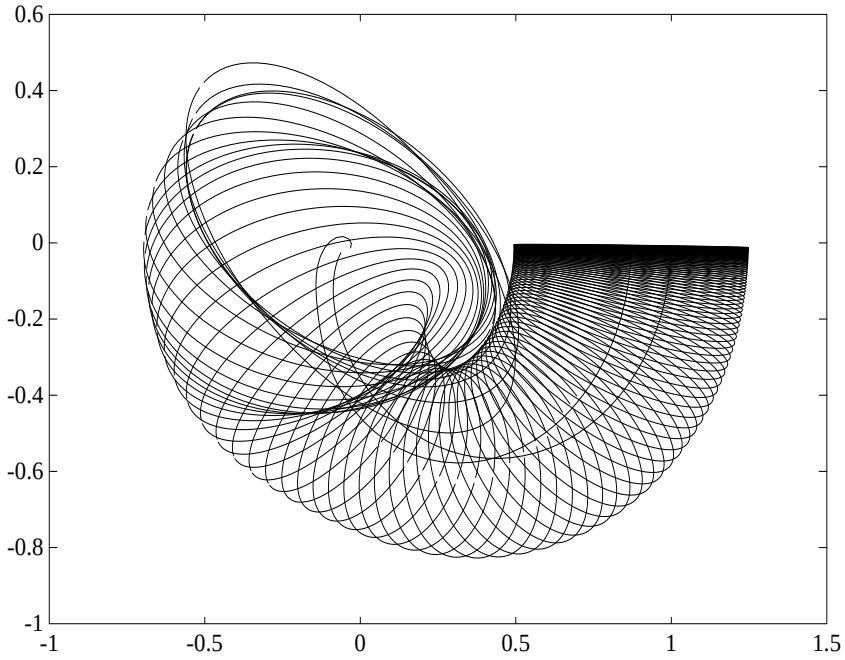

Figure 9: Quantifying the error in estimated frequency responses using bounded error estimation theory.

likelihood function $l(W \mid U, \xi)$ for the observed data as

$$l(W \mid U, \xi) = -\frac{1}{2} \ln \det \Sigma - \frac{1}{2} W^T \Sigma^{-1} W + \text{constant}, \tag{89}$$

where

$$\Sigma = R^T \Psi C_\eta(\alpha, \lambda) \Psi^T R + \sigma_\nu^2 R^T R \tag{90}$$

$$C_\eta(\alpha, \lambda) = \text{diag}\,\{\alpha\lambda, \alpha\lambda^2, \ldots, \alpha\lambda^L\}. \tag{91}$$

That this estimation scheme does in fact work is proved in [1]. Specifically it is proved that if the regressors Φ,Ψ are orthogonalised by Gram-Schmidt, then provided the dimension L of η grows at a rate proportional to the data length N of $\ln N$, then we will have

$$\hat{\xi}_N \xrightarrow{a.s.} \xi_0 \qquad N \to \infty \tag{92}$$

and if P_L is the Fisher information matrix implied by the Likelihood function then we also have

$$\sqrt{N} P_L^{1/2}[\xi_0](\hat{\xi}_N - \xi_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I) \quad N \to \infty. \tag{93}$$

Therefore, the maximum likelihood method of estimation is asymptotically efficient. These results show that if we represent undermodelling via a random variable description, then it is possible to estimate the parameters α,λ that describe the class of undermodellings of which we observe a realisation. At the same time we can measure σ_ν^2 which parameterises a class of disturbances of which we observe a realisation. We can use use these estimates to provide error bounds which apply over the classes of disturbances and undermodellings, and these error bounds are obtained without making any quantitative assumptions. We need only make qualitative assumptions on the nature of undermodelling and disturbances. We believe this dependence only on qualitative assumptions is highly desirable for the purposes of robust control design.

7 Conclusion

In this paper we have examined the rapproachment between a stochastic and deterministic framework for evaluating feasible parameter sets from noise corrupted data. We have shown that in the case of linear regression form model structures the deterministic approach is subsumed by a Bayesian formulation in the stochastic framework. We have also examined the utility of deriving feasible parameter sets from confidence regions arising from application of the central limit theorem in a stochastic setting. In the case where a typical sample path for the disturbance wanders around zero, these latter sets were shown to not only be highly informative, but also to be insensitive to prior quantitative knowledge of disturbances. We concluded by showing how the problem of deriving error bounds accounting for undermodelling can be cast in a stochastic framework. This approach has the attendant advantages of allowing parameters describing the class of undermodellings of which we observe a realisation to be estimated from the available data. We conclude that the rapproachment between the two schools of thought is characterised by a tradeoff between the required accuracy of qualitative knowledge and quantitative knowledge. In a deterministic framework, the nature of a disturbance sequence need not be known, but an accurate quantitative bound is necessary. In a stochastic framework one needs to know the nature of a typical disturbance sample path, but parameters quantifying the ensemble from which it was drawn can be estimated from the data.

Appendix A Proof of Theorem 1

Proof.

$$|y_k - \phi_k^T \theta| < |\nu_k| \tag{A.1}$$

so feasible $\theta \in \mathbb{R}^p$ satisfy

$$\phi_k^T \theta \in [y_k - |\nu_k|, y_k + |\nu_k|] \tag{A.2}$$

For the purposes of forming $\mathcal{D}$, the best possible realizations are $y_k = \phi_k^T \pm \delta$. Therefore, the best bounds obtainable for one sample are given by the hyperplane delineated region in θ space:

$$\mathcal{D}_k \triangleq \left\{\theta \in \mathbb{R}^p : \phi_k^T \theta \in [\phi_k^T \theta_0 + \delta - |\nu_k|, \phi_k^T \theta_0 - \delta + |\nu_k|]\right\} \tag{A.3}$$

But our conservative assumption is $|\nu_k| < \mu$ to give

$$\mathcal{D}_k \triangleq \left\{\theta \in \mathbb{R}^p : \phi_k^T (\theta - \theta_0) \in [\delta - \mu, \mu - \delta]\right\} \tag{A.4}$$

So $\tilde{\theta} \triangleq \theta - \theta_0$ lies between a pair of hyperplanes perpendicular to ϕ_k. Also, if β lies on one hyperplane, then $-\beta$ lies on the other, so the distance between the hyperplanes in the direction β is $2\|\beta\|$. But by the Cauchy-Schwartz inequality:

$$|\mu - \delta| = |\phi_k^T \beta| < \|\phi_k\| \|\beta\| \quad \text{for some } \beta \in \mathbb{R}^p \tag{A.5}$$

so a lower bound on the distance between the two hyperplanes forming $\mathcal{D}_k$ is

$$2\|\beta\| \geq \frac{2|\mu - \delta|}{\|\phi_k\|} \geq \frac{2(\mu - \delta)}{\sigma_\phi} \tag{A.6}$$

But this lower bound is valid $\forall k \in \mathbb{N}$ and

$$\mathcal{D}_N = \bigcap_{k=1}^{N} \mathcal{D}_k \tag{A.7}$$

so $\mathcal{D}_N$ is underbounded by a sphere of diameter d

$$d \geq \frac{2(\mu - \delta)}{\sigma_\phi} \tag{A.8}$$

regardless of how large we make N. □□□

Appendix B Proof of Theorem 2

Proof. The Hermite polynomials defined by

$$H_n(x) = x^n - \left(\begin{array}{c} n \\ 2 \end{array} \right) x^{n-2} + 1.3. \left(\begin{array}{c} n \\ 4 \end{array} \right) x^{n-4} - 1.3.5. \left(\begin{array}{c} n \\ 6 \end{array} \right) x^{n-6} + \cdots \tag{B.1}$$

have the orthogonality property

$$\int_{-\infty}^{\infty} \frac{1}{\sigma} \Phi \left(\frac{x}{\sigma} \right) H_n \left(\frac{x}{\sigma} \right) H_m \left(\frac{x}{\sigma} \right) dx = \left\{ \begin{array}{ll} n! & ; m = n \\ 0 & ; m \neq n \end{array} \right. \tag{B.2}$$

so if we write $f_{\tilde{\theta}}(\theta)$ as a Fourier series in $H_k(x)$:

$$f_{\tilde{\theta}}(\theta) = \frac{1}{\sigma} \Phi \left(\frac{\theta}{\sigma} \right) \sum_{k=1}^{\infty} \alpha_n H_n \left(\frac{\theta}{\sigma} \right) \tag{B.3}$$

then we have

$$\int_{-\infty}^{\infty} f_{\tilde{\theta}}(\theta) H_m \left(\frac{\theta}{\sigma} \right) d\theta = \sum_{k=1}^{\infty} \alpha_n \int_{-\infty}^{\infty} \frac{1}{\sigma} \Phi \left(\frac{\theta}{\sigma} \right) H_n \left(\frac{\theta}{\sigma} \right) H_m \left(\frac{\theta}{\sigma} \right) d\theta \tag{B.4}$$

to give

$$\alpha_n = \frac{1}{n!} \int_{-\infty}^{\infty} f_{\tilde{\theta}}(\theta) H_n \left(\frac{\theta}{\sigma} \right) d\theta \tag{B.5}$$

Then using (B.1) gives:

$$\alpha_0 = \int_{-\infty}^{\infty} f_{\tilde{\theta}}(\theta) d\theta = 1 \tag{B.6}$$

$$\alpha_1 = \int_{-\infty}^{\infty} \frac{\theta}{\sigma} f_{\tilde{\theta}}(\theta) d\theta = 0 \quad \text{by Lemma 15} \tag{B.7}$$

$$\alpha_2 = \frac{1}{2} \int_{-\infty}^{\infty} \left(\frac{\theta^2}{\sigma^2} - 1 \right) f_{\tilde{\theta}}(\theta) d\theta = 0 \quad \text{by Definition of } \sigma \tag{B.8}$$

$$\alpha_3 = \frac{1}{6} \int_{-\infty}^{\infty} \left(\frac{\theta^3}{\sigma^3} - 3\frac{\theta}{\sigma} \right) f_{\tilde{\theta}}(\theta) d\theta = 0 \quad \text{by symmetry of } f_{\nu} \tag{B.9}$$

$$\alpha_4 = \frac{1}{24} \int_{-\infty}^{\infty} \left(\frac{\theta^4}{\sigma^4} - 6\frac{\theta^2}{\sigma^2} + 3 \right) f_{\tilde{\theta}}(\theta) d\theta = \frac{1}{24} \left(\frac{m_4}{\sigma^4} - 3 \right) \tag{B.10}$$

where

$$m_4 \triangleq \int_{-\infty}^{\infty} \theta^4 f_{\tilde{\theta}}(\theta) d\theta \tag{B.11}$$

We elect to terminate the Fourier expansion for $f_{\tilde{\theta}}(\theta)$ at α_4 for the purposes of our approximation. We are justified in this by knowing that [5]:

$$\lim_{n \to \infty} \alpha_n = 0 \tag{B.12}$$

and that the Fourier co-efficients minimize the L_2 error between $f_{\tilde{\theta}}(\theta)$ and our approximation [5]. This gives:

$$f_{\tilde{\theta}}(\theta) \approx \frac{1}{\sigma} \Phi \left(\frac{\theta}{\sigma} \right) \left[1 + \frac{1}{24} \left(\frac{m_4}{\sigma^4} - 3 \right) \left(\left(\frac{\theta}{\sigma} \right)^4 - 6 \left(\frac{\theta}{\sigma} \right)^2 + 3 \right) \right] \tag{B.13}$$

Now by Lemma B.1

$$m_4 = \left(\frac{\delta}{P_N} \right)^4 \left[\frac{1}{5} \sum_{k=1}^{N} \phi_k^4 + \frac{2}{3} \sum_{k=1}^{N} \sum_{j>k}^{N} \phi_k^2 \phi_j^2 (1 - 3 \ln \phi_k^2 \phi_j^2) \right] \tag{B.14}$$

also

$$\sigma^2 = \frac{\delta^2}{3P_N} \tag{B.15}$$

Substituting this into (B.13) then gives the result. □□□

Lemma B.1.

$$m_4 = \left(\frac{\delta}{P_N}\right)^4 \left[\frac{1}{5}\sum_{k=1}^{N}\phi_k^4 + \frac{2}{3}\sum_{k=1}^{N}\sum_{j>k}^{N}\phi_k^2\phi_j^2(1 - 3\ln\phi_k^2\phi_j^2)\right] \tag{B.16}$$

Proof.

$$\tilde{\theta} = \frac{1}{P_N}\sum_{k=1}^{N}\phi_k\nu_k \qquad P_N \triangleq \sum_{k=1}^{N}\phi_k^2 \tag{B.17}$$

So

$$m_4 = \int_{-\infty}^{\infty}\theta^4 f_{\tilde{\theta}}(\theta)d\theta \tag{B.18}$$

$$= \frac{1}{P_N^4}\sum_{k=1}^{N}\phi_k^4\int_\Omega \nu_k^4(\omega)d\mathbb{P} + \frac{6}{P_N^4}\sum_{k=1}^{N}\sum_{j>k}^{N}\phi_k^2\phi_j^2\int_\Omega \nu_k^2(\omega)\nu_j^2(\omega)d\mathbb{P} \tag{B.19}$$

so by Lemma B.2

$$m_4 = \left(\frac{\delta}{P_N}\right)^4 \left[\frac{1}{5}\sum_{k=1}^{N}\phi_k^4 + \frac{2}{3}\sum_{k=1}^{N}\sum_{j>k}^{N}\phi_k^2\phi_j^2(1 - 3\ln\phi_k^2\phi_j^2)\right] \tag{B.20}$$

□□□

Lemma B.2. *Let $\nu_k(\omega)$ have the density function given in (4) and let $\phi_k \in \mathbb{R}$. Then*

$$\phi_k^2\phi_j^2\int_\Omega \nu_k^2(\omega)\nu_j^2(\omega)d\mathbb{P} = \begin{cases} \frac{\phi_k^4\delta^4}{5} & ;k = j \\ \frac{\phi_k^2\phi_j^2\delta^4}{9}(1 - 3\ln\phi_k^2\phi_j^2) & ;k \neq j \end{cases} \tag{B.21}$$

Proof. $\underline{k = j}$ The characteristic function $\varphi(s)$ for $\phi_k\nu_k$ is [38]:

$$\varphi(s) = \frac{\sin\phi_k\delta s}{\phi_k\delta s} \tag{B.22}$$

so we have [38]:

$$\phi_k^2 4\int_\Omega \nu_k^4(\omega)d\mathbb{P} = \lim_{s\to 0}\frac{\partial^4}{\partial s^4}\frac{\sin\phi_k\delta s}{\phi_k\delta s} = \frac{\phi_k^4\delta^4}{5} \tag{B.23}$$

$\underline{k \neq j}$ By Lemma B.3, the density function for $\nu_k(\omega)\nu_j(\omega)$ is

$$f_{\nu_k\nu_j}(x) = \frac{1}{2\delta^2}\ln\frac{\delta^2}{|x|} \quad ;|x| < \delta^2 \tag{B.24}$$

So

$$\phi_k^2\phi_j^2\int_\Omega \nu_k^2(\omega)\nu_j^2(\omega)d\mathbb{P} = \int_{-\phi_k\phi_j\delta^2}^{\phi_k\phi_j\delta^2}\frac{x^2}{2\phi_k\phi_j\delta^2}\ln\frac{\delta^2}{|\phi_k\phi_j x|}dx = \frac{\phi_k^2\phi_j^2\delta^4}{9}(1 - 3\ln\phi_k^2\phi_j^2) \tag{B.25}$$

□□□

Lemma B.3. *Let $X(\omega)$ and $Y(\omega)$ be independent and both have the same density function (4). Then if $Z = XY$, the density function $f_Z(z)$ is*

$$f_Z(z) = \frac{1}{2\delta^2} \ln \frac{\delta^2}{|z|} \quad \textit{for } |z| \leq \delta^2 \tag{B.26}$$

Proof. Define the Mellin Transform $F_M(s)$ of a positive function $f(x)$ as:

$$F_M(s) = \int_0^\infty x^{s-1} f(x) dx \qquad s \in \mathbb{C} \tag{B.27}$$

Split $f_X(x) = f_Y(y)$ into a positive function and a negative function:

$$f_X(x) = f_X^+(x) + f_X^-(x) \tag{B.28}$$

Then

$$F_X^+(s) = \frac{1}{2\delta} \int_0^\delta x^{s-1} dx = \frac{\delta^s}{2\delta s} \tag{B.29}$$

This gives [34]

$$F_Z^+(s) = 2\left[F_X^+(s)\right]^2 = \frac{1}{2\delta^2}\left(\frac{\delta^{2s}}{s^2}\right) \tag{B.30}$$

Then $f_Z^+(z)$ is given by the inverse Mellin Transform [34]:

$$2\delta^2 f_Z^+(z) \frac{1}{2\pi j} \int_{c-j\infty}^{c+j\infty} \frac{z^{-s}\delta^{2s}}{s^2} ds \tag{B.31}$$

which can be evaluated by residues as:

$$\begin{aligned} f_Z^+(z) &= \frac{1}{2\delta^2} \lim_{s\to 0} \frac{\partial}{\partial s} z^{-s}\delta^{2s} && \text{(B.32)} \\ &= \frac{1}{2\delta^2} \ln \frac{\delta^2}{|z|} \qquad z \in [0, \delta^2] && \text{(B.33)} \end{aligned}$$

By symmetry $f_Z^+(z) = f_Z^-(-z)$ therefore

$$f_Z(z) = \frac{1}{2\delta^2} \ln \frac{\delta^2}{|z|} \qquad z \in [-\delta^2, \delta^2] \tag{B.34}$$

□□□

Appendix C Proof of Theorem 3

Proof. From (63):

$$G_T(e^{-j2\pi f}) - G(e^{-j2\pi f}, \hat{\theta}_N) = \Lambda(\theta_0 - \hat{\theta}_N) + \Pi\eta \tag{C.1}$$

However by the definition of Q we have:

$$\theta_0 - \hat{\theta}_N = -Q\Psi\eta - QV \tag{C.2}$$

Furthermore

$$\tilde{g}(e^{-j\omega}) = \Gamma(e^{-j2\pi f})(\rho_0 - \hat{\beta}_N) \tag{C.3}$$

where

$$\begin{aligned} \hat{\beta}_N &\triangleq [\hat{\theta}_N^T, 0^T]^T && \text{(C.4)} \\ \beta_0 &\triangleq [\theta_0^T, \eta^T]^T && \text{(C.5)} \end{aligned}$$

Using (C.2) and remembering that V and η are independent then gives $\mathrm{Cov}\{\rho_0 - \hat{\rho}_N\} = \Upsilon$. This, combined with (C.3), then gives (68). □□□

REFERENCES

[1] B.M.Ninness, *Estimation of variance components in linear models with applications to stochastic embedding*, Technical Report EE9226, Department of Electrical and Computer Engineering, University of Newcastle, AUSTRALIA, (1992).

[2] B.Wahlberg, *On estimation of transfer function error bounds*, Proceedings of 1st ECC, 2 (1991), pp. 1378–1383.

[3] B.Wahlberg and L.Ljung, *Design variables for bias distribution in transfer function estimation*, IEEE Trans.Autom.Control,, AC-31 (1986), pp. 134–144.

[4] D. E. Catlin, *Estimation, Control, and the Discrete Kalman Filter*, Springer-Verlag, 1989.

[5] R. Deutsch, *System Analysis Techniques*, Prentice Hall Inc., Englewood Cliffs N.J., 1969.

[6] 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.

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

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

[9] G.C.Goodwin, M.Gevers, and B.M.Ninness, *Quantifying the error in estimated transfer functions with application to model order selection*, IEEE Transactions on Automatic Control., (July 1992).

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

[11] ——, *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.

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

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

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

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

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

[17] J.M.Krause and P.P.Khargonekar, *A comparison of classical stochastic estimation and deterministic robust estimation*, IEEE Transactions on Automatic Control, AC-37 (1992), pp. 994–1000.

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

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

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

[21] R. L. Maire, L. Valavani, M. Athans, and G. Stein, *A frequency domain estimator for use in adaptive control systems*, Automatica, 27 (1991), pp. 23–38.

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

[23] M.Milanese, *Robustness in Identification and Control*, Plenum Press, New York, 1989.

[24] 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.

[25] S. Mo and J. Norton, *Recursive parameter bounding algorithms which compute polytope bounds*, Proceedings of 12th IMACS World Congress, Paris, (1988).

[26] P.M.Mäkila, *Approximation of stable systems by laguerre filters*, Automatica, 26 (1990), pp. 333–345.

[27] R.C.Younce, *Identification with non-parametric uncertainty*, PhD thesis, University of Notre Dame, Indiana, 1989.

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

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

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

[31] R.L.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.

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

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

[34] M. Springer, *The Algebra of Random Variables*, John Wiley and Sons, New York, 1979.

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

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

[37] 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.

[38] D. Williams, *Probability with Martingales*, Cambridge University Press, Cambridge, UK, 1991.