Quantifying the Error in Estimated Transfer Functions with Application to Model Order Selection

Graham C. Goodwin, *Fellow, IEEE*, Michel Gevers, *Fellow, IEEE*, and Brett Ninness

Abstract—**Previous results on estimating errors or error bounds on identified transfer functions have relied upon prior assumptions about the noise and the unmodeled dynamics. This prior information took the form of parameterized bounding functions or parameterized probability density functions, in the time or frequency domain with known parameters. Here we show that the parameters that quantify this prior information can themselves be estimated from the data using a maximum likelihood technique. This significantly reduces the prior information required to estimate transfer function error bounds. We illustrate the usefulness of the method with a number of simulation examples.**

The paper concludes by showing how the obtained error bounds can be used for intelligent model order selection that takes into account both measurement noise and under-modeling. Another simulation study compares our method to Akaike's well-known FPE and AIC criteria.

I. Introduction

THE starting point of just about any robust control design principle is the assumption that the design engineer possesses not only a nominal plant model, but also precise knowledge of the uncertainty bounds around this nominal model. Both the nominal model and the uncertainty bounds are usually assumed to be given in the frequency domain, for example, in the form of a Nyquist plot of the nominal model with uncertainty bounds around it.

In many practical applications, the nominal model will be obtained as the result of an identification experiment. The need of robust control designers for uncertainty bounded model descriptions is viewed by members of the identification community as a major theoretical challenge, and it so happens that present-day identification theory is not able to deliver the uncertainty bounds that robust control designers require.

Bias Error and Variance Error

The errors in the estimated transfer functions have two components. The first component, often called variance error, is caused by the noise in the data that make up the particular realization that is used for identification purposes. The second component, often called the bias error, is caused by the fact that the parameterized model structure is, at best, a simplified (low order) version of the true system. That is, given the restricted complexity of the nominal model, there is no parameter value for which the nominal transfer function can equal the true transfer function at every frequency.

A key tool used for the computation of the first component, i.e., variance errors, is the Cramér–Rao lower bound on the estimated parameters. *In the case of exact model structure*, this tool produces reasonable variance error expressions for the estimated transfer functions; see, e.g., [17] and [8]. This variance error typically decreases like $1/N$, where N is the number of data. *In the case of restricted complexity model structures*, the parameters of the model have essentially no meaning: they converge to values that bear no connection with the parameters of the true system transfer function if such objects exist. The classical Cramér–Rao expression does not apply. However, recent work has produced an asymptotic procedure for the computation of variance errors on the model parameters in this situation [13].

We now turn to the estimation of the second component; bias errors in the case of restricted complexity models. *In the case of noiseless data*, this is essentially a trivial problem. Indeed, when there is no noise and given sufficient excitation, one can estimate as many parameters as there are data points, say N. If N is large enough, such a high-order model will be as close as desired to the true system, assuming it is linear. If a low-order model is then extracted for control design purposes, the exact bias in that low-order model, rather than just a bias error bound, can be computed by reference to the known high-order model.

The characterization of the bias error *in the case of a finite set of noisy data* is much more difficult. Asymptotically, of course, the same argument applies as in the noiseless case since the noise is averaged out, which allows for the estimation of very accurate high-order models. But our ambition in this paper is to handle the case of estimation of restricted complexity models from a *finite* noisy data record. The first results on a characterization of the bias error are due to Wahlberg and Ljung [28], who provide an implicit description of the bias error using Parseval's formula. This formula allows for an interesting qualitative discussion of the factors affecting bias, but it does not provide an explicit expression of the bias errors or a bound on it.

Manuscript received January 31, 1991; revised November 8, 1991. Paper recommended by Associate Editor at Large, M. P. Polis. This work was supported in part by the scholarship assistance of the Australian Telecommunications and Electronics Research Board.

G. C. Goodwin and B. Ninness are with the Department of Electrical and Computer Engineering, The University of Newcastle, Callaghan, Australia.

M. Gevers is with the Laboratoire d'Automatique, Dynamique et Analyse des Systèmes, Louvain-la-Neuve, Belgium.

IEEE Log Number 9200621.

To summarize our discussion so far, the estimation of the error (or of an error bound) on an identified model is only difficult, and indeed unsolved, in the case where both the model structure is of lower complexity than the true system, and where the data record is finite and noisy [32].

Error Quantification: Review of Existing Methodologies

The mainstream of thought to date has been to derive hard bounds on the transfer function error on the basis of assumed prior knowledge on the noise (a known distribution or a known hard bound) and of assumed prior magnitude and smoothness bounds on the unmodeled dynamics. For example, in [1], a nominal parametric model is fitted to the empirical transfer function estimate (ETFE), for which hard-error bounds are derived on assumptions of smoothness of the true frequency response and bounds on the Gibbs effect due to the finite data windowing used in calculating the ETFE. In [24] and [25], a similar approach is taken save that a Kalman filter is used to calculate the ETFE, and FIR models are then fitted to this in the frequency domain using the interpolation theory of Lagrange. In [14]–[16], [30], and [3] bounds are calculated in the parameter space, again based on assumptions on the smoothness of the true transfer function of the unmodeled dynamics. Compact supports for the distributions of stochastic components are also assumed and the resultant parameter space bounds are transformed to the frequency domain. Finally, in [29], the ideas of set membership estimation developed in [5], [6], and [23] are used to provide hard bounds in the parameter space, which are then transformed to the frequency domain.

The hard bounding approaches to quantifying errors on estimated transfer functions suffer from several drawbacks, leading to overly conservative error bounds. We shall address these limitations at the end of this Introduction.

This conservatism can be avoided by assuming a stochastic prior model for the distribution of the unmodeled dynamics with a distribution of noncompact support, as opposed to the uniform compact support distribution imposed by hard-bound models. Such stochastic description of the unmodeled dynamics is consistent with the stochastic prior model that is typically assumed for the noise. The idea of stochastic embedding was introduced in [12] and subsequently developed in [11], [21], [22], and [8]. In similar spirit to the hard-bounding work, it required prior specification of *likely* smoothness and magnitude parameters of the true-system frequency response, via a parameterized prior distribution with known parameters. This prior distribution was updated to a posterior one using the data and the prior noise distribution, and this gave confidence regions for the estimated frequency response.

Our New Contribution

The major criticism one could level at the methods we have briefly reviewed is that complete prior knowledge was assumed on the noise bound and on the magnitude and smoothness bound for the unmodeled dynamics in the hard-bound approach, or alternatively on the distribution of the noise and dynamics in the stochastic embedding approach. For example, a typical hard-bound prior assumption on the unmodeled dynamics transfer function is that the magnitude of its impulse response is bounded by a first-order exponential $\alpha\lambda^k$, while a typical stochastic prior assumption is that the variance of the unmodeled dynamics is bounded by $\alpha\lambda^k$. In both cases, α and λ had to be fully known. Our major new contribution in this paper is to show that, provided a parameterized structure is chosen for the prior assumption (e.g., a first-order decaying exponential for the variance of the unmodeled dynamics), then the parameters that specify this prior (α and λ in the above examples) can be estimated from the data using maximum likelihood. We believe this to be a significant step forward, since the required prior information now reduces to specifying *the structure* of some parameterized probability density function for the undermodeling and for the noise, while the parameters are left free to be estimated. As our simulations will show, the precise form of the structural assumption on the undermodeling does not appear to be essential, the only requirement being that the undermodeling transfer function be stable; a reasonable stance. *The requirement on prior information is thus reduced from a quantitative one to a qualitative one.* We believe this to be a major advantage. Simulation studies show the resultant confidence regions to be highly realistic and discriminatory.

We conclude the paper by showing how the error bounds obtained may be applied to the problem of model order selection with finite data. The optimal order is obtained by minimizing some suitable criterion of the total mean square error between the true transfer function $G_T(e^{-j\omega})$ and the estimated model $G(e^{-j\omega}, \hat{\theta}_N)$ based on N data. Depending on the application, the criterion could be, for example, the weighted integral of this error over all frequencies, or the supremum of this error weighted over frequency, or any other suitable criterion. We introduce a new criterion, called the generalized information criterion (GIC), which is based on minimizing the mean square output prediction error. We show that, in the presence of undermodeling and with finite data, this new criterion performs better than the classical final prediction error (FPE) and AIC criteria.

The mean square error between $G_T(e^{-j\omega})$ and $G(e^{-j\omega}, \hat{\theta}_N)$ is shown to be the sum of two terms; a bias term that decreases with model order and a variance term that increases with this order. The minimum overall model orders will therefore be well defined. It is important to understand that this optimal model order is also an increasing function of the number of data N, since the variance error contribution decreases with N. We should like to make it very clear that, contrary to popular belief, with finite noisy data the optimal model order is typically smaller than the exact model order if such an exact order exists, and that the traditional quest for a true model order on the basis of finite data is a misguided pursuit.

Why Stochastic Embedding, or the Soft-Versus-Hard-Bound Debate

We conclude this Introduction with a thorough motivation for our stochastic embedding approach. We believe the hard bounding approach to error quantification suffers from two key limitations. First, it is philosophically objectionable to

abruptly abandon a stochastic paradigm in favor of hard-bound models on the noise as soon as undermodeling becomes present. A hard-bound noise model is a very coarse (worst case) model for physical reality since every value within a compact domain is considered as likely as any other. A distribution, with noncompact support, is a model of reality in which the noise values are assumed to be on average centered around some mean value without precluding the possibility of the occasional outlier. This appears to be a much more reasonable model than the worst case assumption that the noise can be at the outlier value at every time.

Second, and precisely because the compact support assumptions on the distribution of the stochastic components have to include the occasional but unlikely outlier, the prior bounds will of necessity be very large. This in turn, implies that the resultant hard-error bounds on the transfer function will also be overly conservative. In addition, for the parameter space bounding methods, the membership sets are only overbounding approximations to the true sets, while the transformations from parameter space to transfer function domain are again overbounds and conservative. This conservatism is illustrated in the diagrams of [29].

By embedding the description of the undermodeling, and of course also that of the noise, in stochastic distributions having noncompact support, we avoid this conservatism. Of course, the resulting frequency-domain bounds then become confidence regions rather than hard bounds. However, we argue that this is appropriate since prior assumptions can never be specified with absolute certainty. Indeed, we also 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. Therefore, control engineers always work with a tradeoff of uncertainty versus performance. Consequently, while the estimation community has a mandate to provide transfer function estimates together with error bounds, the robust control community has a responsibility to accept this information in a realistic format which almost certainly precludes bounds which are absolute.

Furthermore, we contend that the stochastic embedding approach is a very appropriate one to choose because of the nature of undermodeling. That is, undermodeling typically arises because of physical manifestations that are too complicated to exactly describe. The best that can be hoped for is to capture the *on-average* properties of the undermodeling so that its most likely manifestation can be predicted. In this case, a probability density function is an appropriate choice for describing the undermodeling. Indeed, we would argue that the common assumption of measurement noise existing and being modeled by a stochastic process is an equivalent injection of a probabilistic framework on an essentially deterministic underlying problem. Our approach thus has an obvious antecedent in the whole paradigm of stochastic estimation theory.

II. Problem Description

We consider the problem of estimating a model for a dynamic system on the basis of the observation of an N point input–output data sequence $Z^N = [\{u_k\}, \{y_k\}]$ where we assume that the observed data Z^∞ is generated by the system $\mathcal{S}$ according to

$$\mathcal{S}: y_k = G_T(q^{-1})u_k + H(q^{-1})e_k. \tag{1}$$

Here $G_T(q^{-1})$ and $H(q^{-1})$ are rational transfer functions in the backward shift operator q^{-1}. We assume that both are strictly stable and have no poles in $|q| \geq 1$. The disturbance sequence $\{e_k\}$ is assumed to be an i.i.d. stochastic process defined on some probability space $\{\Omega, \mathcal{F}, \mathbb{P}\}$ with $\mathcal{E}\{e_k\} = 0$ where $\mathcal{E}\{\cdot\}$ denotes expectation with respect to the measure $\mathbb{P}$ on Ω. Finally, we assume that $\{u_k\}$ is a quasi-stationary sequence in the sense of [17] and that $\{e_k\}$ is independent of $\{u_k\}$ so that closed-loop collection of the data is ruled out. Consider a predictor for $\{y_k\}$ parameterized by a vector $\theta \in \mathbb{R}^p$

$$\hat{y}_k(\theta) = G(q^{-1}, \theta)u_k \tag{2}$$

where the prediction model $G(q^{-1}, \theta)$ is a member of the model set $\mathcal{M}_p^*$

$$\mathcal{M}_p^* = \{G(q^{-1}, \theta): \theta \in D_{\mathcal{M}} \subseteq \mathbb{R}^p\} \tag{3}$$

and there exists a smooth mapping $\mathcal{M}$ between $\theta \in D_{\mathcal{M}} \subseteq \mathbb{R}^p$ and $\mathcal{M}_p^*$

$$\mathcal{M}: \theta \to \{G(q^{-1}, \theta) \in \mathcal{M}_p^*\}. \tag{4}$$

Suppose, for ease of presentation, that we estimate θ from the data Z_N via the classical least-squares estimate

$$\hat{\theta}_N = \arg\min_\theta \frac{1}{N}\sum_{k=1}^{N} \epsilon_k^2(\theta) \tag{5}$$

$$\epsilon_k(\theta) = y_k - \hat{y}_k(\theta). \tag{6}$$

In [18], Ljung showed that under reasonable conditions

$$\hat{\theta}_N \to \theta^* \tag{7}$$

where

$$\theta^* = \arg\min_\theta \frac{1}{N}\sum_{k=1}^{N} \mathcal{E}\{\epsilon_k^2(\theta)\}. \tag{8}$$

With this definition of θ^* we can now examine the *total error* between the true transfer function $G_T(q^{-1})$, and the estimated one $G(q^{-1}, \hat{\theta}_N)$, at any single-frequency point, and decompose it as follows:

$$G_T(e^{-j\omega}) - G(e^{-j\omega}, \hat{\theta}_N) = G_T(e^{-j\omega}) - G(e^{-j\omega}, \theta^*) + G(e^{-j\omega}, \theta^*) - G(e^{-j\omega}, \hat{\theta}_N). \tag{9}$$

The first contribution, $G_T(e^{-j\omega}) - G(e^{-j\omega}, \theta^*)$, is called the *bias error*. In classical identification theory, it is a deterministic quantity. This bias error will be present, at least at some frequencies, as long as the model set $\mathcal{M}_p^*$ has lower complexity than the true system [i.e., when there is no value of θ for which $G_T(q^{-1}) = G(q^{-1}, \theta)$]. The second contribution $G(e^{-j\omega}, \theta^*) - G(e^{-j\omega}, \hat{\theta}_N)$, is called the *noise error or variance error*. It is a random variable with respect to the probability space of the noise distribution. it vanishes when there is no noise or when the number of data tends to infinity.

III. Solution via Stochastic Embedding

As stated in the Introduction, our ambition in this paper is to obtain an estimate for the on-average characteristics of the total error. The technical tool for doing so is to also *make the bias error a random variable* by ascribing a prior distribution to it. We therefore assume, following [26], [21], [11], [12], [22], and [8], that the true transfer function $G_T(e^{-j\omega})$ is a stochastic process indexed by the variable ω. We further assume that, for the given choice of model set $\mathcal{M}_p^*$, and for *some value* θ_0, it can be decomposed as

$$G_T(e^{-j\omega}) = G(e^{-j\omega}, \theta_0) + G_\Delta(e^{-j\omega}) \quad \text{with } \mathcal{E}\{G_T(e^{-j\omega})\} = G(e^{-j\omega}, \theta_0). \tag{10}$$

It follows that $G_\Delta(e^{-j\omega})$ is a zero mean stochastic process

$$\mathcal{E}\{G_\Delta(e^{-j\omega})\} = 0. \tag{11}$$

Note that $\mathcal{E}\{\cdot\}$ means averaging over different realizations of the undermodeling. Of course, for any given system we will have just one realization. This is analogous to the embedding of the single noise realization in a stochastic process for the purpose of analysis.

We assume that $\{\nu_k\}$ and G_Δ are independent. The rest of the assumptions in G_Δ are contained in the probability density function (pdf) that we choose to associate with it. Call this pdf $f_\Delta(G_\Delta, \beta)$, where β is a real vector parameterizing $f_\Delta(\cdot,\cdot)$. This pdf and β are chosen to describe the likelihood of various realizations of G_Δ being observed. One thrust of this paper will be to show that, once such a parameterized structure has been chosen for $f_\Delta(G_\Delta, \beta)$, then the parameter vector β can be estimated from the data. There need be no conjecture about the undermodeling being random, when we know it to be deterministic since, to quote [4], "A random variable 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." In fact, it is a function that maps how the state of nature manifests itself in observations. A probability density function is then associated with the random variable to center attention on a particular class of manifestations.

Note that within our stochastic embedding paradigm lies the class of hard bounding solutions proposed in the literature. All that is required is to specify $f_\Delta(G_\Delta, \beta)$ with compact support. However, we have added an extra degree of freedom in our formulation by not constraining $f_\Delta(G_\Delta, \beta)$ to be a uniform pdf. This allows a finer structure to be injected into the description of the class of undermodelings considered likely. As will be shown by simulation, the result is that a finer structure can be obtained in the uncertainty quantification. We also assume that the form of the pdf $f_\nu(\nu_k, \gamma)$ which is parameterized by the real vector γ can be specified, for example, $\nu_k \sim \mathcal{N}(0, \sigma_\nu^2)$.

Example 1: An example of a possible assumption on $f_\Delta(G_\Delta, \beta)$ would be that $G_\Delta(e^{-j\omega})$ is a zero mean Gaussian distributed stochastic process in the frequency domain with

$$\mathcal{E}\{G_\Delta(e^{-j\omega_1})G_\Delta(e^{j\omega_2})\} = \frac{\alpha\lambda e^{-j\omega}}{1-\lambda e^{-j\omega}}; \qquad \omega \triangleq \omega_1 - \omega_2 \tag{12}$$

and hence $\beta^T = [\alpha, \lambda]$. Now, without the loss of generality, we can use an IIR expression for $G_\Delta(q^{-1})$

$$G_\Delta(q^{-1}) = \sum_{k=1}^{\infty} \eta_k q^{-k}. \tag{13}$$

Since $\{G_\Delta(e^{-j\omega})\}$ is a zero mean Gaussian and covariance stationary process, the impulse response sequence $\{\eta_k\}$ is a Gaussian and independent process with [26]

$$\mathcal{E}\{\eta_k^2\} = \alpha\lambda^k \tag{14}$$

so, the prior assumption on the undermodeling is equivalent to the fact that its impulse response dies at a rate faster than $\alpha\lambda^k$.

Furthermore, from (12)

$$\mathcal{E}\left\{\left|G_\Delta(e^{-j\omega_1}) - G_\Delta(e^{-j\omega_2})\right|^2\right\} = \frac{2\alpha\lambda(1+\lambda)(1-\cos\omega)}{(1-\lambda)(1+\lambda^2-2\lambda\cos\omega)} \tag{15}$$

$$\leq \left[\frac{\alpha\lambda(1+\lambda)}{(1-\lambda)^3}\right]\omega^2 \tag{16}$$

where $\omega \triangleq (\omega_1 - \omega_2)$. The prior assumption is thus also equivalent to the fact that its frequency response satisfies the Lipschitz smoothness condition (16) as well as a magnitude condition $\mathcal{E}\{|G_\Delta(e^{j\omega})|^2\} = (\alpha/1-\lambda)$. ■

This example illustrates that we can impose the stochastic embedding of G_Δ by specifying a prior probability distribution for $G_\Delta(e^{-j\omega})$ or for the impulse response sequence $\{\eta_k\}$ of $G_\Delta(q^{-1})$. In the rest of this paper, we will choose to specify the distribution on η. However, in order to use the stochastic embedding model $G_\Delta(q^{-1})$ to calculate cov $\{\hat{\theta}_N\}$ and hence, cov $\{G_T(q^{-1}) - G(q^{-1}, \hat{\theta}_N)\}$, it is necessary to add the final assumption that $G_\Delta(q^{-1})$ can be approximated sufficiently closely by an FIR model of order $L \leq N$.

$$G_\Delta(e^{-j\omega}) = \Pi(e^{-j\omega})\eta \tag{17}$$

where

$$\Pi(e^{-j\omega}) \triangleq \left[e^{-j\omega}, \cdots, e^{-jL\omega}\right] \tag{18}$$

$$\eta^T \triangleq [\eta_1, \cdots, \eta_L]. \tag{19}$$

In order to obtain tractable expressions for the second-order properties of $\hat{\theta}_N$, we introduce our final assumption that the model structure $\mathcal{M}$ is a mapping to rational transfer functions $G(q^{-1}, \theta)$ with a fixed denominator; that is, only the numerator is parameterized by θ. FIR models and the Laguerre models studied in [19], [20], [31], and [27] are examples of such model structures.[1] With this assumption, the nominal model can be written as

$$G(e^{-j\omega}, \theta) = \Lambda(e^{-j\omega})\theta \tag{20}$$

where

$$\Lambda(e^{-j\omega}) \triangleq \left[\Lambda_1(e^{-j\omega}), \cdots, \Lambda_p(e^{-j\omega})\right]. \tag{21}$$

With these assumptions on the nominal model and on the unmodeled dynamics, the system equation (1) can now be

[1] More general ARMA models are discussed in [9].

rewritten in signal form as follows:

$$y_k = \phi_k^T \theta_0 + \psi_k^T \eta + \nu_k \tag{22}$$

with

$$\nu_k \triangleq H(q^{-1}) e_k \tag{23}$$

$$\psi_k^T \triangleq [u_{k-1}, \cdots, u_{k-L}] \tag{24}$$

where θ_0 has been defined in (10) and where ϕ_k is a vector containing filtered versions of the input signal.

If we employ the notation

$$\Phi^T \triangleq [\phi_1, \phi_2, \cdots, \phi_N] \tag{25}$$

$$Y^T \triangleq [y_1, y_2, \cdots, y_N] \tag{26}$$

$$V^T \triangleq [\nu_1, \nu_2, \cdots, \nu_N] \tag{27}$$

then $\hat{\theta}_n$ satisfying (5) is given by

$$\hat{\theta}_N = (\Phi^T \Phi)^{-1} \Phi^T Y. \tag{28}$$

Since $\{\eta_k\}$ is assumed independent of $\{\nu_k\}$, it is straightforward to conclude from (22) and (28) that

$$\operatorname{cov}(\hat{\theta}_N - \theta_0) \triangleq P_\theta = (\Phi^T \Phi)^{-1} \Phi^T (\Psi C_\eta \Psi^T + C_\nu) \Phi (\Phi^T \Phi)^{-1} \tag{29}$$

where

$$C_\eta \triangleq \mathcal{E}\{\eta \eta^T\} \tag{30}$$

$$C_\nu \triangleq \mathcal{E}\{V V^T\} \tag{31}$$

$$\Psi^T \triangleq [\psi_1, \cdots, \psi_N] \tag{32}$$

and so the second-order properties of the parameter estimates can be quantified to give a guide to uncertainty.

Prior assumptions on the likely nature of the undermodeling can thus be translated into probable influences on $\hat{\theta}_N$ via (29). The quantification of probable influences on $G(e^{-j\omega}, \hat{\theta}_N)$ may be obtained by the following theorem.

Theorem 1: Define

$$\tilde{g}(e^{-j\omega}) \triangleq \begin{bmatrix} \operatorname{Re}\{G_T(e^{-j\omega}) - G(e^{-j\omega}, \hat{\theta}_N)\} \\ \operatorname{Im}\{G_T(e^{-j\omega}) - G(e^{-j\omega}, \hat{\theta}_N)\} \end{bmatrix} \tag{33}$$

$$Q \triangleq (\Phi^T \Phi)^{-1} \Phi^T \tag{34}$$

$$\Upsilon \triangleq \begin{bmatrix} Q(\Psi C_\eta \Psi^T + C_\nu) Q^T & -Q \Psi C_\eta \\ -C_\eta \Psi^T Q^T & C_\eta \end{bmatrix} \tag{35}$$

$$\Gamma(e^{-j\omega}) \triangleq \begin{bmatrix} \operatorname{Re}\{\Lambda(e^{-j\omega}), \Pi(e^{-j\omega})\} \\ \operatorname{Im}\{\Lambda(e^{-j\omega}), \Pi(e^{-j\omega})\} \end{bmatrix}. \tag{36}$$

Then under the stochastic embedding model for $G_\Delta(q^{-1})$

$$G_T(e^{-j\omega}) - G(e^{-j\omega}, \hat{\theta}_N) = (\Pi - \Lambda Q \Psi)\eta - \Lambda Q V \tag{37}$$

and furthermore

$$\mathcal{E}\{\tilde{g}(e^{-j\omega})\} = 0 \tag{38}$$

$$\mathcal{E}\{\tilde{g}(e^{-j\omega}) \tilde{g}(e^{-j\omega})^T\} \triangleq P_{\tilde{g}} \tag{39}$$

$$= \Gamma(e^{-j\omega}) \Upsilon \Gamma^T(e^{-j\omega}) \tag{40}$$

$$\mathcal{E}\{|G(e^{-j\omega}, \hat{\theta}_N) - G_T(e^{-j\omega})|^2\} = \operatorname{tr}\{P_{\tilde{g}}\} \tag{41}$$

$$= (\Pi - \Lambda Q \Psi) C_\eta (\Pi - \Lambda Q \Psi)^* + \Lambda Q C_\nu Q^T \Lambda^* \tag{42}$$

where * denotes conjugate transpose. The expression (37) is a key result of the stochastic embedding approach [26], [21], [11], [12], [22], and [8].

Proof: Using the expressions (10) and (17), we have

$$G_T(e^{-j\omega}) = \Lambda \theta_0 + \Pi \eta \tag{43}$$

$$G(e^{-j\omega}, \hat{\theta}_N) = \Lambda \hat{\theta}_N \tag{44}$$

to give

$$G_T(e^{-j\omega}) - G(e^{-j\omega}, \hat{\theta}_N) = \Lambda(\theta_0 - \hat{\theta}_N) + \Pi \eta. \tag{45}$$

However, from (22) and (28) and the definition of Q we have

$$\theta_0 - \hat{\theta}_N = -Q \Psi \eta - Q V. \tag{46}$$

Substituting this in (45) then gives (37). Furthermore,

$$\tilde{g}(e^{-j\omega}) = \Gamma(e^{-j\omega})(\rho_0 - \hat{\rho}_N) \tag{47}$$

where

$$\hat{\rho}_N \triangleq [\hat{\theta}_N^T, 0^T]^T \tag{48}$$

$$\rho_0 \triangleq [\theta_0^T, \eta^T]^T. \tag{49}$$

Using (46), and remembering that V and η are independent then gives $\operatorname{cov}\{\rho_0 - \hat{\rho}_N\} = \Upsilon$. This, combined with (47), then gives (40). The result (42) follows from the expression of $P_{\tilde{g}}$, or more simply, from (37). □□□

If the prior distributions f_Δ and f_ν are chosen to be Gaussian, then $\hat{\theta}_N$ and $\tilde{g}(e^{-j\omega})$ are Gaussian distributed

$$\hat{\theta}_N \sim \mathcal{N}(\theta_0, P_\theta) \qquad \tilde{g}(e^{-j\omega}) \sim \mathcal{N}(0, P_{\tilde{g}}) \tag{50}$$

and hence quadratic functions of $\tilde{g}$ have χ^2 distributions. In particular

$$\tilde{g}(e^{-j\omega})^T P_{\tilde{g}}^{-1} \tilde{g}(e^{-j\omega}) \sim \chi^2_{(2)}. \tag{51}$$

We can use this to give confidence ellipses in the complex plane for the frequency response estimate $G(e^{-j\omega}, \hat{\theta}_N)$.

Comment: The expression (37) is a key result of the stochastic embedding approach [26], [21], [11], [12], [22], and [8]. On the right-hand side, the quantities Π and Λ are known functions of the frequency ω, while Q and Ψ are known functions of the signals. Hence, (37) expresses the total error as a known linear combination of two independent random vectors η and V. The expression clearly separates out the undermodeling error $(\Pi - \Lambda Q \Psi)\eta$ and the noise induced error $\Lambda Q V$. The term Πn is the prior estimate of the undermodeling [see (17)] while the term $\Lambda Q \Psi \eta$ is a data-induced correction to this prior due to the shift from θ_0 to $\hat{\theta}_N$. As for the mean square error expression, we note that all the quantities on the right-hand side of (42) are known except for the covariances C_η and C_ν, which are known functions of the unknown parameter vectors β and γ, respectively. In the next section, we will show that these parameter vectors can be estimated from the data. By replacing β and γ by their estimates in (42), we will then have obtained com-

putable estimates of the mean square error on the estimated transfer functions. The same comment applies, of course, to the expression of $P_{\tilde{g}}$.

IV. Estimation of the Parameterization of the Noise and Undermodeling

In order to use Theorem 1 to quantify errors, it is necessary to specify the form of the distribution f_Δ and also its parameterization β. In [21], [11], [12], and [22], a Bayesian stance was adopted, whereby β was specified prior to the identification experiment. It was proposed that this could be done by consideration of the magnitude and smoothness constraints given in Example 1 in (12) and (16).

In this paper our new contribution is to abandon the Bayesian framework and propose that the parameters β be estimated after the experiment from the data. In this case only the *form* of the pdf f_Δ need be specified *a priori*. Note that since we have constrained $G_\Delta(q^{-1})$ to have zero mean value, the parameterization β of f_Δ affects only the second and higher order properties of G_Δ. Therefore, estimating β from the data does *not* amount to estimating $G_\Delta(q^{-1})$. It amounts to estimating the likely class of $G_\Delta(q^{-1})$'s, of which we observe a realization. This idea has its obvious analog in estimating the variance of $\{\nu_k\}$ from the prediction residuals —a well-known technique. Indeed, we follow this paradigm and propose that β and γ, the parameters characterizing f_Δ and f_ν, be estimated form the residuals. Therefore, we define the N-vector of residuals

$$\epsilon \triangleq Y - \Phi\hat{\theta} \tag{52}$$

$$= \left[I - \Phi(\Phi^T\Phi)^{-1}\Phi^T\right]Y \triangleq PY. \tag{53}$$

The matrix in (53) 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 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

$$W \triangleq R^T\epsilon. \tag{54}$$

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

$$W = R^TY = R^T\Psi\eta + R^TV. \tag{55}$$

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 parameter vectors β and γ. We can therefore compute the probability density function of W, conditioned on the observed input data vector U, and on $\xi^T \triangleq (\beta^T, \gamma^T)$. 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

$$\hat{\xi} = \arg\max_{\xi} \{\mathcal{L}(W \mid U, \xi)\}. \tag{56}$$

We investigate the properties of $\hat{\xi}$ for Gaussian assumptions on f_Δ and f_ν in the following example.

Example 2: Consider the special case of the stochastic embedding assumptions being the same as in Example 1. That is

$$\eta \sim \mathcal{N}(0, C_\eta(\beta)) \tag{57}$$

$$C_\eta(\beta) = \operatorname*{diag}_{1\le k\le L}\{\alpha\lambda^k\}. \tag{58}$$

In addition, assume that the noise $\{\nu_k\}$ is iid, independent of η and with $\nu_k \sim \mathcal{N}(0, \sigma_\nu^2)$. In this case $\beta^T = [\alpha, \lambda]$ and $\xi^T = [\alpha, \lambda, \sigma_\nu^2]$. The Gaussian assumptions on the distributions f_Δ and f_ν give the log likelihood function $l(W \mid U, \xi)$ for the observed data as

$$l(W \mid U, \xi) = -\frac{1}{2}\ln\det\Sigma - \frac{1}{2}W^T\Sigma^{-1}W + \text{constant} \tag{59}$$

where

$$\Sigma = R^T\Psi C_\eta(\alpha, \lambda)\Psi^TR + \sigma_\nu^2R^TR \tag{60}$$

$$C_\eta(\alpha, \lambda) = \operatorname{diag}\{\alpha\lambda, \alpha\lambda^2, \cdots, \alpha\lambda^L\}. \tag{61}$$

It is well known [10] that the covariance of an unbiased estimate $\bar{\xi}$ of ξ is bounded below by the Cramér–Rao lower bound.

$$\operatorname{cov}\{\bar{\xi}\} \ge M_\xi^{-1} \triangleq \underline{\operatorname{cov}}(\bar{\xi}) \tag{62}$$

That is, $\underline{\operatorname{cov}}(\bar{\xi}) = M_\xi^{-1}$ is a lower bound on the covariance of $\bar{\xi}$ where M_ξ^{-1} is the Fisher information matrix

$$[M_\xi]_{ij} = \mathcal{E}\left\{\frac{\partial l}{\partial\xi_i}\frac{\partial l}{\partial\xi_j}\right\}. \tag{63}$$

Experience shows that $\underline{\operatorname{cov}}$ is often a good guide to the covariance of estimators which are not actually unbiased. For the case of Gaussian assumptions in this example, we can compute an explicit expression for the information matrix.

Result: The information matrix M_ξ has the following form:

$$M_\xi = \frac{1}{2}\begin{bmatrix} \operatorname{tr}\left[(\Sigma^{-1}T)^2\right] & \operatorname{tr}\left[\Sigma^{-1}T\Sigma^{-1}K\right] & \operatorname{tr}\left[\Sigma^{-1}T\Sigma^{-1}\Delta\right] \\ \operatorname{tr}\left[\Sigma^{-1}T\Sigma^{-1}K\right] & \operatorname{tr}\left[(\Sigma^{-1}K)^2\right] & \operatorname{tr}\left[\Sigma^{-1}K\Sigma^{-1}\Delta\right] \\ \operatorname{tr}\left[\Sigma^{-1}T\Sigma^{-1}\Delta\right] & \operatorname{tr}\left[\Sigma^{-1}K\Sigma^{-1}\Delta\right] & \operatorname{tr}\left[(\Sigma^{-1}\Delta)^2\right] \end{bmatrix} \tag{64}$$

where

$$T \triangleq R^T\Psi\Upsilon\Psi^TR = \frac{\partial\Sigma}{\partial\alpha} \tag{65}$$

$$K \triangleq R^T\Psi\Xi\Psi^TR = \frac{\partial\Sigma}{\partial\lambda} \tag{66}$$

$$\Delta \triangleq R^T R = \frac{\partial \Sigma}{\partial \sigma_v^2} \tag{67}$$

$$\Upsilon \triangleq \operatorname{diag}(\lambda, \lambda^2, \cdots, \lambda^L) = \frac{\partial C_\eta}{\partial \alpha} \tag{68}$$

$$\Xi \triangleq \operatorname{diag}(\alpha, 2\alpha\lambda, \cdots, L\alpha\lambda^{L-1}) = \frac{\partial C_\eta}{\partial \lambda}. \tag{69}$$

Proof: Using (59) we have, on applying Lemmas 1 and 2 of the Appendix

$$\frac{\partial l(\xi)}{\partial \xi_i} = -\frac{1}{2}\operatorname{tr}\left[\Sigma^{-1}\frac{\partial \Sigma}{\partial \xi_i}\right] + \frac{1}{2}W^T\Sigma^{-1}\left[\frac{\partial \Sigma}{\partial \xi_i}\right]\Sigma^{-1}W. \tag{70}$$

Then, using Lemma 3 from the Appendix

$$\mathcal{E}\left\{\frac{\partial l(\xi)}{\partial \xi_i}\frac{\partial l(\xi)}{\partial \xi_j}\right\} = \frac{1}{2}\operatorname{tr}\left[\Sigma^{-1}\frac{\partial \Sigma}{\partial \xi_i}\Sigma^{-1}\frac{\partial \Sigma}{\partial \xi_j}\right]. \tag{71}$$

Replacing in (63) yields the desired result. □□□ ■

The availability of M_ξ and the simple form of $l(W - VU, \xi)$ in the case of Gaussian embedding motivates us to use the Gaussian assumption in practice. This leads to an algorithm which can, of course, also be applied in the non-Gaussian case as we shall do in the examples later. We examine more closely the properties of $\hat{\xi}_N$ based on the Gaussian embedding.

Example 3: Under the stochastic embedding assumptions of Example 2, we can use the previous results to examine the properties of the estimate $\hat{\xi}$ for the case where nominal models are finite impulse response models, FIR(p), and where the input sequence is deterministic and has the following orthogonality property:

$$\frac{1}{N}(\Psi^T\Psi) = \sigma_u^2 I \tag{72}$$

where σ_u^2 is the input power. In this case we have

$$\Phi = \begin{bmatrix} u_0 & \cdots & u_{1-p} \\ \vdots & & \\ u_{N-1} & \cdots & u_{N-p} \end{bmatrix},$$

$$\Psi = \begin{bmatrix} u_0 & \cdots & u_{1-N} \\ \vdots & \ddots & \vdots \\ u_{N-1} & \cdots & u_0 \end{bmatrix}. \tag{73}$$

Without loss of generality, we have taken $L = N$. In line with the requirement that R be orthogonal to Φ, we let R be the last $N - p$ columns of Ψ

$$R = \begin{bmatrix} u_{1-p-1} & \cdots & u_{1-N} \\ \vdots & & \vdots \\ u_{N-p-1} & \cdots & u_0 \end{bmatrix}. \tag{74}$$

By (72), $R^T\Phi = 0$. The new data vector W, defined by (54) or (55), is now

$$W = R^T\Psi\eta + R^T V = \left[0 \vdots N\sigma_u^2 I_{N-p}\right]\eta + R^T V. \tag{75}$$

Notice that $W \in \mathbb{R}^{N-p}$ depends only on $\eta_{p+1}, \cdots, \eta_N$, i.e., that part of the infinite impulse response that is included in the nominal model is eliminated from the W data. For this simple undermodeling case, Lemma A.4 uses the result in Example 2 to give the following.

- An unbiased estimate $\bar{\sigma}_v^2$ of σ_v^2 is asymptotically decoupled from unbiased estimates $\bar{\alpha}$ of α and $\bar{\lambda}$ of λ.
- $$\underline{\operatorname{cov}}\begin{pmatrix}\bar{\alpha}\\ \bar{\lambda}\end{pmatrix} = \mathcal{O}\left(\frac{1}{\ln N}\right) \quad \text{as } N \to \infty \tag{76}$$

$$\underline{\operatorname{cov}}\begin{pmatrix}\bar{\lambda}\\ \bar{\lambda}\end{pmatrix} = \mathcal{O}\left(\frac{1}{(\ln N)^3}\right) \quad \text{as } N \to \infty \tag{77}$$

$$\underline{\operatorname{cov}}\begin{pmatrix}\bar{\sigma}_v^2\\ \sigma_0^2\end{pmatrix} = \mathcal{O}\left(\frac{1}{N}\right) \quad \text{as } N \to \infty. \tag{78}$$

These equations are significant since they show that the variance of the estimates decays with increasing data length and therefore we can expect the estimates to converge quickly to their true values.

These asymptotic properties of the maximum likelihood estimation of ξ were examined via Monte-Carlo simulation. In this case, 800 trials were conducted. In each trial, the true impulse response sequences $\{\eta_k\}$ were randomly generated with distribution

$$\eta_k \sim N(0, 0.9^k).$$

These random impulse responses were then convolved with 1000 samples of a zero mean, unit variance, Gaussian distributed white noise sequence $\{u_k\}$. Finally, these convolved signals were corrupted by white noise sequences $\{v_k\}$ distributed as

$$v_k \sim N(0, 50).$$

The model for the data generation was, therefore,

$$y_k = \sum_{j=1}^{L}\eta_j u_{k-j+1} + v_k \tag{79}$$

where $L = 30$ was used. Following the previous result showing that the estimation of σ_v^2 is asymptotically decoupled from the estimates of α and λ for each trial, we estimated $\hat{\sigma}_v^2$ by fitting an FIR (30) model to the data and computing $\hat{\sigma}_v^2$ as $(1/N - 30)(Y - \Psi\hat{\eta})(Y - \Psi\hat{\eta})$, where η is the least-squares estimate of the FIR (30) model. The parameters α and λ were then estimated from the full data vector $Y = \Psi\eta + V$ using (59). The results for all 800 trials are shown in Fig. 1. The sample means, sample standard deviations, and the theoretical Cramér–Rao lower bounds from $M_\xi^{-1}(800)$ calculated using the result in Example 2 are given in Table I. As can be seen, the sample standard deviations are approaching the Cramér–Rao lower bounds, indicating that the estimator proposed in this paper is in practice an efficient estimator for the characteristics of the undermodeling. ■

V. Error Bounding Simulations

A simulation study was conducted to examine the use of the stochastic embedding paradigm in quantifying the estima-

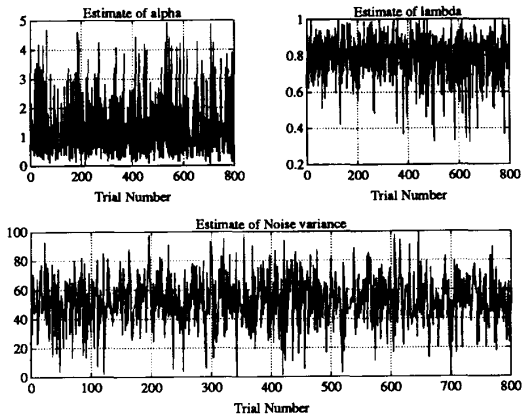


Fig. 1. Monte-Carlo testing of estimates of α, λ, and σ^2.

TABLE I
SAMPLE MEANS AND STANDARD DEVIATIONS FROM MONTE-CARLO SIMULATION COMPARED TO CRAMER–RAO LOWER BOUNDS

Values	True of Estimates	Sample Means of Estimates	Sample S.D.'s Bounds	C.R.
α	1.0	1.14	0.83	0.62
λ	0.9	0.83	0.064	0.046
σ^2	50.0	50.02	2.38	1.68

tion errors when rational, fixed denominator models were fitted to the data. In this case, the following continuous-time system, sampled with period 1 s, 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 $\{v_k\}$ distributed as $v_k \sim \mathcal{N}(0, 0.005)$. 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 second-order model of the form

$$y_k = \left(\frac{\theta_1 q^{-1}}{(1 + \xi q^{-1})} + \frac{\theta_{2q}^{-1}(1 - (2 + \xi)q^{-1})}{(1 + \xi q^{-1})^2} \right) u_k$$

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 [19]. The resulting least-squares estimates were

$$\hat{\theta}_1 = 0.1245 \qquad \hat{\theta}_2 = -0.0715.$$

The response of the estimated model $G(q^{-1}, \hat{\theta}_N)$ to the observed input is shown dashed against the noise corrupted true response in the upper left of Fig. 2. The true (full line) and estimated (dashed line) frequency responses are shown in the upper right corner of Fig. 2.

Next, the parameters of the distributions of the measurement noise and undermodeling were estimated from the data. The stochastic embedding chosen for the undermodeling and

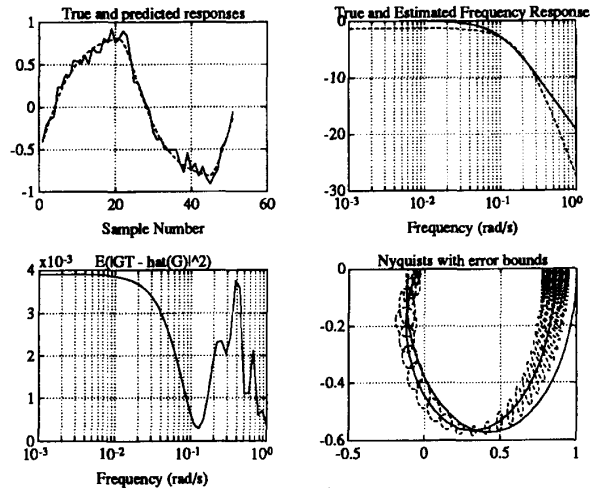


Fig. 2. Results of fitting a second-order, rational Laguerre model to the data. No time delay in system.

noise was that given in Example 2. That is

$$v_k \sim \mathcal{N}(0, \sigma_v^2) \tag{80}$$

$$\eta \sim \mathcal{N}(0, C_\eta) \tag{81}$$

$$C_\eta = \operatorname*{diag}_{1 \le k \le L} \{\alpha \lambda^k\} \tag{82}$$

$\dim(\eta) = L = 40$ was chosen for the FIR model G_Δ. Because of the Gaussian assumptions for the undermodeling, and the Gaussian distribution for the measurement noise, the log likelihood function for the observed data is as given in (59). This was maximized to find the estimates

$$\hat{\alpha} = 0.187 \qquad \hat{\lambda} = 0.4064 \qquad \hat{\sigma}_v^2 = 0.0052.$$

Substituting these estimates in (80) and (82) then gives estimates of C_η and $C_v = \sigma_v^2 I$. These were then used to derive error bounds for the frequency response estimation error. Error bounds on magnitude estimation were calculated via (41) of Theorem 1, replacing C_η and C_v by their estimates and are shown in the lower left of Fig. 2. Uncertainty ellipsoids in the complex plane were calculated similarly via Theorem 1 and (51) and are shown superimposed on the true and estimated Nyquist diagrams in the lower right of Fig. 2. The uncertainty ellipsoids are one standard deviation ellipses; that is, the locus of points satisfying

$$\tilde{g}(e^{-j\omega})^T P_{\tilde{g}}^{-1} \tilde{g}(e^{-j\omega}) = 2. \tag{83}$$

As can be seen, the error bounds give a very good indication of the true modeling errors in the frequency domain. Note in particular from the lower left diagram in Fig. 2 that the estimated error in the estimation of the system magnitude response is small at the fundamental frequency ($0.02 \times 2\pi$ rad/s) of the input signal square wave and at the odd harmonics. This concurs with the analysis of [28].

An experiment in which the undermodeling was more severe was conducted. In this case, the setup was the same as for the results shown in Fig. 2, but a time delay of 2 s was

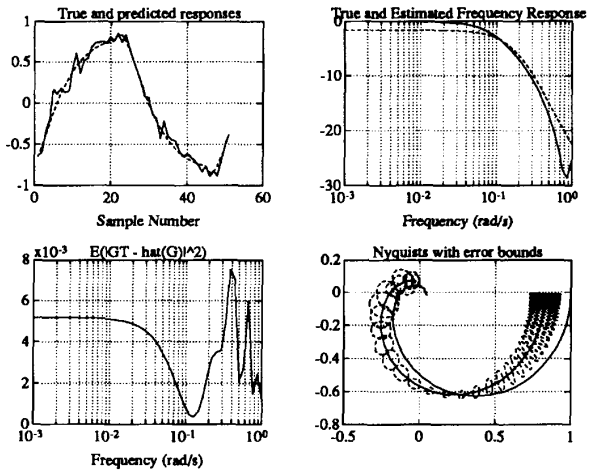


Fig. 3. Results of fitting a second-order, rational Laguerre model to the data. Time delay of 2 seconds in system.

also added to the plant. The estimates were

$$\hat{\theta}_1 = 0.0653 \qquad \hat{\theta}_2 = -0.1013$$

$$\hat{\alpha} = 0.217 \qquad \hat{\lambda} = 0.488 \qquad \hat{\sigma}_v^2 = 0.0061.$$

The results are displayed in Fig. 3. Note that once again the error bounds are highly informative.

Finally, although the theory developed in this paper does not pertain to ARMAX modeling, an ARMAX simulation to test the robustness of the maximum likelihood estimator to the assumptions was conducted. A first-order ARMAX model was fitted to the system with no time delay and measurement noise variance reduced to $\sigma_v^2 = 0.001$. The resulting estimated model was

$$\hat{G}(\delta) = \frac{1.1347q^{-1}}{12.049 - 11.049q^{-1}}. \tag{84}$$

The estimates for the parameterization of the undermodeling using the same stochastic embedding as in the previous simulations were

$$\hat{\alpha} = 0.0251 \qquad \hat{\lambda} = 0.8017 \qquad \hat{\sigma}_v^2 = 0.0015.$$

Note that for the ARMAX case $G(e^{-j\omega}, \theta)$ is not linear in θ. However, we can form an approximately linear expression by using Taylor's Theorem. Specifically, we note that G_Δ is parameterized by the impulse response vector η. Therefore $G_T(e^{-j\omega})$ is parameterized not only by θ for the nominal model component but also by η or the undermodeling component to give

$$G_T(e^{-j\omega}) = G(e^{-j\omega}, \theta_0) + \Pi(e^{-j\omega})\eta \tag{85}$$

$$\approx G(e^{-j\omega}, \hat{\theta}_N) + \left.\frac{\partial G(e^{-j\omega}, \theta)}{\partial \theta^T}\right|_{\theta=\hat{\theta}_N} (\theta_0 - \hat{\theta}_N) + \Pi(e^{-j\omega})\eta. \tag{86}$$

Hence, in Theorem 1 we have

$$\tilde{g}(e^{-j\omega}) \approx \bar{\Gamma}(e^{-j\omega})(\rho_0 - \hat{\rho}_N) \tag{87}$$

with ρ_0 and $\hat{\rho}_n$ defined in (48) and (49) and with

$$\bar{\Gamma}(e^{-j\omega}) \triangleq \begin{bmatrix} \mathrm{Re}\left\{ \left.\frac{\partial G(e^{-j\omega}, \theta)}{\partial \theta^T}\right|_{\theta=\hat{\theta}_N}, \Pi(e^{-j\omega}) \right\} \\ \mathrm{Im}\left\{ \left.\frac{\partial G(e^{-j\omega}, \theta)}{\partial \theta^T}\right|_{\theta=\hat{\theta}_N}, \Pi(e^{-j\omega}) \right\} \end{bmatrix}. \tag{88}$$

Substituting $\bar{\Gamma}(e^{-j\omega})$ for $\Gamma(e^{-j\omega})$ in Theorem 1 allows it to hold approximately due to the approximate linearization in (86). Hence, for the ARMAX case, approximate bounds may be derived and are given in Fig. 4 to illustrate that the method may be successfully applied to the ARMAX modeling case even though the case does not fit the assumptions. The theoretical basis for the ARMAX modeling case is treated in [9].

VI. Model Structure Selection

In this section, we show how the quantified error bounds in the form of the ensemble mean square error $\mathcal{E}\{|G_T(e^{-j\omega}) - G(e^{-j\omega}, \hat{\theta}_N)|^2\}$ of the transfer function estimate can be used for the selection of an optimal model structure for the nominal model. A variety of possible model structure selection criteria will be examined, all of them functions of this ensemble mean square error. In particular, if the family of candidate nominal models is a sequence of models of increasing order, then this yields an optimal model order selection criterion. We shall theoretically, and in simulation compare this criterion to Akaike's final prediction error (FPE) criterion.

Consider first that an optimal model is to be selected among a finite family of r candidate nominal models, all of them linear in the parameters. Denote the model structures $\mathcal{M}_1(\theta_1), \cdots, \mathcal{M}_r(\theta_r)$. Note that $\theta_1, \theta_2, \cdots, \theta_r$ may or may not have different dimensions. For each model structure $\mathcal{M}_i(\theta_i)$, one can estimate $\hat{\theta}_i$ by least squares. With the assumed prior distribution for $G_\Delta(q^{-1})$ and for $\{\nu_k\}$, we can then compute, for each estimated nominal model, the corresponding maximum likelihood estimates of β and γ. We shall denote by $\hat{\beta}_i$ and $\hat{\gamma}_i$ the estimates corresponding to $\mathcal{M}_i(\hat{\theta}_i)$, and by $\hat{V}_i(\omega)$ the estimate of the mean square error $\mathcal{E}\{|G_T(e^{-j\omega}) - G(e^{-j\omega}, \hat{\theta}_N)|^2\}$ in (41) in which $C_\eta(\beta)$ and $C_v(\gamma)$ are replaced by $C_\eta(\hat{\beta}_i)$ and $C_v(\hat{\sigma}_v^2) = \hat{\sigma}_v^2) = \hat{\sigma}_v^2 I$.

To select among the r candidate models, we shall now consider any one of the following three criteria:

$$J_i^1 = \sup_i \hat{V}_i(\omega), \qquad i = 1, \cdots, r \tag{89}$$

$$J_i^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{V}_i(\omega)\, d\omega, \qquad i = 1, \cdots, r \tag{90}$$

$$J_i^3 = \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{V}_i(\omega) S_u(\omega)\, d\omega, \qquad i = 1, \cdots, r. \tag{91}$$

Here $S_u(\omega)$ denotes the power spectral density of a possibly new input sequence to which the model will be applied. The three criteria obviously cover three different applications in which the model may be used. Other criteria can easily be formulated. Depending on the application, the optimal struc-

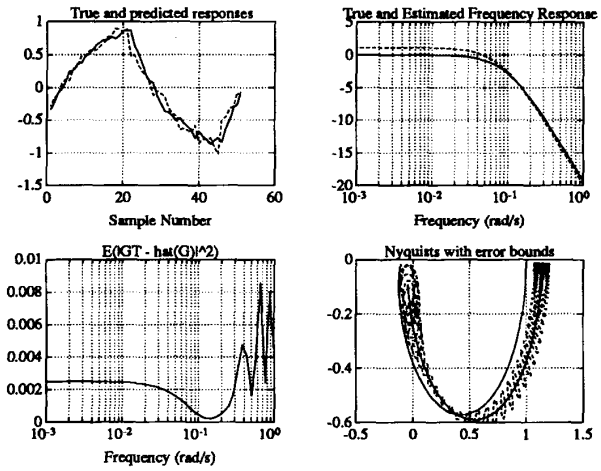


Fig. 4. Results of fitting a first-order ARMAX model to the data. No time delay in system.

ture will be obtained as $\mathcal{M}_{i^*}$ where

$$i^* = \arg\min_{i=1,\cdots,r} J_i^k \qquad k = 1, 2, \text{ or } 3. \tag{92}$$

In the rest of this section, we shall restrict discussion to one of the criterion (89) to (91). Namely, a modification of (91) that we call the generalized information criterion (GIC). In the presentation of GIC we shall also restrict ourselves to the case where $\{v_k\}$ is a white noise sequence of variance σ_v^2 and the model structures $\mathcal{M}_k$ correspond to the choice of fixed denominator models so that their indexing k can be taken to be the dimension p of the vector θ parameterizing them. In this case, GIC is defined as

$$\text{GIC}(p) \triangleq \hat{\sigma}_v^2 + \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{V}_p(\omega) S_u(\omega)\, d\omega \tag{93}$$

where we take (see Theorem 1)

$$\hat{V}_p(\omega) = \text{tr}\{P_{\tilde{g}}\} \tag{94}$$

$$= (\Pi - \Lambda Q\Psi) C_\eta(\hat{\beta}_p)(\Pi - \Lambda Q\Psi)^* + \Lambda Q C_v(\hat{\sigma}_v^2) Q^T \Lambda^*. \tag{95}$$

It is important to observe that in our definition of GIC (p), $\hat{\sigma}_v^2$ is obtained independently of the nominal model under consideration (see below), while $\hat{\beta}_p$ is the maximum likelihood estimate of β for the model with p parameters, hence the index p. By adding an input power into the criterion J^2 and an additional component $\hat{\sigma}_v^2$, we have converted a frequency averaged mean square error criterion on the transfer function estimate J^2 into a mean square output prediction error criterion for a new set of data with the same inputs but a new noise realization.

Lemma 6.1: An alternative frequency domain expression for GIC (p) is

$$\text{GIC}(p) \approx \hat{\sigma}_v^2 + \frac{p}{N}\hat{\sigma}_v^2 + \frac{1}{2\pi}\int_{-\pi}^{\pi} (\Pi - \Lambda Q\Psi) C_\eta(\hat{\beta}_p) \cdot (\Pi - \Lambda Q\Psi)^* S_u(\omega)\, d\omega. \tag{96}$$

Proof: Substituting (95) in (93) we obtain by using $C_v(\hat{\sigma}_v^2) = \hat{\sigma}_v^2 I$

$$\text{GIC}(p) = \hat{\sigma}_v^2 + \frac{1}{2\pi}\int_{\pi} \cdot\Big[(\Pi - \Lambda Q\Psi) C_\eta(\hat{\beta}_p)(\Pi - \Lambda Q\Psi)^* + \hat{\sigma}_v^2 \text{ Trace}\Big\{(\Phi^T\Phi)^{-1}\Lambda^*\Lambda\Big\}\Big] S_u(\omega)\, d\omega. \tag{97}$$

Furthermore, by Parseval's Theorem

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} (\Lambda^*\Lambda) S_u(\omega)\, d\omega = \lim_{N\to\infty} \frac{1}{N}(\Phi^T\Phi) \approx \frac{1}{N}(\Phi^T\Phi) \quad \text{for large } N. \tag{98}$$

Substituting (98) in (97) then gives the result. □□□

The three terms in (96) are, respectively, an estimate of the effect on the prediction error of the variance of the new noise realization, an estimate of the effect on the prediction error of the parameter errors due to noise in the identification data, and an estimate of the effect on the prediction error of the undermodeling. Akaike considers a similar criterion in his final prediction error (FPE) and Akaike information criterion (AIC) tests for model order selection [17]. In these criteria, Akaike seeks a model order that will perform well on average for data other than was used for estimation. The criteria for performance use a quadratic prediction error one for FPE, and a log-likelihood error one for AIC. In both cases, however, the Akaike criterion development does not explicitly acknowledge the presence of undermodeling and only averages over stochastic components in the data assuming the same inputs to the system.

In contrast, through the inclusion of the input spectral density term $S_u(\omega)$, our GIC criterion can be used to estimate a cross validation criterion for a different input realization. The spectral density of the desired cross validation input is simply included in (96). Furthermore, in our GIC criterion, the value used for $\hat{\sigma}_v^2$ is obtained independently of the particular nominal model under consideration. It could, for example, be obtained as the average of the estimates of σ_v^2 obtained for all different model dimensions. In practice, we have found that an accurate estimate of σ_v^2 is obtained as $(1/N - M)(Y - \Phi\hat{\theta}_M)^T(Y - \Phi\hat{\theta}_M)$, where $\phi^T\theta_M$ is a high-dimensional regression model, i.e., $\dim \theta_M = M$ with M large. On the other hand, Akaike's criterion explicitly depends on obtaining a new estimate of σ_v^2, denoted $\hat{\sigma}_p^2$, for each model dimension using

$$\hat{\sigma}_p^2 = \frac{1}{N-p}(Y - \Phi\hat{\theta}_p)^T(Y - \Phi\hat{\theta}_p). \tag{99}$$

This leads to the FPE criterion

$$\text{FPE}(p) = \frac{N+p}{N-p} \times \frac{N-p}{N}\hat{\sigma}_p^2 \tag{100}$$

and the AIC criterion

$$\text{AIC}(p) = \log\left(\frac{N-p}{N}\hat{\sigma}_p^2\right) + \frac{2p}{N}. \tag{101}$$

The estimate (99) would be an unbiased estimate of σ_v^2 if there were no undermodeling. Our rationale for using a high-dimensional model for the estimation of σ_v^2 is to ensure that undermodeling does not affect our estimate. Our criterion explicitly and, we believe, correctly accounts for undermodeling through the third term in (96). Even though, as we have just argued, undermodeling does implicitly affect the estimate of $\hat{\sigma}_p^2$ in Akaike's calculation, this results on average in a quantification of the undermodeling which is not entirely correct, as we now show.

Lemma 6.2: Under the stochastic embedding assumptions $\mathcal{E}\{\mathrm{FPE}(p)\}$ is given by

$$\mathcal{E}\{\mathrm{FPE}(p)\} \approx \sigma_v^2 + \frac{p}{N}\sigma_v^2 + \left(\frac{N+p}{N-p}\right)\frac{1}{2\pi}\int_{-\pi}^{\pi} \cdot(\Pi - \Lambda Q\Psi)C_\eta(\Pi - \Lambda Q\Psi)^* S_u(\omega)\, d\omega. \quad (102)$$

Proof: We first note that

$$Y - \Phi\hat{\theta}_p = \left[I - \Phi(\Phi^T\Phi)^{-1}\Phi^T\right](\Psi\eta + V). \quad (103)$$

Substituting (103) into (99) and (100), and taking expected value with respect to the noise and undermodeling yields

$$\mathcal{E}\{\mathrm{FPE}(p)\} = \frac{N+p}{N-p} \times \frac{1}{N}\Big\{(N-p)\sigma_v^2 + \mathrm{tr}\left(I - \Phi(\Phi^T\Phi)^{-1}\Phi^T\right)\Psi C_\eta \Psi^T \cdot\left(I - \Phi(\Phi^T\Phi)^{-1}\Phi^T\right)\Big\} \quad (104)$$

$$= \sigma_v^2 + \frac{p}{N}\sigma_v^2 + \frac{N+p}{N-p} \times \frac{1}{N}\,\mathrm{tr}\left\{\left(I - \Phi(\Phi^T\Phi)^{-1}\Phi^T\right)\Psi C_\eta \Psi^T \cdot\left(I - \Phi(\Phi^T\Phi)^{-1}\Phi^T\right)^T\right\}. \quad (105)$$

Applying Parseval's Theorem as in (98) then gives the result. □□□

Comparing the first two terms in (102) and (96) shows that the FPE criterion *on-average* captures the variance effects correctly. However, the bias term is incorrectly scaled by a factor $(N + p/N - p)$. To illustrate the consequence of this scaling, we compare GIC to FPE and AIC in a simulation example. Further insight into the GIC criterion is obtained by considering a special case previously studied in Example 3.

Example 4 – Example 3 Continued: Consider the problem setup in Example 3. Consider first the criterion J^2. We have, by the FIR model structure

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\left|G_T(e^{-j\omega}) - G(e^{-j\omega}, \hat{\theta})\right|^2 d\omega = \sum_{k=1}^{p}\left(\theta_k - \hat{\theta}_k\right)^2 + \sum_{k=p+1}^{N}\eta_k^2.$$

Therefore, by the orthogonality of the input sequence

$$J^2(p) = \frac{p\hat{\sigma}_v^2}{N\sigma_u^2} + \sum_{k=p+1}^{N}\hat{\alpha}.\hat{\lambda}^k = \frac{p\hat{\sigma}^2}{N\sigma_u^2} + \frac{\hat{\alpha}\hat{\lambda}^{p+1}(1 - \hat{\lambda}^{N-p})}{1 - \hat{\lambda}}. \quad (106)$$

Finally then, since $S_u(\omega) = \sigma_u^2$

$$\mathrm{GIC}(p) = \hat{\sigma}_v^2 + \frac{p\hat{\sigma}_v^2}{N} + \frac{\sigma_u^2\hat{\alpha}\hat{\lambda}^{p+1}(1 - \hat{\lambda}^{N-p})}{1 - \hat{\lambda}}. \quad (107)$$

Notice that the variance contribution increases linearly with p while the bias term decreases with p. The optimal model dimension, for this special case of FIR models, is obtained by the following result.

Result: For FIR(p) models, with an orthogonal input sequence and with G_Δ and $\{v_k\}$ stochastically distributed as in Example 2, the optimal model order with respect to GIC(p) is

$$p^* = \left[\ln\left(\frac{\hat{\sigma}_v^2}{N}\left(\frac{\hat{\lambda} - 1}{\hat{\lambda}}\right)\left(\frac{1}{\sigma_u^2\hat{\alpha}\ln\hat{\lambda}}\right)\right)\right]\left[\ln\hat{\lambda}\right]^{-1}. \quad (108)$$

Proof: Differentiating (107) with respect to p yields

$$\frac{\partial\,\mathrm{GIC}}{\partial p} = \frac{\hat{\sigma}_v^2}{N} + \left(\frac{\hat{\alpha}}{1 - \hat{\lambda}}\right)\sigma_u^2\hat{\lambda}^{p+1}\ln\hat{\lambda}. \quad (109)$$

Setting this to zero yields the desired result. □□□

A rough approximation to the above result is obtained by replacing the derivative, $(\partial GIC/\partial p)$, by the difference as p is increased by one

$$\frac{\partial\,\mathrm{GIC}}{\partial p} \simeq \frac{\hat{\sigma}_v^2}{N} + \sigma_u^2\frac{\hat{\alpha}(\hat{\lambda}^{p+1} - \hat{\lambda}^p)}{1 - \hat{\lambda}} = \frac{\hat{\sigma}_v^2}{N} - \sigma_u^2\hat{\alpha}\hat{\lambda}^p. \quad (110)$$

Hence, we have, approximately

$$\hat{\alpha}\hat{\lambda}^{p^*} \approx \frac{\hat{\sigma}_v^2}{N\sigma_u^2}. \quad (111)$$

This value of p^* then is precisely the point where the prior expected mean square value (averaged over the population of G_Δ) of the pth impulse response element is equal to the experimental variance of the pth impulse response estimate due to the noise, i.e., where $\mathcal{E}_v\{(\hat{h}_{p^*} - h_{p^*})^2\} = \mathcal{E}_\Delta\{h_{p^*}^2\}$, where h_k are the impulse response elements of the system. ■

Finally, it is interesting to compare our GIC criterion to the respected cross validation procedure for model order selection [17]. In the latter procedure, one tests the estimated model on a new set of data. The GIC criterion obviates the need for this second experiment by computing the expected performance on new data with either the same spectral distribution, or even on data with a different spectral distribution if that is desired.

VII. Simulation Study of Model Order Selection

Simulation examples to study GIC, FPE, and AIC were conducted for single realizations of 1000 data with the inputs and noise generated from zero mean white Gaussian noise sequences with variances $\sigma_u^2 = 1$ and $\sigma_v^2 = 50$, respectively. The true systems were the unit period sampled versions of the following continuous-time systems.

System 1:

$$\frac{1}{\tau s + 1} \qquad \tau = 0.1.$$

System 2:

$$\frac{e^{-\Delta s}}{\tau s + 1} \qquad \tau = 0.1, \qquad \Delta = 5.$$

System 3:

$$\frac{\omega_c^2}{s^2 + 2\zeta\omega_c s + \omega_c^2} \qquad \omega_c = 0.1, \qquad \zeta = 0.1.$$

For each of these three systems, and for a particular noise realization, a Gaussian prior model was assumed for G_Δ and for the noise as in Example 2, and the following calculations were made.

1) A range of nominal FIR (p) models was fitted to the data by least squares for $p = 1, \cdots, 30$. Since the inputs are uncorrelated, the estimated impulse response coefficients remain unchanged when the model order is increased. Therefore, a full FIR (30) model was estimated and the lower order estimated nominal models obtained from it by truncation.

2) As explained before, the parameter σ_v^2 was estimated from a high-order FIR model, here FIR (30), while estimates $\hat{\alpha}_p$, $\hat{\lambda}_p$ were obtained by maximization of the log likelihood function (59) for each nominal model-order p. These estimates were then used to evaluate the J^2 criterion (90) via (106). Note that $S_u(\omega)$ is constant here. Note also that with $N = 1000$ we have $1/N\Psi^T\Psi \approx \sigma_u^2 I$. That is, the input sequence is approximately orthogonal.

3) The integrals of the true squared frequency errors for various model orders were calculated for comparison to the estimates in 2 using the known $G_T(e^{-j\omega})$.

4) Akaike's AIC and FPE model-order determination criteria were calculated from the data.

Typical results for these calculations for the three different systems 1, 2, and 3 are shown in Figs. 5–7. The top left quadrants of each figure show the true impulse response together with the full 30th-order model found via least squares. The true impulse response is the solid line, and the estimated impulse response is the dashed line. The Akaike information criterion and final prediction error tests are shown in the bottom left- and right-hand corners of each figure. As can be seen, considering the vertical axis scales, they give inconclusive order determination criteria.

The top right plot shows the integral of the squared frequency response error

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \left|G_T(e^{-j\omega}) - \hat{G}(e^{-j\omega})\right|^2 d\omega \qquad (112)$$

versus the model order used to obtain $\hat{G}$ as a solid line, and the estimate of (112) as a dashed line. The estimate was obtained from the estimates $\hat{\theta}$, $\hat{\alpha}$, $\hat{\Lambda}$, and $\hat{\sigma}^2$ via (106). Notice that in all cases, the estimate of the value of the integral is a very good indication of the true value of the integral. Furthermore, notice that the integrals (true and estimated) give a very clear criterion for the best model order

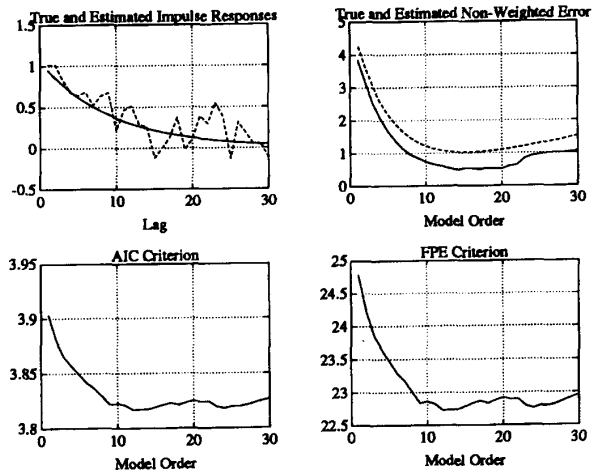


Fig. 5. Full-order estimation, Akaike criteria, and true square error and estimated mean square errors for system 1.

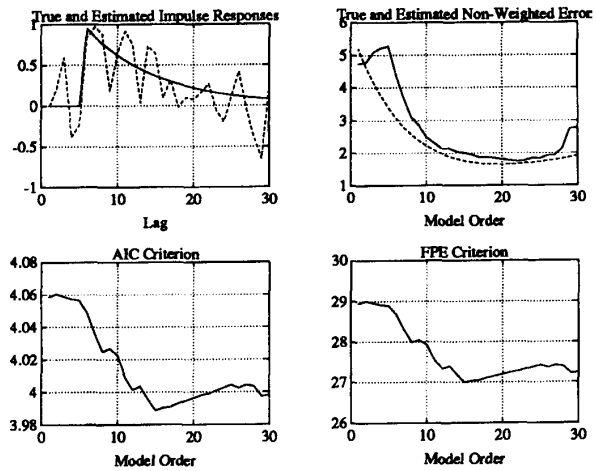


Fig. 6. Full-order estimation, Akaike criteria, and true square error and estimated mean square errors for system 2.

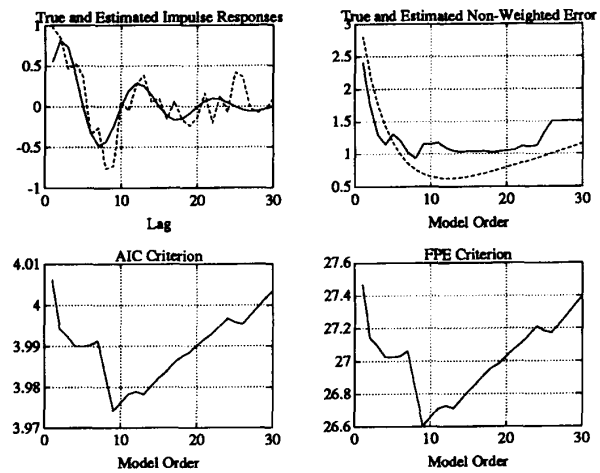


Fig. 7. Full-order estimation, Akaike criteria, and true square error and estimated mean square errors for system 3.

to fit to the data using least squares. It is evident that for finite noisy data, the optimal model order to use for these two common applications is not the true model order. Finally, Fig. 8 shows the same experiment as shown in Fig. 7, save that the measurement noise variance has been tripled. Note that at this higher noise, the upper right diagram in Fig. 8 indicates that lower order models should be fitted to the finite data samples. This is completely in accordance with the bias-variance tradeoff arguments presented earlier in this paper. Note that in the simulations we have not presented GIC (p). This is because here $S_u(\omega) = \sigma_u^2$ are constant so that

$$\text{GIC}\,(p) = J_p^2 \times \sigma_u^2 + \hat{\sigma}_v^2 \tag{113}$$

and hence, in these simulations GIC (p) is parallel to J_p^2, which compares directly to (112).

VIII. Conclusion

In most identification applications, the nominal model is at best, an approximation to the true system whose structure is more complex than that of the parameterized model. This induces an error between the true transfer function and the estimated nominal model which is usually called unmodeled dynamics. One way of treating this error is to estimate it by further parameterizing it as in [7], but this amounts to replacing the nominal model by a more complex one; it amounts to modeling the unmodeled dynamics. In the no noise case treated in [7], increasing the model order is not a problem if the input spectrum is rich enough. However, in the noise corrupted case treated in this paper, increasing the model order increases the error in the estimated parameters and this may increase the total model error.

In this paper we have provided bounds on the undermodeling errors produced by truncating the model order fitted to noisy data. We have done this by assuming that the unmodeled dynamics is a realization of a stochastic process described by a parametrized probability density function. With this stochastic embedding model for the data production mechanism, the transfer function for the unmodeled component is not explicitly represented. Instead, the *class* of functions that the unmodeled transfer function is *likely* to come from is represented in the model. The parameterization of this class can be translated into likely regions in the complex plane in which the frequency response of the true system may lie.

Under the stochastic embedding model, we have shown how the parameters of the stochastic distribution may be estimated from the data. For the case of Gaussian probability density assumptions, simulation shows the resultant error bounds to be highly discriminatory and informative.

Our procedure produces an estimate of the mean square error between the true and estimated nominal transfer functions. This estimate is the sum of two clearly distinguishable terms, one due to the undermodeling (which decreases with model complexity) and one due to noise in the data (which increases with model complexity and decreases with the number of data). Our motivation for generating these mean square errors has been to link identification with robust control.

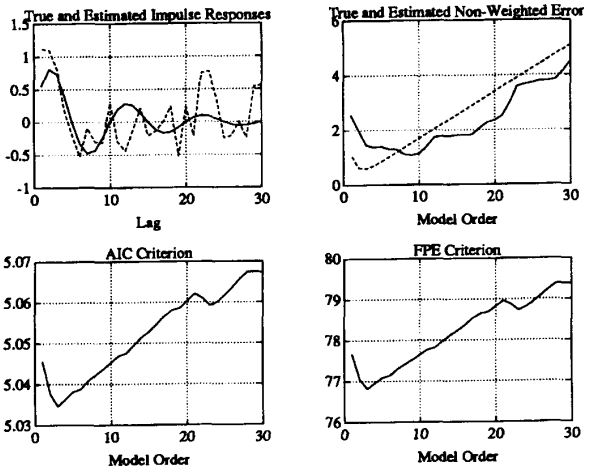


Fig. 8. Full-order estimation, Akaike criterions, and true and estimated mean square errors for system 3 with high noise.

Our expressions for the mean square error have also allowed us to develop a new optimal model-order estimation criterion, GIC. This criterion, explicitly and, we believe, correctly incorporates the effect of undermodeling. It compares favorably with Akaike's FPE, as is demonstrated by both our theoretical analysis and simulations.

Appendix

Lemma A.1: Consider an invertible square matrix Σ parameterized by a set of scalars $\{\alpha_1, \cdots, \alpha_n\}$. The derivative of $\ln \det(\Sigma)$ with respect to one of the parameterizing constants α_k can then be written

$$\frac{\partial \ln \det(\Sigma)}{\partial \alpha_k} = \text{tr}\,\{\Sigma^{-1}\Lambda\}$$

where

$$\Lambda \triangleq \frac{\partial \Sigma}{\partial \alpha_k}.$$

Proof: Application of the chain rule gives

$$\frac{\partial \ln \det(\Sigma)}{\partial \alpha_k} = \sum_i \sum_j \left(\frac{\partial \ln \det(\Sigma)}{\partial \Sigma_{i,j}} \frac{\partial \Sigma_{i,j}}{\partial \alpha_k} \right)$$

$$= \sum_i \sum_j \left(\frac{(\text{Adj}\,\Sigma)_{ji}}{\det(\Sigma)} \lambda_{ji} \right)$$

$$= \text{tr}\,[\Sigma^{-1}\Lambda]. \qquad (\blacksquare)$$

Lemma A.2: Consider an invertible square matrix Σ parameterized by a set of scalars $\{\alpha_1, \cdots, \alpha_n\}$. Then the derivative of Σ^{-1} with respect to any α_k, $k \in [1, n]$ is given by

$$\frac{\partial \Sigma^{-1}}{\partial \alpha_k} = -\Sigma^{-1} \frac{\partial \Sigma}{\partial \alpha_k} \Sigma^{-1}.$$

Proof:

$$\Sigma\Sigma^{-1} = I$$

$$\text{Therefore:} \quad \frac{\partial \Sigma}{\partial \alpha_k}\Sigma^{-1} + \Sigma\frac{\partial \Sigma^{-1}}{\partial \alpha_k} = 0$$

$$\text{Therefore:} \quad \frac{\partial \Sigma^{-1}}{\partial \alpha_k} = -\Sigma^{-1}\frac{\partial \Sigma}{\partial \alpha_k}\Sigma^{-1}. \quad \blacksquare$$

Lemma A.3: Consider a random vector ϵ distributed as

$$\epsilon \sim N(0, \Sigma). \tag{A1}$$

Then

$$\mathcal{E}\{(\epsilon^T\Sigma^{-1}\Lambda\Sigma^{-1}\epsilon)(\epsilon^T\Sigma^{-1}\Gamma\Sigma^{-1}\epsilon)\} = \operatorname{tr}(\Sigma^{-1}\Lambda)\operatorname{tr}(\Sigma^{-1}\Gamma) + 2\operatorname{tr}(\Sigma^{-1}\Lambda\Sigma^{-1}\Gamma)$$

where Λ and Γ are square matrices and $\mathcal{E}\{\cdot\}$ denotes expectation over the probability space on which ϵ is defined.

Proof:

$$\mathcal{E}\{(\epsilon^T\Sigma^{-1}\Lambda\Sigma^{-1}\epsilon)(\epsilon^T\Sigma^{-1}\Gamma\Sigma^{-1}\epsilon)\} = \mathcal{E}\left\{\sum_m\sum_n\sum_j\sum_k \lambda_{mn}\gamma_{jk}\epsilon_m\epsilon_n\epsilon_k\epsilon_j\right\}$$

where

$$(\Sigma^{-1}\Lambda\Sigma^{-1})_{mn} = \lambda_{mn}$$
$$(\Sigma^{-1}\Gamma\Sigma^{-1})_{jk} = \gamma_{jk}$$
$$(\epsilon)_k = \epsilon_k.$$

Furthermore, if $(\Sigma)_{jk} = \sigma_{jk}$, then since ϵ is distributed as in (A1)

$$\mathcal{E}\{\epsilon_m\epsilon_n\epsilon_k\epsilon_j\} = \sigma_{mn}\sigma_{jk} + \sigma_{jm}\sigma_{nk} + \sigma_{jn}\sigma_{mk}.$$

Therefore,

$$\begin{aligned}\mathcal{E}\{&(\epsilon^T\Sigma^{-1}\Lambda\Sigma^{-1}\epsilon)(\epsilon^T\Sigma^{-1}\Gamma\Sigma^{-1}\epsilon)\} \\ &= \sum_m\sum_n \lambda_{mn}\sigma_{mn}\sum_j\sum_k \gamma_{jk}\sigma_{jk} \\ &\quad + \sum_m\sum_n \lambda_{mn}\sum_j \sigma_{jm}\sum_k \gamma_{jk}\sigma_{nk} \\ &\quad + \sum_m\sum_n \lambda_{mn}\sum_j \sigma_{jn}\sum_k \sigma_{mk}\gamma_{jk} \\ &= \operatorname{tr}(\Sigma^{-1}\Lambda)\operatorname{tr}(\Sigma\Gamma) + 2\,\mathrm{tR}(\Sigma^{-1}\Lambda\Sigma^{-1}\Gamma). \quad \blacksquare\end{aligned}$$

Lemma A.4: For Example 3, an unbiased estimate $\bar{\xi}$ of ξ has the following asymptotic properties:

$$\underline{\operatorname{cov}}\left(\frac{\bar{\alpha}}{\alpha}\right) = \mathcal{O}\left(\frac{1}{\ln N}\right) \quad \text{as } N \to \infty \tag{114}$$

$$\underline{\operatorname{cov}}\left(\frac{\bar{\lambda}}{\lambda}\right) = \mathcal{O}\left(\frac{1}{(\ln N)^3}\right) \quad \text{as } N \to \infty \tag{115}$$

$$\underline{\operatorname{cov}}\left(\frac{\bar{\sigma}_v^2}{\sigma_0^2}\right) = \mathcal{O}\left(\frac{1}{N}\right) \quad \text{as } N \to \infty \tag{116}$$

and the estimate $\bar{\sigma}_v^2$ is asymptotically decoupled from the estimates $\bar{\alpha}$ and $\bar{\lambda}$.

Proof: A lower bound on the covariance of an unbiased estimate $\bar{\xi}$ of ξ is found by the Fisher's information matrix $M_\xi(N)$ for N data points which is found by Result 2 with the substitution of

$$T = \operatorname*{diag}_{1\le k\le N}\{N^2\sigma_u^4\lambda^k\} \tag{117}$$

$$K = \operatorname*{diag}_{1\le k\le N}\{N^2\sigma_u^4\alpha k\lambda^{k-1}\} \tag{118}$$

$$\Sigma = \operatorname*{diag}_{1\le k\le N}\{N^2\sigma_u^4\alpha\lambda^k + N\sigma_u^2\sigma_v^2\} \tag{119}$$

$$\Delta = N\sigma_u^2 I \tag{120}$$

as

$$M_\xi(N) = \frac{1}{2}\sum_{k=1}^{N}\frac{1}{\left(\alpha\lambda^k + \dfrac{\sigma_v^2}{N\sigma_u^2}\right)^2} \cdot \begin{bmatrix} \lambda^{2k} & \alpha k\lambda^{2k-1} & \dfrac{\lambda^k}{N\sigma_u^2} \\ \alpha k\lambda^{2k-1} & \alpha^2k^2\lambda^{2k-2} & \dfrac{\alpha k\lambda^{k-1}}{N\sigma_u^2} \\ \dfrac{\lambda^k}{N\sigma_u^2} & \dfrac{\alpha k\lambda^{k-1}}{N\sigma_u^2} & \dfrac{1}{N^2\sigma_u^4} \end{bmatrix}. \tag{121}$$

Now, if we define $\bar{\mu}_N$ by

$$\bar{\mu}_N^T \triangleq \left[\sqrt{\frac{-\ln N}{\ln\lambda}}\,\frac{\hat{\alpha}}{\alpha}, \left(\frac{-\ln N}{\ln\lambda}\right)^{3/2}\frac{\hat{\lambda}}{\lambda}, \sqrt{N}\,\frac{\hat{\sigma}_v^2}{\sigma_v^2}\right] \tag{122}$$

then, if $\bar{\mu}_N$ is an unbiased estimate of μ, the covariance in the estimate satisfies the Cramér–Rao lower bound P_μ where P_μ is the inverse of the limit as $N \to \infty$ of $M_\mu(N)$, Fisher's information matrix for μ. By the definitions for μ and ξ we have

$$\mu = S\xi \tag{123}$$

where

$$S = \begin{bmatrix} \dfrac{1}{\alpha}\sqrt{\dfrac{-\ln N}{\ln\lambda}} & 0 & 0 \\ 0 & \dfrac{1}{\lambda}\left(\dfrac{-\ln N}{\ln\lambda}\right)^{3/2} & 0 \\ 0 & 0 & \dfrac{\sqrt{N}}{\sigma_v^2} \end{bmatrix}. \tag{124}$$

Therefore,

$$M_\mu(N) = S^{-1}M_\xi(N)S^{-1}. \tag{125}$$

Now

$$S^{-1}M_\xi(N)S^{-1} \to \frac{1}{2}\begin{bmatrix} 1 & \dfrac{1}{2} & 0 \\ \dfrac{1}{2} & \dfrac{1}{3} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

pointwise as $N \to \infty$. (126)

A proof of (126) may be found in [2]. Therefore, for an unbiased estimate $\bar{\mu}_N$ of μ we have

$$\operatorname{cov}\{\bar{\mu}_N\} \geq P_\mu \tag{127}$$

where

$$P_\mu = 2\begin{bmatrix} 4 & -6 & 0 \\ -6 & 12 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{128}$$

■

Acknowledgment

The results presented in this paper have been obtained with the framework of the Interuniversity Attraction Poles initiated by the Belgian State, Prime Minister's Office, Science Policy Programming. The Scientific responsibility rests with its authors.

The first two authors would like to thank L. Ljung, H. Hjalmarsson, and B. Wahlberg for an inspirational meeting in the Swedish countryside that got their minds properly focused on questions of bias, variance, undermodeling, and other noisy issues.

References

[1] R. O. La Maire, L. Valavani, M. Athans, and G. Stein, "A frequency domain estimator for use in adaptive control systems," in *Proc. ACC*, 1987, pp. 238-244.

[2] B. M. Ninness, "Robust estimation," Ph.D. dissertation, Univ. Newcastle, New South Wales, Australia, 1992.

[3] B. Younce, "Identification with Nonparametric Uncertainty," Ph.D. dissertation, Univ. Notre Dame, Notre Dame, IN, 1989.

[4] Donald E. Catlin, *Estimation, Control, and the Discrete Kalman Filter*. New York: Springer-Verlag, 1989.

[5] E. Fogel, "System identification via membership set constraints with energy constrained noise," *IEEE Trans. Automat. Contr.*, vol. AC-24, no. 5, pp. 615-622, 1979.

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

[7] E. W. Bai, "Adaptive quantification of model uncertainties by rational approximation," *IEEE Trans. Automat. Contr.*, vol. 36, no. 4, pp. 441-453, 1991.

[8] G. C. Goodwin, M. Gevers, and D. Q. Mayne, "Bias and variance distribution in transfer function estimates," in *Proc. 9th IFAC Symp. Identif. Syst. Parameter Estimation*, Budapest, July 1991.

[9] G. C. Goodwin and B. M. Ninness, "Model error quantification for robust control based on quasi-bayesian estimation in closed loop," in *Proc. CDC*, 1991.

[10] G. C. Goodwin and R. L. Payne, *Dynamic System Identification*. New York: Academic, 1977.

[11] G. C. Goodwin and M. Salgado, "Quantification of uncertainty in estimation using an embedding principle," in *Proc. of ACC*, Pittsburgh, PA, 1989.

[12] G. C. Goodwin and M. Salgado, "A stochastic embedding approach for quantifying uncertainty in the estimation of restricted complexity models," *Int. J. Adaptive Contr. Signal Processing*, vol. 3, no. 4, pp. 333-356, 1989.

[13] H. Hjalmarsson, "On estimation of model quality in system identification," *Licentiate Thesis* LIU-TEK-LIC-1990:51, Linkoping Univ., Linkoping, Sweden, 1990.

[14] R. L. Kosut, "Adaptive control via parameter set estimation," *Int. J. Adaptive Contr. Signal Processing*, vol. 2, no. 4, pp. 371-400, 1988.

[15] ——, "Adaptive robust control via transfer function uncertainty estimation," in *Proc. ACC*, Atlanta, GA, 1988.

[16] R. L. Kosut, M. Lau, and S. Boyd, "Identification of systems with parametric and nonparametric uncertainty," in *Proc. Amer. Contr. Conf.*, 1990, pp. 2412-2417.

[17] L. Ljung, *System Identification: Theory for the User*. Englewood Cliffs, NJ: Prentice-Hall, 1987.

[18] ——, "Asymptotic variance expressions for identified black box transfer function models," *IEEE Trans. Automat. Contr.*, vol. AC-31, no. 2, pp. 134-144, 1986.

[19] P. M. Makila, "Approximation of stable systems by Laguerre filters," *Automatica*, vol. 26, pp. 333-345, 1990.

[20] ——, "Laguerre series approximation of infinite dimensional systems," *Automatica*, vol. 26, pp. 985-995, 1990.

[21] G. C. Goodwin, D. Q. Mayne, and M. Salgado, "Uncertainty, information and estimation," presented at the IFAC Symp. Adaptive Contr. Signal Processing, 1989.

[22] G. C. Goodwin, B. M. Ninness, and M. Salgado, "Quantification of uncertainty in estimation," in *Proc. Amer. Contr. Conf.*, 1990, pp. 2400-2405.

[23] J. P. Norton, "Identification of parameter bounds of armax models from records with bounded noises," *Int. J. Contr.*, vol. 42, pp. 375-390, 1987.

[24] P. J. Parker, "Frequency domain descriptions of linear systems," Ph.D. dissertation Australian National Univ., Canberra, 1988.

[25] P. J. Parker and R. R. Bitmead, "Adaptive frequency response identification," in *Proc. 26th Conf. Decision Contr.*, 1987, pp. 348-353.

[26] M. E. Salgado, "Issues in robust identification," Ph.D. dissertation, Univ. Newcastle, New South Wales, Australia, 1989.

[27] B. Wahlberg, "System identification using laguerre models," *IEEE Trans. Automat. Contr.*, vol. 36, no. 5, 1991.

[28] B. Wahlberg and L. Ljung, "Design variables for bias distribution in transfer function estimation," *IEEE Trans. Automat. Contr.*, vol. AC-31, pp. 134-144, 1986.

[29] B. Wahlberg and L. Ljung, "Hard frequency-domain model error bounds from least squares like identification techniques," Dep. Elec. Eng., Linkoping Univ., Linkoping, Sweden, Tech. Rep. LITH-ISY-1144, 1990.

[30] R. C. Younce and C. E. Rohrs, "identification with parameteric and nonparametric uncertainty," in *Proc. Int. Conf. Circ. Syst.*, 1990.

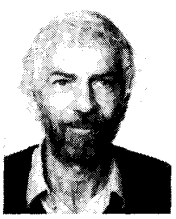
[31] C. C. Zervos and G. A. Dumont, "Deterministic adaptive control based on laguerre series representation," *Int. J. Contr.*, vol. 48, pp. 2333-2359, 1988.

[32] M. Gevers, "Connecting identification and robust control: A new challenge," preprint, presented at the IFAC Symp. Indent., Budapest, 1991.

Graham C. Goodwin (M'74-SM'84-F'86) was born in Broken Hill, Australia, in 1945. He received the B.Sc degree in physics, the B.E. degree in electrical engineering, and the Ph.D. degree, from the University of New South Wales, Australia.

From 1970 until 1974 he was a lecturer in the Department of Computing and Control, Imperial College, London. Since 1974 he has been with the Department of Electrical Engineering and Computer Science, University of Newcastle, Australia.

Dr. Goodwin is a Fellow of the Australian Academy of Technology, Science and Engineering. He is the recipient of several international prizes including a Best Paper Award by IEEE TRANSACTIONS ON AUTOMATIC CONTROL, and Best Engineering Textbook Award from the International Federation of Automatic Control. He is the coauthor of four books: *Control Theory* (Oliver and Boyd, 1970), *Dynamic System Identification* (New York: Academic, 1977), *Adaptive Filtering, Prediction and Control* (Englewood Cliffs, NJ: Prentice-Hall, 1984), and *Digital Control and Estimation* (Englewood Cliffs, NJ: Prentice-Hall, 1989). He is currently Professor of Electrical Engineering and Director of the Centre for Industrial Control Science at the University of Newcastle, New South Wales, Australia.

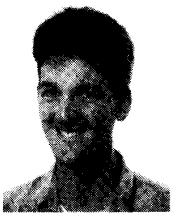


Michel Gevers (S'66-S'70-M'72-SM'86-F'90) was born in Antwerp, Belgium, in 1945. He received the electrical engineering degree from Louvain University, Louvain-la-Neuve, Belgium, in 1968, and the Ph.D. degree from Stanford University, Stanford, CA, in 1972.

He is currently a Professor and Head of the Laboratoire d'Automatique, Dynamique et Analyse des Systèmes at Louvain University, Louvain-la-Neuve, Belgium. He has spent long-term visits in several universities, including the University of Newcastle, New South Wales, Australia, the Technical University of Vi-

enna, and a three-year term at the Australian National University. His research interests are in system identification, adaptive estimation and control, multivariable system theory, optimal control and filtering, and the numerical aspects of filter and controller design. He is a coauthor with R. R. Bitmead and V. Wertz of *Adaptive Optimal Control—The Thinking Man's GPC* (Englewood Cliffs, NJ: Prentice-Hall, 1990) and with G. Li of *Parametrizations in Control, Estimation and Filtering Problems: Accuracy Aspects* (Communication and Control Engineering Series) (New York: Springer-Verlag, 1992).

Prof. Gevers has been the Associate Editor of *Automatica* and the IEEE TRANSACTIONS ON AUTOMATIC CONTROL, and is presently the Associate Editor of *Mathematics of Control, Signals, and Systems*.



Brett Ninness was born in Singleton, Australia, in 1963. He completed the B.E. degree in 1986, the M.E. degree in 1990, and is currently pursuing the Ph.D. degree all in electrical engineering at the University of Newcastle, New South Wales, Australia. In the time between the B.E. and M.E. degrees he studied medicine for a short time and completed the B.Math course while pursuing the M.E. and Ph.D degrees.

His research interests are adaptive control and signal processing and he is particularly interested in their application to the field of medicine.