The Utility of Orthonormal Bases

Brett Ninness*

Abstract

There has been recent interest in using orthonormalised forms of fixed denominator model structures for system identification. However, modulo numerical conditioning considerations, the transfer function estimates obtained by using these sometimes complex to implement structures are *identical* to that obtained by simply pre-filtering the input with an all pole denominator (with poles the same as the orthonormal form) and using pre-existing software for FIR identification to estimate the numerator coefficients. In recognition of this, the report here provides detailed comment on the utility of using orthonormally parameterised model structures in a system identification setting.

Technical Report EE9802, Department of Electrical and Computer Engineering, University of Newcastle,AUSTRALIA

1 Introduction

There has been a recent explosion of interest in the use of rational orthonormal bases in system theoretic settings, and in particular in a system identification setting. See, for example, the work [19, 79, 54, 70, 74, 34, 63, 56, 6, 15, 4]. The purpose of this report is to closely examine the utility of some of these ideas.

In particular, let us consider the Devil's advocate position by posing the provocative challenge:

> *Surely 'orthonormal' is just an impressive sounding word. All that is really being done is that very simple model structures are being*

*This work was supported by the Australian Research Council and the Centre for Integrated Dynamics and Control (CIDAC). It was partly completed while on leave at the Department of Sensors, Signals and Systems-Automatic Control, The Royal Institute of Technology, S-100 44 Stockholm, Sweden. This author is with the Department of Electrical and Computer Engineering, University of Newcastle, Australia and can be contacted at email:brett@ee.newcastle.edu.au or FAX: +61 2 49 21 69 93

> *reparameterised in an equivalent form. Who cares whether this form is 'orthonormal' or not? Let's not confuse changes in model structure with fundamental changes in methods of estimation.*

For example, historically right from the 1950's much of the motivation for using orthonormal bases in a system identification or system approximation setting has arisen from the desire to provide parsimonious representation of systems by a strategy of fixing poles near where the poles of the underlying dynamics are believed to lie.

However, once this idea is accepted then one soon realises that actually, the orthonormality property is not responsible for the advantage of achieving parsimony. It is achieved solely by the idea of fixing poles at arbitrary points. But this can be achieved in very simple ways. For example, it can be realised by using the simple model structure ($\beta^T = [b_1, b_2, \cdots, b_n]$ is a vector of parameters)

$$G(q, \beta) = \frac{b_1 + b_2 q + \cdots + b_n q^{n-1}}{D_n(q)} \tag{1}$$

where

$$D_n(q) = \prod_{k=1}^{n} (q - \xi_k)$$

is a fixed denominator. With the extra considerations (see for example [56, 63]) required to derive the orthonormal bases $\{\mathcal{B}_k(q)\}$, the strategy of using (1) may be much simpler than using the equivalent orthonormalised form

$$G(q, \theta) = \sum_{k=1}^{n} \theta_k \mathcal{B}_k(q) \tag{2}$$

where the $\{\mathcal{B}_k(q)\}$ also have fixed poles taken from the set $\{\xi_1, \cdots, \xi_n\}$.

Just how equivalent are these model structures (1) and (2)? Well, if least squares estimation is performed relative to them, then although different parameter space estimates $\widehat{\beta}_N^n$ and $\widehat{\theta}_N^n$ will ensue, the frequency domain estimates will be *identical*:

$$G(e^{j\omega}, \widehat{\beta}_N^n) = G(e^{j\omega}, \widehat{\theta}_N^n). \tag{3}$$

Additionally, these estimates can be found with respect to the simple structure (1) wherein a numerator is estimated by just using pre-existing software for estimation of FIR models (the numerator) after having pre-filtered the input u_t via the all-pole filter $1/D_n(q)$. However, if the estimate $\widehat{\theta}_N^n$ is to be found with respect to the orthonormal structure (2), then specialised software needs to be written. Yet (from (3)) it will lead to the same frequency response estimate as if the 'simpler' structure (1) had been used.

However, in spite of its directness, this Devils advocate argument is seriously flawed, and the flaws are the reason why consideration of orthonormalised forms such as (2) are so important. To be specific,

Numerical Considerations: The above argument of equivalence of estimates was made modulo numerical conditioning considerations. In fact, it turns out that estimating the frequency response $G(e^{j\omega}, \widehat{\beta}_N^n)$ with respect to the model structure (1) is very often horribly ill conditioned to the point where it fails on modern high performance workstations. In these same cases, the estimation problem with respect to the orthonormalised form is well conditioned. Therefore, implementation with respect to the orthonormal form is very often the only way of implementing the idea of fixing the poles in a model structure.

Analysis of Estimation Error: As just explained, the whole point of employing model structures with fixed poles is to arrive at more efficient parameterisations. Why is this desirable? Well, it is well known that the sensitivity of a frequency response estimate $G(e^{j\omega}, \widehat{\theta}_N^n)$ is proportional to the model order n. If this can be made as small as possible, then the noise induced error can be minimised. At the same time, this is no good if by decreasing n the model structure is so restricted that it cannot approximate the true dynamics $G(q)$ very well. Fixing poles near where it is believed the true poles of $G(q)$ is meant to avoid this.

But surely if this pole fixing strategy is to be implemented, then we should provide some analysis to:

1. Show how the sensitivity to noise depends on the choice of the fixed poles $\{\xi_k\}$ and the model order n.
2. Show how the undermodelling induced error (bias error) is affected by the choice of these same fixed poles and the model order n.

It turns out that providing these answers is impossible if one attempts the analysis of the simple 'fixed denominator' structure (1) directly. However, it is possible if one attempts the error analysis with respect to the orthonormal structure (2) and *exploits the orthogonality*. So the orthonormal parameterisation is very useful as an *analysis tool* whether or not it is chosen as an implementational option. However, based on the numerical considerations just mentioned, we would suggest it be used as the implementational tool as well.

Extensibility to other Model Structures. Continuing on this theme of analytical expediency, as we'll illustrate, even when you employ model structures that seem to have nothing at all to do with the orthonormal structure

(2), the orthonormal parameterisation is still vitally important. For example, suppose you use an ARX or Box-Jenkins structure where the poles are not fixed, but in fact are estimated. Then this would certainly seem to have little to do with the orthonormal parameterisation (2) where the poles are fixed.

However, it turns out that to most accurately quantify the sensitivity of *any* prediction error estimate with respect to *any* model structure in such a way as to accurately account for the effect of data pre-filtering (which is a very common and useful operation), it is necessary to use an expression which is couched in terms of an orthonormal basis $\{\mathcal{B}_k(q)\}$ and in fact is derived by consideration of an equivalent orthonormally parameterised situation.

The purpose of this report is to delve into these three key factors (although there are others, see for example [5]) contributing to the utility of rational orthonormal structures in a system identification setting. The detailed examination conducted here is meant to most clearly expose the vital role of orthonormal bases. At the same time, a more detailed exposition requires greater effort on the part of the reader. In recognition of this, a brief synopsis of the main points to be made are as follows.

Pages 8 to 12 Take two model structures, one orthonormal and one not, and denote the associated co-variance matrices as (respectively) R_ϕ and R_γ. The matrices are $n \times n$ if there are n parameters estimated, they depend on the spectral density $\Phi_u(\omega)$ of u_t and the numerical conditioning associated with finding the estimates is governed by the condition numbers $\kappa(R_\phi)$ and $\kappa(R_\gamma)$ of R_ϕ and R_γ. By the orthonormal property, when the input is white, $\kappa(R_\phi) = 1$ which is perfect numerical conditioning.

Page 12 For non-white Φ_u an upper bound on the condition number $\kappa(R_\phi)$ associated with the orthonormal model structure exists, but no such bound exists for $\kappa(R_\gamma)$ pertaining to the non-orthonormal structure.

Page 14 Although R_ϕ and R_γ depend on Φ_u which is infinite dimensional, they can only be 'moved' by choice of Φ_u in an n dimensional manifold of the $n(n+1)/2$ dimensional manifold of symmetric $n \times n$ matrices. This highly restricted mobility helps provide orthonormal bases with improved numerical conditioning across a range of coloured input spectrums.

Page 15 For R_ϕ associated with the orthonormal structure (2) and R_γ associated with the non-orthonormal structure (1) the following approximations hold

$$\kappa(R_\phi) \approx \frac{\max_\omega \Phi_u(\omega))|^2}{\min_\omega \Phi_u(\omega)}., \qquad \kappa(R_\gamma) \approx \frac{\max_\omega \Phi_u(\omega)/|D_n(e^{j\omega})|^2}{\min_\omega \Phi_u(\omega)/|D_n(e^{j\omega})|^2}.$$

Page 18 The total estimation error between the n'th order estimate $G(e^{j\omega}, \widehat{\theta}_N^n)$ and the true response $G(e^{j\omega})$ can be split into two parts as

$$G(e^{j\omega}, \widehat{\theta}_N^n) - G(e^{j\omega}) = G_b(e^{j\omega}) + G_\nu(e^{j\omega})$$

where $G_b(e^{j\omega})$ called the 'bias error' is due only to undermodelling and $G_\nu(e^{j\omega})$ called the 'variance error' is due only to the measurement noise v_t.

Page 19 By virtue of orthonormality, when using any model structure like (1) or (2) the bias error is such that on average over frequency and weighted with $\Phi_u(\omega)$ the magnitude of the frequency response is underestimated.

Page 20 By introducing the idea of Kernel functions, which are defined in terms of orthonormal bases, the bias error can be decomposed into two terms, one which is dependent only on the model structure, and one which is dependent on the estimation method.

Page 23 A very important characterisation of the bias error involved with estimating

$$G(z) = \sum_{i=0}^{m-1} \frac{\alpha_i}{z - p_i^0}$$

when Φ_u is white was given as

$$|G_b(e^{j\omega})| \leq \sum_{i=0}^{m-1} \left| \frac{\alpha_i}{e^{j\omega} - p_i^0} \right| \prod_{k=0}^{n-1} \left| \frac{p_i^0 - \xi_k}{1 - \overline{\xi_k} p_i^0} \right| .$$

On page 15 the derivation of this result is outlined in order to illustrate the vital role that orthornormal bases play in providing it. The starting point is the decomposition of presented page 20.

Page 24 An algorithm is presented for numerically calculating the bias error for any pre-supposed $G(z)$, $\Phi_u(\omega)$ and pole choices $\{\xi_k\}$.

Page 28 A famous result due to Lennart Ljung is that the variance error can be approximated as

$$\mathbf{E}\left\{|G_\nu(e^{j\omega})|^2\right\} \approx \frac{n}{N} \frac{\Phi_\nu(\omega)}{\Phi_u(\omega)} \tag{4}$$

and that this holds for a very wide range of model structures. On page 28 we point out that this expression can be a very poor approximation for the

model structures considered in these notes where many poles $\{\xi_k\}$ may not be chosen at the origin, and that instead the approximation

$$\mathbf{E}\left\{|G_\nu(e^{j\omega})|^2\right\} \approx \frac{1}{N}\frac{\Phi_\nu(\omega)}{\Phi_u(\omega)}\sum_{k=1}^{n}|\mathcal{B}_k(e^{j\omega})|^2. \tag{5}$$

should be used. This gives (4 as a special case if all the poles are chosen at the origin.

Pages 38 to 40 The most general formulation of model structures for system identification is that the relationship of input data u_t to output data y_t and measurement noise $v_t = H(q)e_t$ be modelled as

$$F(q)y_t = G(q,\theta)F(q)u_t + H(q,\theta)e_t.$$

where $F(q)$ is a data pre-filter. It is pointed out on page 40 that *regardless of the model structure* used for $G(q,\theta)$ and $H(q,\theta)$, in any case where the data pre-filter is of the form $1/D_n(q)$ then (5) (with poles in the orthonormal $\{\mathcal{B}_k(q)\}$ chosen the same as in $1/D_n(q)$) should be used in place of (4). This is illustrated numerically on pages 38 to 40 for the ARX model structure case of

$$G(q,\theta) = \frac{B(q,\theta)}{A(q,\theta)}, \qquad H(q,\theta) = \frac{1}{A(q,\theta)}.$$

This illustrates that orthonormal parameterisations play a *fundamental and intrinsic role* in system identification, since regardless of whether they are used for the actual implementation of the algorithm, they quantify the bias and variance error.

Pages 32 to 36 Some fundamental tools for using general orthonormal bases for the purposes of analysing system theoretic problems are over-viewed. Namely, it is first acknowledged that the trigonometric basis $\{\mathcal{B}_n(z) = z^{-n}\}$ is already widely used for this purpose by defining $n \times n$ Toeplitz matrices $T_n(f)$ via a spectral density $f(\omega)$ as

$$[T_n(f)]_{m,\ell} = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-j\omega m}e^{j\omega\ell}f(\omega)\,\mathrm{d}\omega$$

and then using the algebraic properties that for g also a spectral density

$$T_n(f)T_n(g) \approx T_n(fg), \qquad T_n^{-1}(f) \approx T_n(1/f).$$

As well, the following Fourier reconstruction property is vital

$$\lim_{n\to\infty}\frac{1}{n}\sum_{m=1}^{n}\sum_{\ell=1}^{n}e^{j\omega m}e^{-j\omega\ell}[T_n(f)]_{m,\ell} = f(\omega).$$

On the pages around page 36 it is outlined that this can be generalised to the case of matrices $M_n(f)$ defined as

$$[M_n(f)]_{m,\ell} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathcal{B}_m(e^{j\omega})\overline{\mathcal{B}_\ell(e^{j\omega})} f(\omega)\,\mathrm{d}\omega$$

for which the algebraic properties

$$M_n(f)M_n(g) \approx M_n(fg), \qquad M_n^{-1}(f) \approx M_n(1/f)$$

can also be shown to hold and also the generalised Fourier reconstruction property

$$\lim_{n\to\infty}\left(\sum_{k=1}^{n}|\mathcal{B}_k(e^{j\omega})|^2\right)^{-1}\sum_{m=1}^{n}\sum_{\ell=1}^{n}\overline{\mathcal{B}_m(e^{j\omega})}\mathcal{B}_\ell(e^{j\omega})[M_n(f)]_{m,\ell} = f(\omega).$$

holds. The classical results arise from these generalised ones as special cases of $\xi_k = 0$. The existence of such generalised results are considered a key reason for considering orthonormal parameterisations in system theoretic settings.

With this introductory and overview material dispensed with, a more detailed treatment begins with the consideration of the role of orthonormal bases in providing improved numerical conditioning.

2 Numerical Conditioning Issues

The focus in this set of notes is on estimation problems where one has available N point data records of an input sequence u_t and output sequence y_t of a linear time invariant system and it is assumed that this data is generated as follows

$$y_t = G(q)u_t + \nu_t.$$

Here $G(q)$ is a stable (unknown) transfer function describing the system dynamics that are to be identified by means of the observations u_t, y_t, and the sequence ν_t is some sort of possible noise corruption. The input sequence u_t is assumed to be quasi–stationary in the sense used by Ljung [44]. This essentially means that certain sample time averages of the signal u_t exist as follows

$$R_u(\tau) \triangleq \lim_{N\to\infty} \frac{1}{2N} \sum_{t=-N}^{N} \mathbf{E}\left\{u_t u(t-\tau)\right\}.$$

This allows the 'spectral density' (distribution of energy with respect to frequency) $\Phi_u(\omega)$ of the signal u_t to be defined as

$$\Phi_u(\omega) \triangleq \sum_{\tau=-\infty}^{\infty} R_u(\tau) e^{-j\omega\tau}.$$

Not that by construction $\Phi_u(\omega)$ can never by negative (it represents energy after all), however in this report, it is also assumed that $\Phi_u(\omega) \neq 0$ for any ω.

The method of estimating the dynamics $G(q)$ which is of interest here is one wherein the following 'fixed denominator' model structure is used

$$G(q,\beta) = \sum_{k=1}^{n} b_k \mathcal{G}_k(q) \tag{6}$$

where the $\{b_k\}$ are real valued co-efficients and the transfer functions $\{\mathcal{G}_k(q)\}$ may be chosen in various ways, but in every case the poles of the transfer functions $\{\mathcal{G}_k(q)\}$ are selected from the set $\{\xi_1, \xi_1, \cdots, \xi_n\} \subset \mathbf{D}$.

For example, the structure (1) corresponds to

$$\mathcal{G}_k(q) = \frac{q^k}{D_n(q)}, \qquad D_n(q) = \prod_{k=1}^{n} (q - \xi_k) \tag{7}$$

for $k = 0, 1, \cdots, n-1$. The fixed poles $\{\xi_k\}$ are chosen by the user to reflect prior knowledge of the nature of $G(q)$. That is, in the interests of improved estimation accuracy, they are chosen as close as possible to where it is believed the true poles lie. See the papers [70, 34, 58] for detailed discussion of this, which is revisited in § 3 of this report.

An advantage of the simple model structure (6) is that it is linearly parameterised in $\{b_k\}$, so that with $\beta \triangleq [b_1, b_2, \cdots, b_n]^T$ then the least-squares estimate

$$\widehat{\beta}_N^n = \arg\min_{\beta \in \mathbf{R}^n} \left\{ \frac{1}{N} \sum_{t=1}^{N} [y_t - G(q,\beta)u_t]^2 \right\} \tag{8}$$

is easily computed. Specifically, the solution $\widehat{\beta}_N^n$ to (8) can be written in closed form once the model structure (6) is cast in familiar linear regressor form notation as $G(q,\beta)u_t = \gamma_t^T \beta$ where

$$\gamma_t = \Lambda_n(q)\, u_t, \qquad \Lambda_n(q) \triangleq [\mathcal{G}_1(q), \mathcal{G}_1(q), \cdots, \mathcal{G}_n(q)]^T \tag{9}$$

so that (8) is solved as

$$\widehat{\beta}_N^n = \left(\sum_{t=1}^{N} \gamma_t \gamma_t^T \right)^{-1} \sum_{t=1}^{N} \gamma_t y_t \tag{10}$$

provided that the input is persistently exciting enough for the indicated inverse to exist.

However, a large literature [70, 74, 34, 63, 4, 56] (of which these notes are intended to provide a tutorial overview) has developed suggesting that instead of using the model structure (6), one should instead use its so-called 'orthonormal' form. That is, the model structure (6) should be re-parameterised as

$$G(q,\theta) = \sum_{k=1}^{n} \theta_k \mathcal{B}_k(q) \tag{11}$$

where now the $\{\mathcal{B}_k(q)\}$ are transfer functions such that

$$\text{Span}\{\mathcal{G}_1, \mathcal{G}_1, \cdots, \mathcal{G}_n\} = \text{Span}\{\mathcal{B}_1, \mathcal{B}_1, \cdots, \mathcal{B}_n\} \tag{12}$$

but also such that the $\{\mathcal{B}_k(q)\}$ are orthonormal with respect to the inner product

$$\langle \mathcal{B}_n, \mathcal{B}_m \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathcal{B}_n(e^{j\omega}) \overline{\mathcal{B}_m(e^{j\omega})}\, \mathrm{d}\omega = \frac{1}{2\pi j} \oint_{\mathbf{T}} \mathcal{B}_n(z) \overline{\mathcal{B}_m(z)}\, \frac{\mathrm{d}z}{z} = \begin{cases} 1 & ; m = n \\ 0 & ; m \neq n \end{cases} \tag{13}$$

where $\mathbf{T} \triangleq \{z \in \mathbf{C} : |z| = 1\}$ is the complex unit circle.

As illustrated by the work [63, 56, 4], there are a variety of ways of generating rational basis functions that have poles $\{\xi_k\}$ at pre-specified points in $\mathbf{D}$ and also satisfy the orthonormality condition (13). However, in this report, the particular choice analysed in [56]

$$\mathcal{B}_n(q) = \frac{\sqrt{1 - |\xi_n|^2}}{q - \xi_n} \prod_{k=1}^{n} \left(\frac{1 - \overline{\xi}_k q}{q - \xi_k} \right) \tag{14}$$

will be considered. The reason behind using the formulation (14) is that some calculations will be required that depend upon having an explicit formula such as (14) for the $\{\mathcal{B}_k(q)\}$ basis functions, rather than having them defined via a state space construction. Note, however, that all the orthonormal basis construction presented in [56, 63, 4] (that span the same space) are unitarily related and hence possess identical numerical conditioning properties.

In any event, defining in a manner analogous to (9) the 'regression vector' ϕ_t and transfer function vector $\Gamma_n(q)$ as

$$\phi_t \triangleq \Gamma_n(q)u_t, \qquad \Gamma_n(q) \triangleq [\mathcal{B}_1(q), \mathcal{B}_1(q), \cdots, \mathcal{B}_n(q)]^T \tag{15}$$

then the least squares estimate with respect to the model structure (11) is given as

$$\widehat{\theta}_N^n = \left(\sum_{t=1}^N \phi_t\phi_t^T\right)^{-1} \sum_{t=1}^N \phi_t y_t. \tag{16}$$

A key point is that under the span preserving condition (12) there is a linear relationship $\phi_t = J\gamma_t$ for some non-singular J. Therefore, $\widehat{\beta}_N^n = J^T\widehat{\theta}_N^n$ and hence modulo numerical issues the least-squares frequency response estimate is invariant to the change in model structure between (6) and (11). Specifically:

$$\begin{aligned}
G(e^{j\omega}, \widehat{\beta}_N^n) &= \Lambda_n^T(e^{j\omega})\widehat{\beta}_N^n \\
&= \Lambda_n^T(e^{j\omega}) \left(\sum_{t=1}^N \gamma_t\gamma_t^T\right)^{-1} \sum_{t=1}^N \gamma_t y_t \\
&= \Lambda_n^T(e^{j\omega}) \left[J^{-1}\left(\sum_{t=1}^n \phi_t\phi_t^T\right)J^{-T}\right]^{-1} J^{-1}\sum_{t=1}^N \phi_t y_t \\
&= [J\Lambda_n(e^{j\omega})]^T \left(\sum_{t=1}^N \phi_t\phi_t^T\right)^{-1} \sum_{t=1}^N \phi_t y_t \\
&= \Gamma_n^T(e^{j\omega})\widehat{\theta}_N^n \\
&= G(e^{j\omega}, \widehat{\theta}_N^n).
\end{aligned}$$

Given this exact equivalence of frequency response estimates, the purpose of this section is to critically examine the motivation for using the structure (14) (which is complicated by the precise definition of the orthonormal bases (14) or whichever other one detailed in [56, 34, 4] is used) in place of some other one such as (6). For example, if the choice (7) is made then (as already mentioned) estimation with respect to this model structure can be done by using pre-existing software for FIR estimation after first pre-filtering u_t with the all-pole filter $1/D_n(q)$.

However, a vital point is that in practice, for any problem of reasonable size, estimates defined by least-squares cost criteria like (8) are *never* solved by explicitly using closed form formula such as (10) or (16) since to do so courts disaster from inaccuracies accrued by the reality of finite precision arithmetic. Instead, the following matrix/vector equation

$$Y = \Phi\theta \Rightarrow \Phi^T Y = \Phi^T \Phi\theta \tag{17}$$

where

$$\begin{aligned} Y^T &\triangleq [y(1), y(2), \cdots, y(N)] \\ \Phi^T &\triangleq [\phi(1), \phi(2), \cdots, \phi(N)] \end{aligned}$$

is solved for θ (with result $\widehat{\theta}_N^n$) by a process of something like QR factorising $\Phi^T\Phi$ and then 'back-substituting' to find $\widehat{\theta}_N^n$. Note that the N equations represented in matrix notation on the left hand side of (17) are commonly referred to as the 'normal equations'.

Regardless of whether this, or some other numerically robust procedure is employed, a theoretical analysis of how errors due to finite precision numerical representation influence the calculation of $\widehat{\theta}_N^n$ may be performed by representing these errors as ϵ perturbations of $\Phi^T\Phi$ and $\Phi^T Y$ in (respectively) directions Ψ and Z as follows

$$\left[\Phi^T\Phi + \epsilon\Psi\right] \widehat{\theta}_N^n(\epsilon) = \Phi^T Y + \epsilon Z$$

where the notation $\widehat{\theta}_N^n(\epsilon)$ indicates the influence of the solution on the error ϵ. In this case, it can be shown that [31]

$$\frac{\|\widehat{\theta}_N^n(\epsilon) - \widehat{\theta}_N^n(0)\|}{\|\widehat{\theta}_N^n(0)\|} \leq \epsilon \|(\Phi^T\Phi)^{-1}\| \|\Phi^T\Phi\| \left(\frac{\|\Psi\|}{\|\Phi^T\Phi\|} + \frac{\|Z\|}{\|\Phi^T Y\|} \right) + O(\epsilon^2) \tag{18}$$

where the matrix norm is the operator one induced by the norm chosen for the vector quantities. Therefore, defining

$$R_\phi(N) \triangleq \frac{1}{N}\Phi^T\Phi = \frac{1}{N}\sum_{t=1}^{N} \phi_t \phi_t^T, \qquad R_\gamma(N) \triangleq \frac{1}{N}\sum_{t=1}^{N} \gamma_t \gamma_t^T$$

(where the vectors γ_t and ϕ_t are defined in (9) and (15) respectively) then according to (18) to a first order (in ϵ) of approximation, the normalised error (left hand side of (18)) associated with least squares estimation using the model structures (6) and (11) is governed by the size of the so-called condition numbers $\kappa(R_\gamma(N))$ and $\kappa(R_\phi(N))$ where the latter is defined for a square matrix R as

$$\kappa(R) \triangleq \|R\| \, \|R^{-1}\|. \tag{19}$$

However, by the quasi-stationarity assumption on u_t and by Parseval's Theorem, the following limits exist

$$R_\gamma \triangleq \lim_{N\to\infty} \frac{1}{N}\sum_{t=1}^{N} \gamma_t \gamma_t^T = \frac{1}{2\pi}\int_{-\pi}^{\pi} \Lambda_n(e^{j\omega})\Lambda_n^\star(e^{j\omega})\Phi_u(\omega)\,\mathrm{d}\omega \tag{20}$$

$$R_\phi \triangleq \lim_{N\to\infty} \frac{1}{N}\sum_{t=1}^{N} \phi_t \phi_t^T = \frac{1}{2\pi}\int_{-\pi}^{\pi} \Gamma_n(e^{j\omega})\Gamma_n^\star(e^{j\omega})\Phi_u(\omega)\,\mathrm{d}\omega, \tag{21}$$

(here $\cdot^\star$ denotes 'conjugate transpose') so that the numerical properties of least squares estimation using the model structures (6) and (11) should be closely related to the condition numbers $\kappa(R_\gamma)$ and $\kappa(R_\phi)$.

Now as exposed in the derivation leading to (18), the condition number definition (19) is clearly dependant on the matrix norm used. Most commonly however, the matrix 2-norm is used [31], which for positive definite symmetric R is the largest positive eigenvalue. In this case $\kappa(R)$ is the ratio of largest to smallest eigenvalue of R.

Now here is the key point: For white input u_t, by definition the spectrum $\Phi_u(\omega)$ is a constant (say α) so that by orthonormality $R_\phi = \alpha I$ by construction and hence for this case the use of the orthonormally parameterised form (11) leads to normal equations (17) that are perfectly numerically conditioned since $K(R_\phi) = 1$.

However, an obvious question concerns how the condition numbers of R_γ and R_ϕ compare for the more commonly encountered coloured input case.

An important result in this context is that purely by virtue of the orthonormality in the structure (11), an upper bound on the conditioning of R_ϕ may be guaranteed for any $\Phi_u(\omega)$ by virtue of the fact that [63, 59] ($\lambda(R)$ denotes the set of eigenvalues of the matrix R.)

$$\min_{\omega\in[-\pi,\pi]} \Phi_u(\omega) \le \lambda(R_\phi) \le \max_{\omega\in[-\pi,\pi]} \Phi_u(\omega). \tag{22}$$

No such bounds are available for the matrix R_γ corresponding to the general (non-orthonormal) structure (6). This suggests that the numerical conditioning associated with (11) might be superior to that of (6) across a range of coloured Φ_u, and not just the white Φ_u that the structure (11) is designed to be perfectly conditioned for.

However, in consideration of this prospect, it would seem natural to also suspect that even though $R_\phi = I$ is designed to occur for unit variance white input, that $R_\gamma = I$ might equally well occur for some particular coloured input. If so, then in this scenario the structure (11) would actually be inferior to (6) in numerical conditioning terms. Therefore, in spite of the guarantee (22), it is not

clear when and why the structure (11) should be preferred over the often-times simpler one (6) on numerical conditioning grounds.

In fact, perhaps surprisingly, it turns out that even though the orthonormal model structure (11) is only specifically designed to give perfect conditioning for white input, in fact it seems to provide better numerical conditioning than virtually any other 'simpler' equivalent model structure across a very wide range of coloured inputs as well!

In the remainder of this section, we'll provide some evidence to back these claims up, but in overview the main points to be made will be these:

- Even though $\Phi_u(\omega)$ is an infinite dimensional quantity, and the manifold of symmetric $n \times n$ matrices is only $n(n+1)/2$ dimensional, it is not possible to find a spectral density $\Phi_u(\omega)$ to assign R_ϕ or R_γ arbitrarily. In fact, it is only possible to assign either matrix in *an n dimensional sub-manifold* of the full $n(n+1)/2$ manifold. Furthermore, although the perfectly conditioned $R_\phi = I$ is guaranteed by construction to be in the manifold of possible R_ϕ *it may not be in the manifold of possible* R_γ, or if it is, it may only be achievable for spectral densities $\Phi_u(\omega)$ that are physically unreasonable (they are increasing at the folding frequency indicating that the underlying continuous time signal responsible for the samples u_t was sampled so slowly that aliasing had occurred.

- Approximate formulas for the condition numbers of R_ϕ and R_γ can be generated that clearly show the numerical superiority of the orthonormal form.

In what follows, all the results will be stated without proof. For readers interested in proof, or in an expanded discussion of the material in this section, it is available as [57]

2.0.1 Existence of Spectra

This section addresses the issue of the existence of a particular coloured Φ_u for which the non-orthonormal model structure (6) leads to perfect conditioning ($R_\gamma = I$) and would thus make it a superior choice on numerical grounds than the 'orthonormal' structure (11). This issue is subsumed by that of designing a $\Phi_u(\omega)$ parameterised via real valued co-efficients $\{c_k\}$ as

$$\Phi_u(\omega) = \sum_{k=-\infty}^{\infty} c_k e^{j\omega k} \tag{23}$$

and so as to achieve an arbitrary symmetric, positive definite R_γ. In turn, this question may be formulated as the search for the solution set $\{\cdots, c_{-2}, c_{-1}, c_0, c_1, c_2, \cdots\}$

such that

$$\sum_{k=-\infty}^{\infty} c_k \left(\frac{1}{2\pi j} \oint_{\mathbf{T}} \Lambda_n(z) \Lambda_n^\star(z) z^k \frac{\mathrm{d}z}{z} \right) = R_\gamma$$

which (on recognising that since Φ_u is necessarily real valued then $c_k = c_{-k}$) may be more conveniently expressed as the linear algebra problem

$$\Pi \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ \vdots \end{bmatrix} = \mathrm{vec}\{R_\gamma\} \tag{24}$$

where the vec$\{\cdot\}$ operator is one which turns a matrix into a vector by stacking its columns on top of one another in a left-to-right sequence and the matrix Π, which will be referred to frequently in the sequel, is defined as

$$\Pi \triangleq \frac{1}{2\pi j} \oint_{\mathbf{T}} [\Lambda_n(z) \otimes I_n] \Lambda_n(z)^\star [1, z + z^{-1}, z + z^{-2}, \cdots] \frac{\mathrm{d}z}{z}. \tag{25}$$

Here $\otimes$ denotes the Kronecker tensor product of matrices defined for an $m \times n$ matrix A and an $\ell \times p$ matrix B to provide the $n\ell \times mp$ matrix $A \otimes B$ as

$$A \otimes B \triangleq \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & & & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{bmatrix}.$$

Now it might be supposed that since (24) is an equation involving $n(n+1)/2$ constraints, but with an infinite number of degrees of freedom in the choice $c_0, c_1, \cdots$ then it should be possible to solve for an arbitrary symmetric positive definite R_γ.

Perhaps surprisingly, this turns out not to be the case, the reason being that:

With Π defined as in (25), and for all bases that maintain the same span as in condition (12) then

$$\mathrm{Rank}\, \boldsymbol{\Pi} = \boldsymbol{n}.$$

In fact therefore, the achievable R_γ live only in a sub-manifold of the $n(n+1)/2$ dimensional manifold of $n \times n$ symmetric matrices, and this sub-manifold *may not contain a perfectly conditioned matrix*. Furthermore, as can be seen by (25), this sub-manifold that the possible R_γ lie in will be completely determined by the choice of the functions $\mathcal{G}_k(z)$ in the model structure (6) and hence also in the definition for $\Lambda_n(z)$ in (9).

2.0.2 Approximate Expressions for Condition Number

As mentioned in the introduction, a key feature of the orthonormal parameterisation (11) is that associated with it is a covariance matrix with numerical conditioning guaranteed by the bounds

$$\min_{\omega\in[-\pi,\pi]} \Phi_u(\omega) \leq \lambda(R_\phi) \leq \max_{\omega\in[-\pi,\pi]} \Phi_u(\omega). \tag{26}$$

A natural question to consider is how tight these bounds are. In [59], this was addressed by a strategy of analysis that is asymptotic in n. Specifically, define $X_\phi \triangleq \lim_{n\to\infty} R_\phi$. In this case, X_ϕ is an operator $\ell_2 \to \ell_2$, so that the eigenvalues of the finite dimensional matrix R, generalize to the continuous spectrum $\lambda(R_\infty)$ of the operator X_ϕ defined as [7]

$$\lambda(X_\phi) = \{\lambda \in \mathbf{R} : \lambda I - M_\phi \text{ is not invertible}\}.$$

This spectrum can be characterized as follows (See [59] for a proof - it depends crucially on exploiting orthonormality).

Suppose that $\sum_{k=1}^{\infty}(1-|\xi_k|) = \infty$. Then

$$\boldsymbol{\lambda(X_\phi) = \text{Range}\{\Phi_u(\omega)\}.} \tag{27}$$

This provides evidence, that at least for large n (when the issue of numerical conditioning is most important), that the bounds (26) are in fact tight, and therefore

$$\kappa(R_\phi) \approx \frac{\max_\omega \Phi_u(\omega)}{\min_\omega \Phi_u(\omega)} \tag{28}$$

might be expected to be a reasonable approximation.

Of course, what would also be desirable is a similar approximation for R_γ. This will depend on the nature of the definition of the $\{\mathcal{G}_k(q)\}$. Again consider the simple one (7). Then for this straightforward implementation of a fixed denominator model structure, it is also possible to develop an approximation of the condition number $\kappa(R_\gamma)$ via the following asymptotic result

Consider the choice for the non-orthonormal $\{\mathcal{G}_k(q)\}$ of (7), (9) with associated R_γ defined via (20). Suppose that only a finite number of the poles $\{\xi_\ell\}$ are chosen away from the origin so that

$$D(\omega) \triangleq \lim_{n\to\infty} \prod_{\ell=1}^{n} |e^{j\omega} - \xi_\ell|^2 \tag{29}$$

exists. Define, in a manner analogous to that pertaining to (27), the operator $X_\gamma : \ell_2 \to \ell_2$ as

$$X_\gamma \triangleq \lim_{n\to\infty} R_\gamma.$$

Then

$$\lambda(X_\gamma) = \text{Range}\left\{\frac{\Phi_u(\omega)}{D(\omega)}\right\}.$$

In analogy with the previous approximation, it is tempting to apply this asymptotic result for finite n to derive the approximation

$$\kappa(R_\gamma) \approx \frac{\max_\omega \Phi_u(\omega)/|D_n(e^{j\omega})|^2}{\min_\omega \Phi_u(\omega)/|D_n(e^{j\omega})|^2}. \tag{30}$$

Now, considering that $|D_n(e^{j\omega})|^2 = \prod_{\ell=1}^{n} |e^{j\omega} - \xi_\ell|^2$ can take on both very small values (especially if some of the ξ_ℓ are close to the unit circle) and also very large values (especially if all the $\{\xi_\ell\}$ are chosen in the right half plane so that aliasing is not being modelled), then the maxima and minima of $\Phi_u(\omega)/|D_n(e^{j\omega})|^2$ will be much more widely separated than those of $\Phi_u(\omega)$.

The approximations (28) and (30) therefore indicate that estimation with respect to the orthonormal form (11) could be expected to be much better conditioned than estimation with respect to the model structure (9) with the simple choice (7) and that this conclusion will apply for a very large class of $\Phi_u(\omega)$ - an obvious exception here would be $\Phi_u(\omega) = |D_n(e^{j\omega})|^2$ for which $R_\gamma = I$.

However, this conclusion depends on the accuracy of applying the asymptotically derived approximations (28) and (30) for finite n. In the absence of theoretical analysis, which appears intractable, simulation study can be pursued. Consider n in the range 2-30 with all the $\{\xi_\ell\}$ chosen at $\xi_\ell = 0.5$, and $\Phi_u(\omega) = 0.36/(1.36 - \cos\omega)$. Then the maximum and minimum eigenvalues for R_γ and R_ϕ are shown as solid lines in the left and (respectively) right diagrams in figure (1). The dash-dot lines in these figures are the approximations (28) and (30). Clearly, in this case the approximations are quite accurate, even for

what might be considered small n. (Note that the minimum eigenvalue of R_γ is shown only up until $n = 18$ since it was numerically impossible to calculate it for higher n). Again, this provides evidence that even though model structures

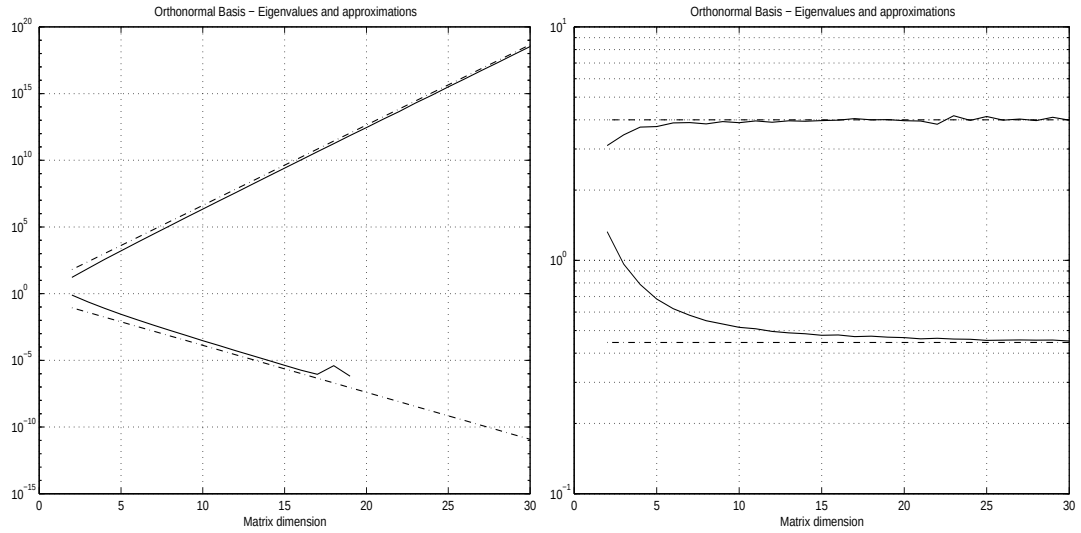

Figure 1: *Solid lines are maximum and minimum eigenvalues of (left figure) R_γ and (right figure) R_ϕ for a range of dimensions n. The dash dot lines are the approximations (28) and (30).*

(11) parameterised in terms of orthonormal $\{\mathcal{B}_k(q)\}$ are only designed to provide superior numerical conditioning properties for white input, they seem to also provide them for a very wide range of coloured inputs as well.

3 Estimation Accuracy - the role of Kernels

Having considered numerical conditioning issues, let us leave them aside and move on to address the question of the role of orthonormality in analysing how the choice of fixed poles in a model structure affects the estimation accuracy.

In order to examine this, it is as well to begin by discussing the general prediction error framework [44, 68, 10] (of which the estimation methods detailed in these notes are a special case) wherein one models the observed input-output behaviour according to the general model structure

$$y_t = G(q, \theta)u_t + H(q, \theta)e_t \tag{31}$$

where e_t is zero mean white noise and (31) is parameterised by a vector $\theta \in \mathbf{R}^n$ such that

$$G(q, \theta) = \sum_{k=1}^{\infty} g_k(\theta)q^{-k}, \quad H(q, \theta) = 1 + \sum_{k=1}^{\infty} h_k(\theta)q^{-k}.$$

() This structure implies the following one step ahead predictor

$$\widehat{y}_t(\theta) = [1 - H^{-1}(q,\theta)]y_t + H^{-1}(q,\theta)G(q,\theta)u_t \tag{32}$$

and associated prediction error

$$\varepsilon_t(\theta) \triangleq y_t - \widehat{y}_t(\theta) \tag{33}$$

so that if the quadratic (least squares) criterion

$$V_N(\theta) \triangleq \frac{1}{2N}\sum_{t=1}^{N}\varepsilon_t^2(\theta)$$

is employed, then based on the N point data observation, a least squares estimate $\widehat{\theta}_N$ of θ may be found as

$$\widehat{\theta}_N^n \triangleq \arg\min_{\theta\in\mathbf{R}^n} V_N(\theta). \tag{34}$$

The theory pertaining to the properties of such a method is very rich. Germane to this report are the properties that with probability one [44, 68, 10]

$$\lim_{N\to\infty}\widehat{\theta}_N^n = \theta_\circ^n$$

where

$$\theta_\circ^n \triangleq \arg\min_{\theta\in\mathbf{R}^n}\lim_{N\to\infty}\mathbf{E}\left\{V_N(\theta)\right\}. \tag{35}$$

With these ideas in hand, note that a very well accepted idea in estimation theory is that the total estimation error consists of two parts, the 'bias error' and the 'variance error'. These are defined according to a decomposition of the total estimation error $G(e^{j\omega}) - G(e^{j\omega},\widehat{\theta}_N)$ as

$$G(e^{j\omega}) - G(e^{j\omega},\widehat{\theta}_N^n) = [G(e^{j\omega}) - G(e^{j\omega},\theta_\circ^n)] + [G(e^{j\omega},\theta_\circ^n) - G(e^{j\omega},\widehat{\theta}_N^n)] \tag{36}$$

with $G_b(e^{j\omega}) \triangleq G(e^{j\omega}) - G(e^{j\omega},\theta_\circ^n)$ resulting from the parsimony of model structure being called the 'bias error' and $G_\nu(e^{j\omega}) \triangleq G(e^{j\omega},\theta_\circ^n) - G(e^{j\omega},\widehat{\theta}_N^n)$ due to measurement noise being called the 'variance error'.

In examining the utility of orthonormal parameterisations in quantifying the size of the total estimation error, these bias and variance error components will be dealt with separately.

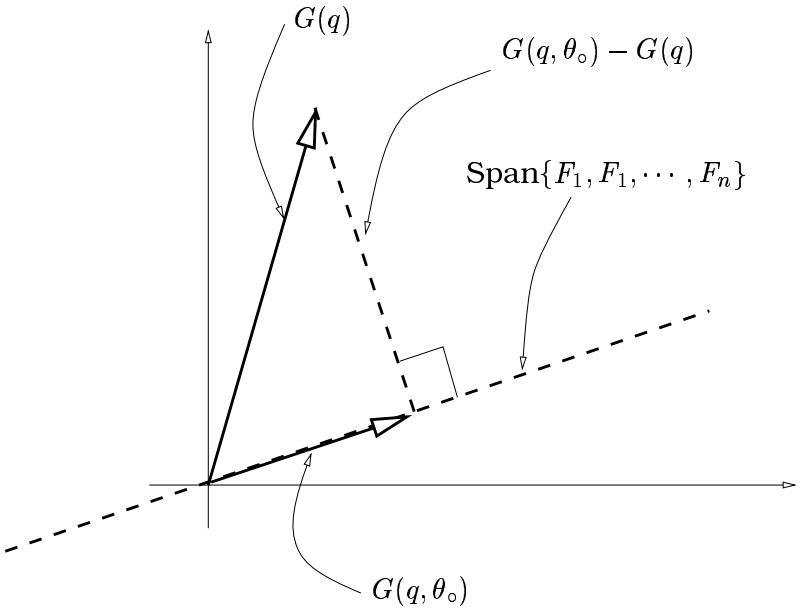

Figure 2: *Geometrical Description of Bias Error*

3.0.3 Bias Error

The limiting estimate $G(e^{j\omega}, \theta_\circ^n)$ which is used to define the bias error $G(e^{j\omega}) - G(e^{j\omega}, \theta_\circ^n)$ can be characterised via (Parseval's Theorem is used to derive this - see [44])

$$G(e^{j\omega}, \theta_\circ^n) = \arg\min_{\theta \in \mathbf{R}^n} \frac{1}{2\pi} \int_{-\pi}^{\pi} |G(e^{j\omega}) - G(e^{j\omega}, \theta)|^2 \Phi_u(\omega)\, \mathrm{d}\omega.$$

But how can this be used to quantify the bias error? Well, if any fixed denominator model structure such as (1) or the orthonormalised form (11) is used, then $G(e^{j\omega}, \theta)$ is parameterised linearly in terms of the elements in θ, so that $G(q, \theta_\circ^n)$ is the element of a subspace closest to $G(q)$, and this means that the error $G(q) - G(q, \theta_\circ^n)$ must be orthogonal to this subspace - see figure 2 for a graphical illustration of this. That is

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \mathcal{B}_k(e^{j\omega}) \overline{[G(e^{j\omega}) - G(e^{j\omega}, \theta_\circ^n)]} \Phi_u(\omega)\, \mathrm{d}\omega = 0, \quad k = 1, 2, \cdots, n \tag{37}$$

This means that the true response $G(e^{j\omega})$ is the hypotenuse of a triangle which has the bias error $G(e^{j\omega}) - G(e^{j\omega}, \theta_\circ^n)$ as one of it's sides. But the hypotenuse is the longest side of a triangle. So averaged over frequency, the bias error is such that true frequency response magnitude is under-estimated

$$\int_{-\pi}^{\pi} |G(e^{j\omega}, \theta_\circ^n)|^2 \Phi_u(\omega)\, \mathrm{d}\omega \leq \int_{-\pi}^{\pi} |G(e^{j\omega})|^2 \Phi_u(\omega)\, \mathrm{d}\omega.$$

What about more precise characterisation of the bias error, such as one which indicates in a frequency dependant fashion what the nature of the bias error is?

To provide this, note that firstly, the frequency response of the limiting estimate $G(e^{j\omega}, \theta_\circ^n)$ is, for ω fixed, *a linear functional* of the whole transfer function $G(z, \theta_\circ^n)$. That is, for some fixed $\omega = \omega_\circ$ there is a linear functional, call it $\mathcal{B}_{\omega_\circ}$, such that

$$\mathcal{B}_{\omega_\circ}[G(q, \theta_\circ^n)] = G(e^{j\omega_\circ}, \theta_\circ^n).$$

It is linear since for any $f, g \in \mathcal{H}_2(\mathbf{E})$ and $\alpha, \beta \in \mathbf{C}$ then

$$\mathcal{B}_{\omega_\circ}[\alpha f + \beta g] = \alpha f(e^{j\omega_\circ}) + \beta g(e^{j\omega_\circ}) = \alpha \mathcal{B}_{\omega_\circ}[f] + \beta \mathcal{B}_{\omega_\circ}[g].$$

Now it is a theorem of functional analysis (the Reisz Representation Theorem), that for every linear functional like this one there is a special function, denote it by $K_n(e^{j\omega}, e^{j\omega_\circ})$, that lives in the same space that $G(z, \theta_\circ^n)$ lives in (in this case $\mathcal{H}_2(\mathbf{E})$) and is such that $\mathcal{B}_{\omega_\circ}[\cdot]$ can be represented via an inner product:

$$\begin{aligned} G(e^{j\omega_\circ}, \theta_\circ^n) = \mathcal{B}_{\omega_\circ}[G(e^{j\omega}, \theta_\circ^n)] &= \langle G(e^{j\omega}, \theta_\circ^n), K_n(e^{j\omega}, e^{j\omega_\circ}) \rangle & (38) \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} G(e^{j\omega}, \theta_\circ^n) \overline{K_n(e^{j\omega}, e^{j\omega_\circ})} \, \mathrm{d}\omega \\ &= \frac{1}{2\pi j} \oint_{\mathbf{T}} G(1/z, \theta_\circ^n) \overline{K_n(1/z, e^{j\omega_\circ})} \, \frac{\mathrm{d}z}{z} & (39) \end{aligned}$$

where in moving to the contour integral formulation, the change of variable $e^{j\omega} \mapsto 1/z$ was used.

As well, if one knows that the true transfer function is stable, then $G(1/z)$ has no poles inside the unit circle $\mathbf{T}$, so for any μ in $\mathbf{D}$ (the interior of $\mathbf{T}$) then by Cauchy's Integral Theorem

$$G(1/\mu) = \frac{1}{2\pi j} \oint_{\mathbf{T}} \frac{1}{z - \mu} G(1/z) \, \mathrm{d}z. \tag{40}$$

Combining (39) and (40) after taking the limit as $\mu \to e^{-j\omega_\circ}$ then gives an expression for the estimation error (bias error):

$$\begin{aligned} G(e^{j\omega_\circ}, \theta_\circ^n) - G(e^{j\omega_\circ}) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} G(e^{j\omega}) \frac{\overline{K_n(e^{j\omega}, e^{j\omega_\circ})} e^{j(\omega - \omega_\circ)}}{1 - e^{j(\omega - \omega_\circ)}} \, \mathrm{d}\omega + & (41) \\ & \frac{1}{2\pi} \int_{-\pi}^{\pi} [G(e^{j\omega}, \theta_\circ^n) - G(e^{j\omega})] \overline{K_n(e^{j\omega}, e^{j\omega_\circ})} \, \mathrm{d}\omega. & (42) \end{aligned}$$

Note that in deriving this, absolutely no assumptions were made on the model structure, or even on the estimation method used.

In fact, the dependence of the estimation error as given by (41), (42) on the choice of model structure is completely encoded in the term $K_n(e^{j\omega}, e^{j\omega_\circ})$ and the choice of estimation method is completely encoded in the second term (42). Because of the obvious importance of this K_n term, both in this context and a wider one of mathematical approximation theory [18], it is given a particular name of 'Reproducing Kernel'.

The critically important issue is that for the case considered in these notes of least squares estimation according to (8) and 'fixed denominator' model structures like (1) or (11) then this reproducing kernel can actually be computed. It is at this stage that the orthonormality issues become crucial, because it is only be exploiting them that this calculation is possible.

Explicitly, regardless of whether a simple model structure like (1) or it's orthonormalised form (11) is actually implemented, it is a fact that for some $\{\theta_k\}$ co-efficients the resultant estimate $\widehat{G}(e^{j\omega})$ (the notation $G(e^{j\omega}, \theta_\circ^n)$, or $G(e^{j\omega}, \widehat{\beta}_N^n)$ is avoided here in order to emphasise that the results do not depend on which of a range of equivalent model structures is used) can be expressed as

$$\widehat{G}(e^{j\omega}) = \sum_{k=1}^{n} \theta_k \mathcal{B}_k(q).$$

Now consider the following formula for the Reproducing Kernel

$$K_n(e^{j\omega}, e^{j\omega_\circ}) = \sum_{m=1}^{n} \mathcal{B}_m(e^{j\omega})\overline{\mathcal{B}_m(e^{j\omega_\circ})} \tag{43}$$

and test to see if this works as follows:

$$\begin{aligned}
\langle \widehat{G}(e^{j\omega}, K_n(e^{j\omega}, e^{j\omega_\circ}))\rangle &= \left\langle \sum_{k=1}^{n} \theta_k \mathcal{B}_k(e^{j\omega}), \sum_{m=1}^{n} \mathcal{B}_m(e^{j\omega})\overline{\mathcal{B}_m(e^{j\omega_\circ})} \right\rangle \\
&= \sum_{k=1}^{n}\sum_{m=1}^{n} \theta_k \mathcal{B}_m(e^{j\omega_\circ})\langle \mathcal{B}_k(e^{j\omega}), \mathcal{B}_m(e^{j\omega})\rangle \\
&= \sum_{k=1}^{n} \theta_k \mathcal{B}_k(e^{j\omega_\circ}) \\
&= \widehat{G}(e^{j\omega_\circ}).
\end{aligned}$$

So the formula (43) does work according to the definition (38). Since it is a fact that $K_n(e^{j\omega}, e^{j\omega_\circ})$ is unique [18], then we have now found a formula for it for the particular case of model structure considered in these notes. Note the crucial role played by the orthonormal bases in deriving $K_n(e^{j\omega}, e^{j\omega_\circ})$ in that it allowed the double sum over k and m to collapse to a single one.

The final step now is to show how this formula for the reproducing kernel can be combined with the expression (41), (42) in order to say something

interesting about the bias error $G(e^{j\omega}) - G(e^{j\omega}, \theta_\circ^n)$. So far, in the derivation of (41), (42) and the formula (43), nothing has been assumed about the estimation method or conditions - only a model structure was supposed. However, in order now to be more explicit about bias error it is finally necessary to be explicit about the estimation method. Note that whatever we specify will only affect the second term (42) in the bias error characterisation since the first term (41) is invariant to the choice of estimation method.

Suppose that we acknowledge that the least squares criterion (8) is to be used, and we also suppose for the moment that the input spectrum $\Phi_u(\omega)$ is white. Then as derived in (37) and shown graphically in figure 2, a consequence of this is that the bias error $G(e^{j\omega}) - G(e^{j\omega}, \theta_\circ^n)$ is then orthogonal to the space spanned by the basis functions. But as shown by the formulation (43), the reproducing kernel $K_n(e^{j\omega}, e^{j\omega_\circ})$ lives in this space. Therefore, the second term (42) must be zero and hence in this case the bias error becomes:

$$G(e^{j\omega_\circ}, \theta_\circ^n) - G(e^{j\omega}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} G(e^{j\omega}) \frac{\overline{K_n(e^{j\omega}, e^{j\omega_\circ})} e^{j(\omega - \omega_\circ)}}{1 - e^{j(\omega - \omega_\circ)}} \, \mathrm{d}\omega. \tag{44}$$

Note also that since the reproducing kernel has been found via (43), then since it is unique it doesn't matter which of the orthonormal basis formulations considered in these notes is used to actually calculate it, since the result is invariant to this choice.

This leaves open the possibility of using whichever is simplest for the application at hand. In the remainder of this section, the applications are analytical ones in which case it is essential to have a closed form formula for $K_n(e^{j\omega}, e^{j\omega_\circ})$. Therefore, it is most expedient to use basis function formulations that themselves are in closed form like (14) because this allows (after some arithmetic - see [59] for details) the kernel $K_n(e^{j\omega}, e^{j\omega_\circ})$ to be written as

$$K_n(e^{j\omega}, e^{j\omega_\circ}) = \frac{1 - \prod_{k=1}^{n} \frac{(e^{j\omega_\circ} - \xi_k)(1 - e^{j\omega}\overline{\xi}_k)}{(1 - e^{j\omega_\circ}\overline{\xi}_k)(e^{j\omega} - \xi_k)}}{1 - e^{j(\omega - \omega_\circ)}}$$

Combining this formulation with the expression (44) then allows the derivation of the following bias error expression:

Suppose $G(z)$ has partial fraction expansion

$$G(z) = \sum_{i=0}^{m-1} \frac{\alpha_i}{z - p_i^0}, \qquad |p_i^0| < 1$$

Suppose that u_t is white ($\Phi_u(\omega) = \text{constant}$). Then

$$|G(e^{j\omega}) - G(e^{j\omega}, \theta_\circ^n)| \leq \sum_{i=0}^{m-1} \left| \frac{\alpha_i}{e^{j\omega} - p_i^0} \right| \prod_{k=0}^{n-1} \left| \frac{p_i^0 - \xi_k}{1 - \overline{\xi}_k p_i^0} \right|. \tag{45}$$

The proof is a simple one that involves re-expressing (44) as a contour integral via the change of variable $e^{j\omega} \mapsto z$ and then using Cauchy's Residue Theorem, and then the Triangle inequality.

These details are not particularly important, but what *is* important is the vital role played by re-parameterising the estimation problem into one in which a model structure is equivalent, but orthonormal. Via this strategy, the utility of orthonormal bases is exposed as being one that greatly facilitates analysis, whether or not one chooses to actually implement an estimation algorithm in terms of the orthonormal basis.

Note however that this only addresses the bias error for white input. For the case of coloured input, things become more complicated, mainly because it seems impossible to cleanly handle the second term (42) in this case. One way of circumventing these problems at least partially is to present a numerical means for calculating the bias error:

The asymptotic frequency response estimate $G(e^{j\omega}, \theta_\circ^n)$ at $\omega = \omega_\circ$ is related to the true frequency response estimate $G(e^{j\omega})$ via

$$G(e^{j\omega_\circ}, \theta_\circ^n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} G(e^{j\omega})\overline{K_n(e^{j\omega}, e^{j\omega_\circ})}\Phi_u(\omega)\,\mathrm{d}\omega \tag{46}$$

where the reproducing kernel $K_n(e^{j\omega}, e^{j\omega_\circ})$ may be calculated as

$$K_n(e^{j\omega}, e^{j\omega_\circ}) = \sum_{k=1}^{n} \alpha_k(\omega_\circ)\mathcal{B}_k(e^{j\omega}) \tag{47}$$

with the $\{\alpha_k(\omega_\circ)\}$ satisfying the n equations

$$\sum_{k=1}^{n} \alpha_k(\omega_\circ)\frac{1}{2\pi}\int_{-\pi}^{\pi} \mathcal{B}_k(e^{j\omega})\overline{\mathcal{B}_\ell(e^{j\omega})}\Phi_u(\omega)\,\mathrm{d}\omega = \overline{\mathcal{B}_\ell(e^{j\omega_\circ})} \quad \ell = 1, 2, \cdots, n. \tag{48}$$

Note that the previously considered case is a special case of this theorem, since it will provide the formula (43) for $K_n(e^{j\omega}, e^{j\omega_\circ})$ when the input spectrum Φ_u happens to be white. This result also makes explicit the relationship between $G(e^{j\omega})$, the choice of the basis functions $\{\mathcal{B}_k(z)\}$, the input spectrum $\Phi_u(\omega)$ and the estimate $G(e^{j\omega}, \theta_\circ^n)$ of $G(e^{j\omega})$. It shows that the estimated response at some frequency $\omega_\circ$ depends on the integral of the whole true frequency response weighted by a kernel $K_n(e^{j\omega}, e^{j\omega_\circ})$. Obviously, we would like the kernel to be as much like a dirac delta function $\delta(\omega - \omega_\circ)$ as possible. We can check if this is so for a particular choice of $\{\mathcal{B}_k(z)\}$ by using the formulae (47),(48).

For example, suppose we envisage an input spectral density as shown in the upper right corner of figure 3, and we are trying to decide between using one of two possible model structures corresponding to the following choice of basis functions:

$$\begin{aligned} &\text{FIR:} && \mathcal{B}_n(z) = \frac{1}{z^n} && ; n = 1, 2, \cdots, 10 \\ &\text{Laguerre:} && \mathcal{B}_n(z) = L_n(z) = \frac{\sqrt{1-\xi^2}}{z-\xi}\left(\frac{1-\xi z}{z-\xi}\right)^{n-1} && ; n = 1, 2, \cdots, 10 \quad \xi = 0.9 \end{aligned} \tag{49}$$

We can then calculate the reproducing kernels $K_n(e^{j\omega}, e^{j\omega_\circ})$ for both these choices via (47),(48) and plot them to see if we can expect good estimates at frequencies of interest to us. This is done for $\omega_\circ$ set to the normalised frequencies of 0.08 rad/s and 0.76 rad/s in the lower left and right of figure 3. The Laguerre

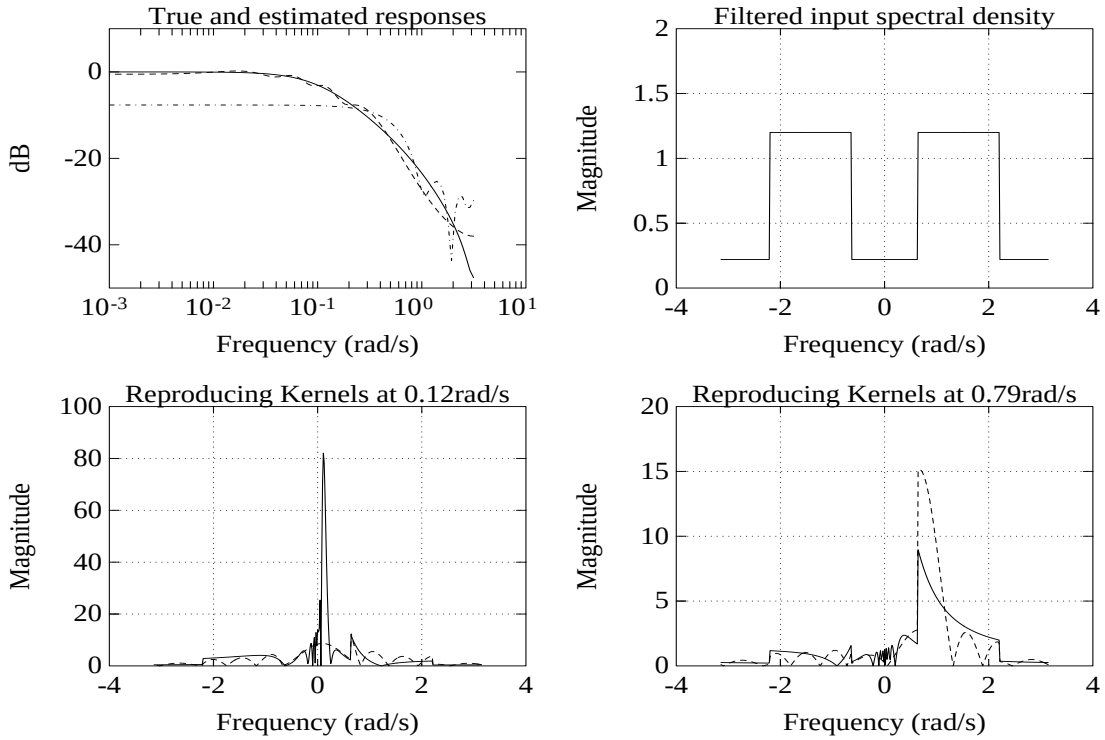

Figure 3: *Comparison of FIR and Laguerre basis function models for frequency response estimation by examination of their corresponding reproducing kernels. In top left plot Laguerre estimate is dashed line, FIR estimate is dash-dot line and true system is the solid line. In bottom plots, FIR kernel is shown as a dashed line and Laguerre kernel as solid line.*

kernel is shown as a solid line and the FIR kernel is shown as a dashed line. As can be seen, at 0.08 rad/s the Laguerre kernel is a good approximation to a delta function at that frequency, so we should expect accurate frequency response estimates from a Laguerre based model at that frequency. In contrast, the FIR kernel is a very poor approximation to a delta function[1], so would not be a good basis function choice if accurate low frequency estimates are required. At the higher frequency of 0.76 rad/s the reproducing kernels for the two basis function choices are shown in the lower right of figure 3.

Once the kernel $K_n(e^{j\omega}, e^{j\omega_\circ})$ has been calculated for a range of $\omega_\circ$ of interest, we can use (46) to calculate what the asymptotic estimate $G(e^{j\omega}, \theta^n_\circ)$ will be for a hypothesised $G(e^{j\omega})$. For example, with a hypothetical $G(z)$ of

$$G(z) = \frac{0.4}{z - 0.6} \tag{50}$$

[1]The FIR kernel, shown as a dashed line, is hard to see in the lower left of figure 3 since it is buried in the low amplitude side-lobes of the Laguerre kernel.

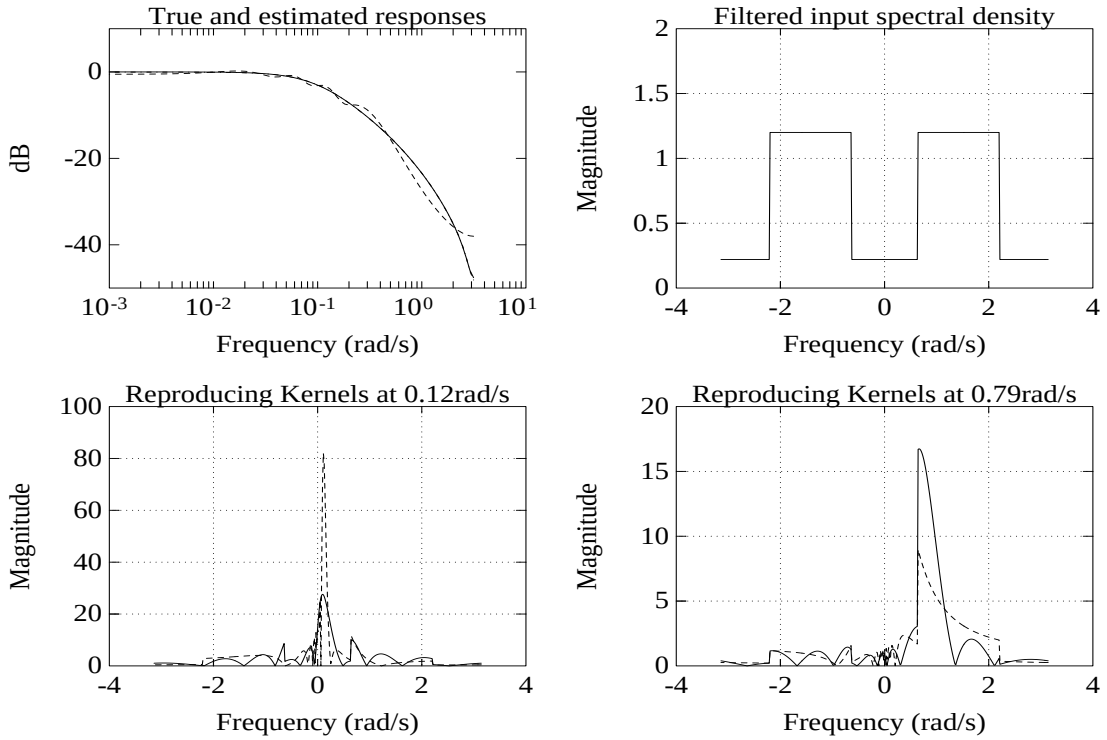

Figure 4: *Comparison of Legendre and Laguerre basis function models for frequency response estimation by examination of their corresponding reproducing kernels. In top left plot Legendre estimate sits directly on top of true system which appears as a solid line, Laguerre estimate is the dashed line. In bottom plots, Legendre kernel is shown as solid line and Laguerre kernel as dashed line.*

whose frequency response is shown as a solid line in the upper left of figure 3, we can predict via (46) what the asymptotic estimates will be for the FIR and Laguerre model structures. The Laguerre is shown as a dashed line and the FIR as a dash-dot line in the upper left of figure 3. As can be seen, the Laguerre model is more accurate at low frequencies and we expected this because its reproducing kernel $K_n(e^{j\omega}, e^{j\omega_\circ})$ was very much like a dirac delta at low frequencies. Note that these plots were not made from an estimation experiment, but were made by numerically evaluating (46)–(48).

Of course, the kernels for other basis function choices can also be calculated and integrated against hypothetical $G(e^{j\omega})$ in an effort to a-priori choose the most suitable basis for frequency response estimation. For example, in the upper left of figure 3 we note that against a hypothetical $G(e^{j\omega})$, a Laguerre basis estimate performs much better than an FIR estimate of the same order. Nevertheless, the Laguerre model is somewhat deficient at high frequencies. Motivated by this we examine the use of the so-called Legendre basis, which is

(14) with the choice of pole location as

$$\xi_k = \frac{2 - \xi(2k+1)}{2 + \xi(2k+1)}$$

for some fixed real ξ. Due to the progression in pole location in this basis progression in pole position might be expected to perform better than a Laguerre basis at high frequencies. This indeed is the case as reference to figure 4 shows. Note that in the upper left plot of figure 4 the Legendre basis function estimate cannot be discriminated from the hypothetical $G(e^{j\omega})$ since it sits directly on top of it.

3.1 Variance Error

Remember that back in equation(36) the total estimation error $G(e^{j\omega}) - G(e^{j\omega}, \widehat{\theta}_N)$ was decomposed into a 'bias error' and a 'variance error' part as

$$G(e^{j\omega}) - G(e^{j\omega}, \widehat{\theta}_N^n) = [G(e^{j\omega}) - G(e^{j\omega}, \theta_\circ^n)] + [G(e^{j\omega}, \theta_\circ^n) - G(e^{j\omega}, \widehat{\theta}_N^n)].$$

The bias error term $G(e^{j\omega}) - G(e^{j\omega}, \theta_\circ^n)$ has just been studied in the previous section. This section will consider the variance error term defined as being $G(e^{j\omega}, \theta_\circ^n) - G(e^{j\omega}, \widehat{\theta}_N^n)$. It represents the error induced by the measurement noise process ν_t.

In order to try to quantify this error, an appropriate place to start would seems to be to consider what results might be available in the 'classical' literature that pertains to general prediction error methods. In this context, a relevant result is that [46, 44] the random variations of the parameter space estimates defined by the general prediction error framework (31)–(34) follow a particular Gaussian distribution for large data length N

$$\sqrt{N}(\widehat{\theta}_N^n - \theta_\circ^n) \xrightarrow{\mathcal{D}} \mathcal{N}(0, P_n), \quad \text{as } N \to \infty \tag{51}$$

where the $\mathcal{D}$ notation means that the distribution of the functions on the left of the arrow converges to the distribution specified on the right. In the specification of this latter distribution

$$P_n \triangleq R_n^{-1} Q_n R_n^{-1}. \tag{52}$$

where with the definition of the prediction error gradient $\psi_t(\theta)$ as

$$\psi_t(\theta) = \frac{\mathrm{d}\widehat{y}_t(\theta)}{\mathrm{d}\theta} = \Gamma_n(q) u_t$$

then for open loop data collection and using Parseval's Theorem

$$\begin{aligned} R_n \triangleq \lim_{N\to\infty} \mathbf{E}\left\{ \left.\frac{\mathrm{d}^2 V_N(\theta)}{\mathrm{d}\theta\mathrm{d}\theta^T}\right|_{\theta=\theta_\circ^n} \right\} &= \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} \mathbf{E}\left\{ \psi_t(\theta_\circ^n) \psi_t^T(\theta_\circ^n) \right\} \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \Gamma_n(e^{j\omega}) \Gamma_n(e^{j\omega}) \Phi_u(\omega)\, \mathrm{d}\omega. \end{aligned} \tag{53}$$

As well, under the simplifying assumption of no undermodelling existing (in the sequel, large model order n is considered)

$$\begin{aligned} Q_n &\triangleq \lim_{N\to\infty} N\mathbf{E}\left\{\frac{\mathrm{d}V_N(\theta_\circ^n)}{\mathrm{d}\theta}\left(\frac{\mathrm{d}V_N(\theta_\circ^n)}{\mathrm{d}\theta}\right)^T\right\} \\ &= \lim_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N}\sum_{\ell=1}^{N} E\left\{\psi_t(\theta_\circ^n)\psi^T(\ell,\theta_\circ^n)v_t v(\ell)\right\} \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\Gamma_n(e^{j\omega})\Gamma_n(e^{j\omega})\Phi_u(\omega)\Phi_\nu(\omega)\,\mathrm{d}\omega. \end{aligned} \tag{54}$$

where the passage to the last line is not trivial (see [58] for details), but again depends crucially on Parseval's Theorem.

Therefore, although (51) gives us quite precise information, at least in parameter space, for the nature of the variance error, the expression for P_n is so complicated that it would seem difficult for us to extract any useful design insight from it.

In recognition of this problem, a quite famous result due to Ljung and coworkers [48, 45, 47, 82] is that in fact (51) can be used as a foundation for deriving a very simple and useful approximation for the variance error.

For example, in [48, 45] it is shown for the special FIR case of $\xi_k = 0$ for all k then for 'large' data length N and model order n

$$\mathbf{E}\left\{|G(e^{j\omega},\theta_\circ^n) - G(e^{j\omega},\widehat{\theta}_N^n)|^2\right\} \approx \frac{n}{N}\frac{\Phi_\nu(\omega)}{\Phi_u(\omega)} \tag{55}$$

where $\Phi_\nu(\omega)$ is the spectral density of the measurement noise ν_t. The purpose of this section is to point out that this is only a good approximation if all (or at least less than 20%) of the poles $\{\xi_k\}$ are fixed at the origin, and that if not a more accurate approximation that includes (55) as a special case is formulated in terms of orthonormal bases as

$$\mathbf{E}\left\{|G(e^{j\omega},\theta_\circ^n) - G(e^{j\omega},\widehat{\theta}_N^n)|^2\right\} \approx \frac{1}{N}\frac{\Phi_\nu(\omega)}{\Phi_u(\omega)}\sum_{k=1}^{n}|\mathcal{B}_k(e^{j\omega})|^2. \tag{56}$$

where the bases $\{\mathcal{B}_k(q)\}$ have poles at the $\{\xi_k\}$ specifying the fixed denominator $D_n(q)$.

To illustrate this perhaps unexpected phenomenon, consider a simple simulation example where an $n = 12$'th order FIR model of the true system (the zero order hold equivalence is calculated using a sampling period of 1 second)

$$G(q) = \mathcal{ZOH}\left\{\frac{1}{(s+1)(2s+1)}\right\} = \frac{0.1548q + 0.0939}{(q-0.6065)(q-0.3679)}$$

is obtained by observing 10000 samples of the true systems its input-output response when the former is a stationary Gaussian process with spectral density $\Phi_u(\omega) = 0.25/(1.25 - \cos\omega)$ and the latter is corrupted by zero mean Gaussian white noise of variance $\sigma^2 = 0.001$. In this case, since both n and N can reasonably be considered 'large', then the approximation (55) could be expected to hold. This can be checked by Monte–Carlo simulation over say, 500 input and noise realisations so as to estimate the variance $\mathbf{E}\left\{|G(e^{j\omega}, \theta_\circ^n) - G(e^{j\omega}, \widehat{\theta}_N^n)|^2\right\}$ by its sample average, which can then be compared to the approximation (55). The results for just such an experiment are shown in Figure 5 with the agreement between (dashed line) the expression (55) and the sample average (solid line) being excellent. Note that in this simulation (and in all the rest to follow in this section), the bias error is negligible, and hence the variance error represents the total error.

This close agreement even for reasonably small N and n is encouraging, and so the approximation (55) is clearly relevant to our goal here of quantifying the bias error involved with various fixed denominator model structures. However the approximation (55) is derived in [48] only for FIR model structures. Nevertheless, since FIR structures are so close to the fixed pole generalisations of them considered in these notes, it would seem reasonable to expect that the original results (55) could be pressed into service without too much trouble.

To pursue this point, consider the continuation of the simulation example in which we suppose that prior knowledge of the poles of $G(q)$ exists, so that, as illustrated in the previous section, in the interests of decreasing the bias error it makes sense to try to incorporate this prior knowledge in the estimation process by fixing some poles in the model near where it is believed the true poles lie.

For example, suppose in the previous simulation it is believed that the true pole is near $z = 0.75$, so that guesses of, say $z = 0.7, 0.72, 0.78, 0.8$ are to be incorporated into the model structure. This can be implemented by simply pre-filtering the input by $F(q) = q^4/(q - 0.7)(q - 0.72)(q - 0.78)(q - 0.8)$ before an FIR 'numerator' model $\widehat{G}(q)$ is estimated, and then the complete system $\widehat{G}'(q) = \widehat{G}(q)F(q)$ may be taken as the fixed-pole estimate of the underlying dynamics.

Since the model order is still large ($n = 12$) then (55) should still provide a quantification of the variability of this new estimate by labeling the FIR estimate as $\widehat{G}$ and reasoning

$$\begin{aligned}
\mathrm{Var}\{\widehat{G}'(e^{j\omega})\} &= |F(e^{j\omega})|^2 \mathbf{E}\left\{|G(e^{j\omega}, \theta_\circ^n) - G(e^{j\omega}, \widehat{\theta}_N^n)|^2\right\} && (57)\\
&\approx \frac{n}{N}|F(e^{j\omega})|^2 \frac{\Phi_\nu(\omega)}{|F(e^{j\omega})|^2\Phi_u(\omega)} && (58)\\
&= \frac{n}{N}\frac{\Phi_\nu(\omega)}{\Phi_u(\omega)} && (59)
\end{aligned}$$

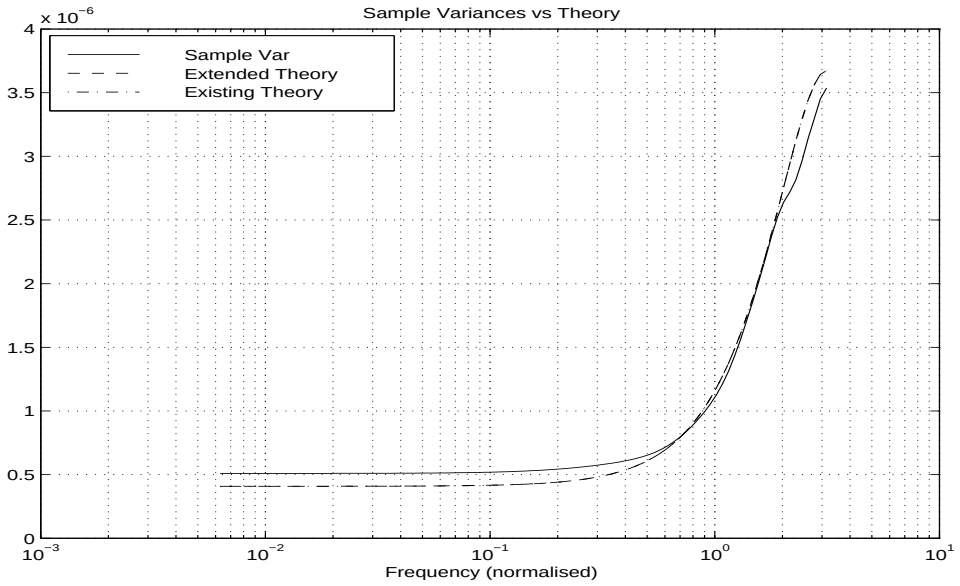

Figure 5: *FIR with all poles at origin. This is a comparison of Monte–Carlo estimate of sample variability (solid line) with (dashed-dot line) the approximate expression (55). Note that this last line obscures a dashed line which is the new approximation (56). The obfuscation occurs because for the case of all poles at the origin the pre-existing approximation (55) and the new one (56) are identical.*

which is unchanged from the normal FIR case. This unchanging nature is also reasonable, since the FIR case can be considered as already incorporating prior knowledge of system poles; namely poles near the origin.

Interestingly, when the expression (59) is compared to Monte–Carlo calculated sample variability as it is in Figure 6, then the agreement between the true variability (solid line) and approximation (59) (dash-dot line) is seen to be not nearly so good as is figure 5. Nevertheless, the expression (59) still provides useful information on the qualitative 'high-pass' nature of how the true variability changes with frequency. The dashed line near the solid one in figure 6 will be commented on in a moment.

Now suppose even more guesses of system poles are made, say at the locations $z = \{0.7, 0.72, 0.78, 0.8, 0.75, 0.85, 0.82, 0.79\}$, with the sample variability again being compared to (59) in figure 7. In this case there is virtually no agreement (even qualitative) between true and predicted variability. Clearly then, the well known approximation (55) can be quite misleading in situations where it would be expected to be reliable - namely when using the fixed denominator model structures considered in these notes. This indicates that the following issues have to be dealt with:

1. Why does the approximation (55) become progressively worse in describing

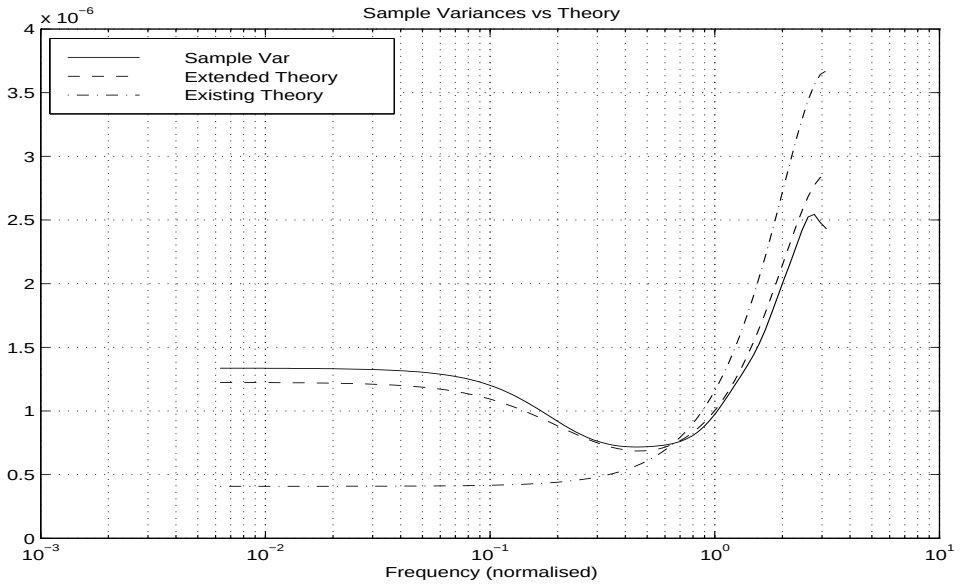

Figure 6: *FIR with 4 poles away from origin. This is a comparison of Monte–Carlo estimate of sample variability (solid line) with (dash-dot line) the approximate expression (55) and (dashed line) the new improved approximation (56) derived in these notes.*

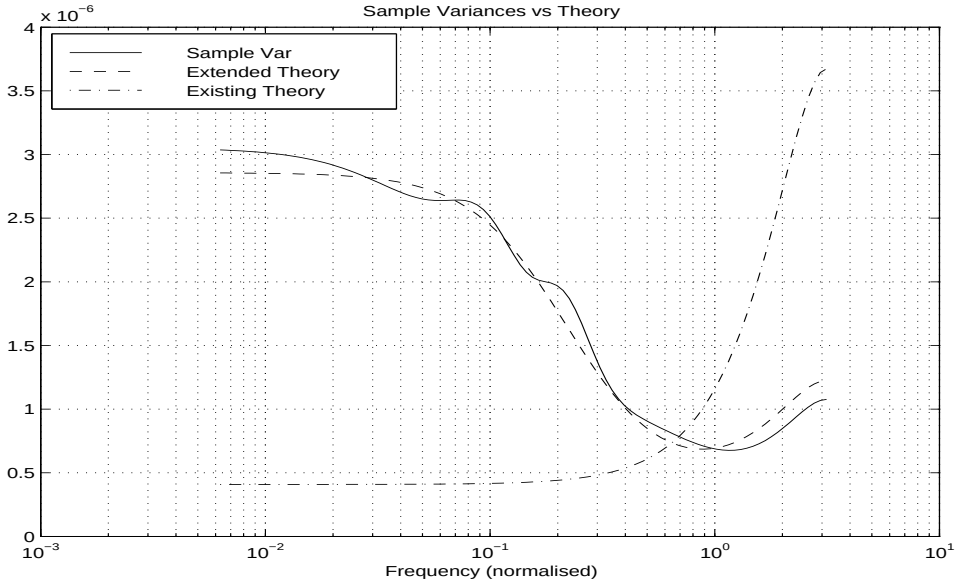

Figure 7: *FIR with 8 poles away from origin.*

the variance error of fixed denominator model structures as more poles $\{\xi_k\}$ are moved away from the origin?

2. How is the new 'improved accuracy' expression (56) obtained?

3. Why do the orthonormal bases $\{\mathcal{B}_k(z)\}$ appear (56) even for cases such as (1) where they are not included in the model structure?

We will address these issues in turn.

3.1.1 Failure of FIR derived Variance Expression

In order to understand why (55) is not a good approximation for arbitrary fixed pole location, it is necessary to understand the rudiments of how (55) is derived. There are two important principles underlying this derivation. Namely, use of the asymptotic nature of Toeplitz matrices and the employment of the principles of Fourier series convergence.

Beginning with Toeplitz matrices, note that any positive function $f : [-\pi, \pi] \to (0, \infty)$ defines an $n \times n$ symmetric Toeplitz matrix denoted as $T_n(f)$ by using $\Gamma_n(z)$ defined in (15) with the special choice of $\xi_k = 0$ for all k so that $\Gamma_n(z)^T = [z^{-1}, \cdots, z^{-n}]$ to provide

$$T_n(f) \triangleq \frac{1}{2\pi} \int_{-\pi}^{\pi} \Gamma_n(e^{j\omega}) \Gamma_n^\star(e^{j\omega}) f(\omega) \, \mathrm{d}\omega \tag{60}$$

As well, recalling the definition of R_n made in (53), not that via it's Parseval derived frequency domain characterisation it may be written as (compare this with (20) and (21) way back in section 2)

$$R_n = T_n(\Phi_u / |D_n|^2). \tag{61}$$

It is also possible, for the case of $H(q) = 1$ to establish via (54) that provided the measurement noise ν_t is white and of variance σ^2 then $Q_n = \sigma^2 R_n$ so that $P_n = \sigma^2 R_n^{-1} = \sigma^2 T_n^{-1}(\Phi_u / |D_n|^2)$.

Furthermore, a very well known result on the asymptotic in n behaviour of Toeplitz matrices [32, 76] is that for any continuous positive f and g

$$T_n(f) T_n(g) \approx T_n(fg)$$

where the precise meaning of the $\approx$ operator is that equality occurs in Hilbert–Schmidt weak matrix norm $|\cdot|$ (defined by $|A_n|^2 = n^{-1}\mathrm{Trace}\{A_n^T A_n\}$) as $n \to \infty$. Therefore, since by orthonormality $T_n(1) = I$ and

$$T_n^{-1}(f) = T_n(1/f) + T_n^{-1}(f)[T_n(1) - T_n(f) T_n(1/f)]$$

then $T_n^{-1}(f) \approx T_n(1/f)$ and hence

$$P_n = \sigma^2 T_n^{-1}(\Phi_u/|D_n|^2) \approx \sigma^2 T_n\left(\frac{|D_n|^2}{\Phi_u}\right).$$

Now, note that from the parameter space distributional result (51)

$$N\mathbf{E}\left\{(\widehat{\theta}_N^n - \theta_\circ^n)(\widehat{\theta}_N^n - \theta_\circ^n)^T\right\} \approx P_n \approx \sigma^2 R_n^{-1} \approx \sigma^2 T_n\left(\frac{|D_n|^2}{\Phi_u}\right) \tag{62}$$

so that since, in the case of using the model structure (1) with fixed denominator $D_n(z)$ which was considered in the previous simulation examples

$$G(e^{j\omega},\theta) = \frac{1}{D_n(e^{j\omega})}\sum_{k=1}^{n}\theta_k e^{-j\omega k}$$

then in fact

$$\begin{aligned}\frac{N}{n}\mathbf{E}\left\{|G(e^{j\omega},\theta_\circ^n) - G(e^{j\omega},\widehat{\theta}_N^n)|^2\right\} &\approx \frac{\sigma^2}{|D_n(e^{j\omega})|^2}\sum_{\ell=1}^{n}\sum_{m=1}^{n} e^{j\omega\ell}e^{-j\omega m}[T_n(|D_n|^2/\Phi_u)]_{m,\ell} \\ &= \frac{\sigma^2}{|D_n(e^{j\omega})|^2}\sum_{k=-n}^{n}\left(1-\frac{|k|}{n}\right)c_k e^{j\omega k} \end{aligned} \tag{63}$$

where the $\{c_k\}$ are the Fourier co-efficients of $|D_n|^2/\Phi_u$ defined by

$$c_k \triangleq \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{|D_n(e^{j\omega})|^2}{\Phi_u(\omega)}e^{-j\omega k}\,\mathrm{d}\omega$$

so that the left hand side of (63) is in fact a Cesàro mean (triangularly windowed) Fourier reconstruction of $|D_n|^2/\Phi_u$ which is known [83], provided $|D_n|^2/\Phi_u$ is continuous, to converge uniformly to $|D_n|^2/\Phi_u$ on its domain and with increasing n. Therefore, it should approximately hold that

$$\sum_{k=-n}^{n}\left(1-\frac{|k|}{n}\right)c_k e^{j\omega k} \approx \frac{|D_n(e^{j\omega})|^2}{\Phi_u(\omega)} \tag{64}$$

so that combining this approximation with the ones leading to (63) provides

$$\frac{N}{n}\mathbf{E}\left\{|G(e^{j\omega},\theta_\circ^n) - G(e^{j\omega},\widehat{\theta}_N^n)|^2\right\} \approx \frac{\sigma^2}{|D_n(e^{j\omega})|^2}\frac{|D_n(e^{j\omega})|^2}{\Phi_u(\omega)}$$

and hence

$$\mathbf{E}\left\{|G(e^{j\omega},\theta_\circ^n) - G(e^{j\omega},\widehat{\theta}_N^n)|^2\right\} \approx \frac{n}{N}\frac{\sigma^2}{\Phi_u(\omega)} \tag{65}$$

which is (55) for $\Phi_\nu(\omega) = \sigma_\nu^2$. Now, in the case where all the poles are fixed at the origin and hence $|D_n(e^{j\omega})|^2 = 1$, then as illustrated in figure 5 these approximating steps work rather well and give an informative indication of the true variance error.

The point to be addressed now is why, as illustrated in figures 6 and 7, the same approximating arguments work so poorly when poles are fixed away from the origin.

The reason is that the analysis leading to approximation in (64) depends on the Fourier series (which (64) is) to have approximately converged. The number of terms n for which this can be expected to have occurred is well known to depend on the smoothness of the function being approximated [41]. But this function is $\Phi_u(\omega)/|D_n(e^{j\omega})|^2$, which clearly becomes less smooth as more poles are moved away from the origin. So the reason why (65) which is (55) fails in the general fixed denominator case is that the underlying Fourier series has not approximately converged.

For the special FIR case of $|D_n(e^{j\omega})| = 1$, the smoothness is fixed as the smoothness of $\Phi_u(\omega)$ so that the approximation can be expected to monotonically improve with increasing n. However, the more poles that are chosen that are away from the origin, and hence the more dependant on n that $|D_n(e^{j\omega})|$ is, the less one should rely on (55) applying for finite n since the less likely it is that the underlying Fourier series has come close to convergence. This is precisely the behaviour demonstrated in figures 5,6 and 7.

3.1.2 Derivation of Improved Variance Expression

Initially, this analysis of the utility of (55) for certain model structures may seem pessimistic since it indicates that much higher model orders n will be necessary before (55) can be used. A main contribution of the use of general rational orthonormal bases is that in fact this pessimism is unfounded, since as pre-cursed by the dashed lines of figures 6,7 it is possible to derive the improved approximation (56) that does not require increased model complexity in order to be accurate.

The key idea behind deriving (56) is to recognise that the asymptotic frequency domain properties of $G(e^{j\omega}, \widehat{\theta}_N^n)$ are invariant to re-parameterisations of the model structure. Therefore, regardless of whether a fixed denominator structure like (1) or (11) is implemented, for the purposes of analytical tractability it is assumed that in fact the orthonormal one (11) is employed.

The machinery behind using this equivalent-parameterisation idea is essentially the same as that just presented, save that the ideas of the asymptotic properties of Toeplitz matrices and the convergence of Fourier series have to be generalised to the case of generalised Toeplitz matrices and general Fourier series with respect to the general rational orthonormal basis $\{\mathcal{B}_k(z)\}$.

These ideas are developed in detail in [63, 59], but the essential points are that firstly, in [59] the classical Toeplitz matrix formulation (60) in generalised to

$$M_n(f) \triangleq \frac{1}{2\pi}\int_{-\pi}^{\pi} \Gamma_n(\omega)\Gamma_n^\star(\omega) f(\omega)\,\mathrm{d}\omega \tag{66}$$

which although formally identical to (66), is functionally quite different in that the underlying orthonormal basis is not fixed at the trigonometric one, but a generalisation obtained by $\Gamma_n(z)$ defined in (15) being allowed to be any rational orthonormal basis of the class formulated in (14):

$$\Gamma_n(z) \triangleq [\mathcal{B}_1(z), \mathcal{B}_1(z), \cdots, \mathcal{B}_n(z)]^T. \tag{67}$$

Matrices defined by (66), (67) are called 'generalised Toeplitz' matrices, with the epithet deriving from the fact that if all the poles are chosen at the origin then $M_n(f) = T_n(f)$ is a bona-fide Toeplitz matrix, but otherwise it is not.

As was proved in [59], a fundamental point is that these generalised Toeplitz matrices also have the pleasant property that for large dimension n and any continuous positive definite f and g:

$$M_n(f)M_n(g) \approx M_n(fg) \tag{68}$$

$$M_n^{-1}(f) \approx M_n(1/f). \tag{69}$$

Additionally, according to the definition of R_n in (53)

$$R_n = M_n(\Phi_u).$$

Similarly, via (54) the matrix Q_n is expressible as

$$Q_n = M_n(\Phi_u\Phi_\nu)$$

so that since the matrix P_n governing the parameter space covariance according to (62) is given by $P_n = R_n^{-1}Q_nR_n^{-1}$ then using the approximations (68) and (69)

$$\begin{aligned} N\mathbf{E}\left\{(\widehat{\theta}_N^n - \theta_\circ^n)(\widehat{\theta}_N^n - \theta_\circ^n)^T\right\} &\approx P_n \\ &\approx M_n^{-1}(\Phi_u)M_n(\Phi_u\Phi_n)M_n^{-1}(\Phi_u) \\ &\approx M_n(\Phi_\nu/\Phi_u). \end{aligned}$$

Proceeding then as in the previous section, then since

$$G(e^{j\omega},\theta) = \Gamma_n^T(e^{j\omega})\theta$$

it should hold that

$$N\mathbf{E}\left\{|G(e^{j\omega},\theta_\circ^n) - G(e^{j\omega},\widehat{\theta}_N^n)|^2\right\} \approx \Gamma_n^\star(e^{j\omega})M_n(\Phi_\nu/\Phi_u)\Gamma_n(e^{j\omega}). \tag{70}$$

In order to further simplify this to some useful form, it is pivotal that a generalised analog of the Fourier convergence of (63) exists. Such a result is provided in [59] where it is shown that provided $\sum(1-|\xi_k|)=\infty$ then for any possibly complex valued but continuous function f

$$\lim_{n\to\infty}\frac{1}{K_n(\omega,\omega)}\Gamma_n^\star(\omega)M_n(f)\Gamma_n(\lambda)=\begin{cases} f(\omega) & ;\omega=\lambda, \\ 0 & ;\omega\neq\lambda \end{cases} \tag{71}$$

Where, remembering that from the previous section on bias error, $K_n(\omega,\sigma)$ formulated in (43) is the so-called 'reproducing kernel' of the space spanned by $\{\mathcal{B}_1,\cdots,\mathcal{B}_n\}$. As will be seen in a moment, as well as playing a key role in quantifying bias error, this reproducing kernel also plays a key role in quantifying variance error.

Note also that by the formulation (43), (14) with $\xi_k=0$ then $K_n(\omega,\omega)=n$ so that the left hand side of (71) becomes the Cesàro mean (63). Because of this, the result (71) can be seen to expand Fourier convergence analysis to the general orthonormal basis (14) which contains the classical trigonometric basis of (63) as a special case of $\xi_k=0$.

Returning to the question of variance error, combining (71) with (70) leads to

$$\begin{aligned}\lim_{n\to\infty}\frac{N}{K_n(\omega,\omega)}\mathbf{E}\left\{|G(e^{j\omega},\theta_\circ^n)-G(e^{j\omega},\widehat{\theta}_N^n)|^2\right\} &= \lim_{n\to\infty}\frac{1}{K_n(\omega,\omega)}\Gamma_n^\star(e^{j\omega})M_n(\Phi_\nu/\Phi_u)\Gamma_n(e^{j\omega}) \\ &= \frac{\Phi_\nu(\omega)}{\Phi_u(\omega)}\end{aligned}$$

so that assuming that convergence has approximately converged for finite n provides the approximation presented in (56) of

$$\mathbf{E}\left\{|\boldsymbol{G}(\boldsymbol{e}^{\boldsymbol{j\omega}},\boldsymbol{\theta}_\circ^{\boldsymbol{n}})-\boldsymbol{G}(\boldsymbol{e}^{\boldsymbol{j\omega}},\widehat{\boldsymbol{\theta}}_{\boldsymbol{N}}^{\boldsymbol{n}})|^{\boldsymbol{2}}\right\}\approx\frac{\boldsymbol{1}}{\boldsymbol{N}}\frac{\boldsymbol{\Phi_\nu(\omega)}}{\boldsymbol{\Phi_u(\omega)}}\sum_{\boldsymbol{k=1}}^{\boldsymbol{n}}|\boldsymbol{\mathcal{B}_k}(\boldsymbol{e}^{\boldsymbol{j\omega}})|^{\boldsymbol{2}}.$$

The dividend of considering the orthonormal reparameterisation (11) rather than the perhaps more natural fixed denominator form (1) is that in so doing (56) is derived from asymptotic analysis of an expression (70) which involves generalised Fourier analysis of an underlying function Φ_ν/Φ_u whose smoothness is invariant to model order, or fixed pole selection.

Aside from the theoretical genesis, implementationally the essential feature imbuing the new expression (56) with greater accuracy than the pre-existing approximation (55) is that the influence of the fixed pole location on

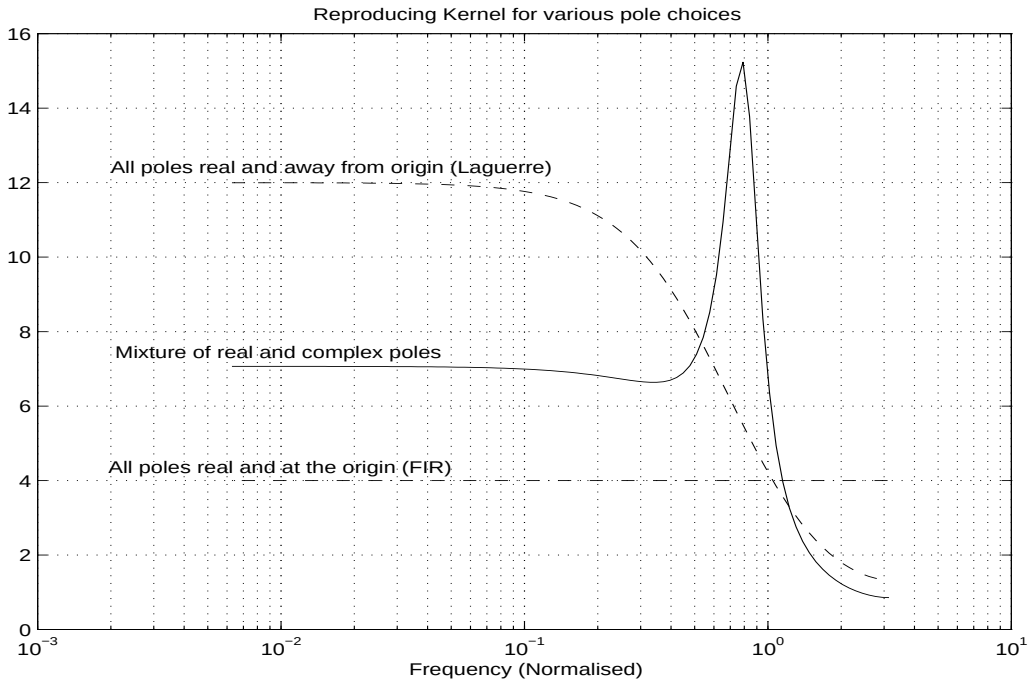

Figure 8: *Plot, for various choices of* $\{\xi_k\}$*, of term* $K_n(\omega,\omega) = \sum_{k=0}^{n-1} |\mathcal{B}_k(e^{j\omega})|^2$ *which captures effect of pole choice* $\{pole_k\}$ *on transfer function estimate sensitivity to measurement noise. Here* $n = 4$*.*

Var$\{G(e^{j\omega}, \widehat{\theta}_N)\}$ is quantified by the reproducing kernel $K_n(\omega,\omega)$. See for example figure 8 where the expression $K_n(\omega,\omega)$ is plotted for a variety of choices of $\{\xi_k\}$. In particular, note that for all poles fixed at the origin, by the formulation (43), (14) then $K_n(\omega,\omega) = n$ so that in this special case of FIR modeling, (56) is identical to (55). However, the more poles that are not fixed at the origin, the more $K_n(\omega,\omega)$ will (being then frequency dependent) differ from n and hence the more the new approximation (56) will, in the interests of improved accuracy, be perturbed from the original approximation (55).

Another point worth emphasising is that the approximation (56) applies for any model structure (of which (11) and (1) are special cases) of the form

$$G(q,\theta^n)u_t = \phi_t^T\theta^n \tag{72}$$

where

$$\phi_{t+1} = A\phi_t + Bu_t \tag{73}$$

with $A \in \mathbf{R}^{n\times n}$ and $B \in \mathbf{R}^{n\times 1}$ arbitrary but such that the eigenvalues of A are $\{\xi_0, \xi_1, \cdots, \xi_{n-1}\}$. It seems quite unexpected that this role for the orthonormal bases $\{\mathcal{B}_k\}$ given by (14) should arise in such a fundamental manner in a problem that can via (72), (73) be ab-initio formulated with no orthonormality in its structure.

3.2 ARX and more general model structures

Although the fixed denominator model structure (11), (1) and its generalisation (72),(73) have many practical advantages, they suffer from the drawback of relying on prior-knowledge for pole location.

A common strategy for avoiding this drawback is to estimate the pole locations of $G(q)$ while still involving a predictor that is linear in θ so that the advantage of simple numerical requirement for finding $\widehat{\theta}_N^n$ is retained. This is done by employing the model structure

$$G(q,\theta) = \frac{B(q,\theta)}{A(q,\theta)}, \quad H(q,\theta) = \frac{D_n(q)}{A(q,\theta)} \tag{74}$$

where

$$\begin{aligned} A(q) &= a_0 + a_1 q + a_2 q^2 + \cdots + a_{n-1} q^{n-1} + q^n, \\ B(q) &= b_0 + b_1 q + b_2 q^2 + \cdots + b_{n-1} q^{n-1} \end{aligned}$$

with

$$\theta = [a_0, b_0, a_1, b_1, \cdots, a_{n-1}, b_{n-1}]^T$$

being the vector of parameters to be estimated and $D_n(q)$ is as previously defined.

Here the dynamics and noise model share parameters in θ. As is well known this can lead to bias if the model structure is not rich enough [72, 44]. The motivation for including the $D_n(q)$ term in the noise model is to avoid this bias by allowing $H(q,\theta_\circ^n) \approx H(q)$ for some $\theta_\circ^n$ while simultaneously, through zeros of $B(q,\theta_\circ^n)$ cancelling parts of $A(q,\theta_\circ^n)$ that pertain only to $H(q,\theta_\circ^n)$, achieving sufficient flexibility for $G(q,\theta_\circ^n) = G(q)$.

Most commonly the model structure (74) appears with the choice $\xi_k = 0$ in which case it is known as the 'equation error' or sometimes 'ARX' model structure and for which the analysis of Ljung [45] provides the well known result (which holds, when at least asymptotically in n, the true system is in the model set) for open loop data collection of

$$\mathbf{E}\left\{ |G(e^{j\omega},\theta_\circ^n) - G(e^{j\omega},\widehat{\theta}_N^n)|^2 \right\} \approx \frac{n}{N} \frac{\Phi_\nu(\omega)}{\Phi_u(\omega)} \tag{75}$$

which is the same as (55) that was just presented as a result only applying to FIR model structures.

A main purpose of this section is to highlight that unfortunately (and in resonance with the fixed denominator case) if the $\{\xi_k\}$ are not all chosen at the origin, then the approximation (75) can be quite inaccurate, even for large model order and data length.

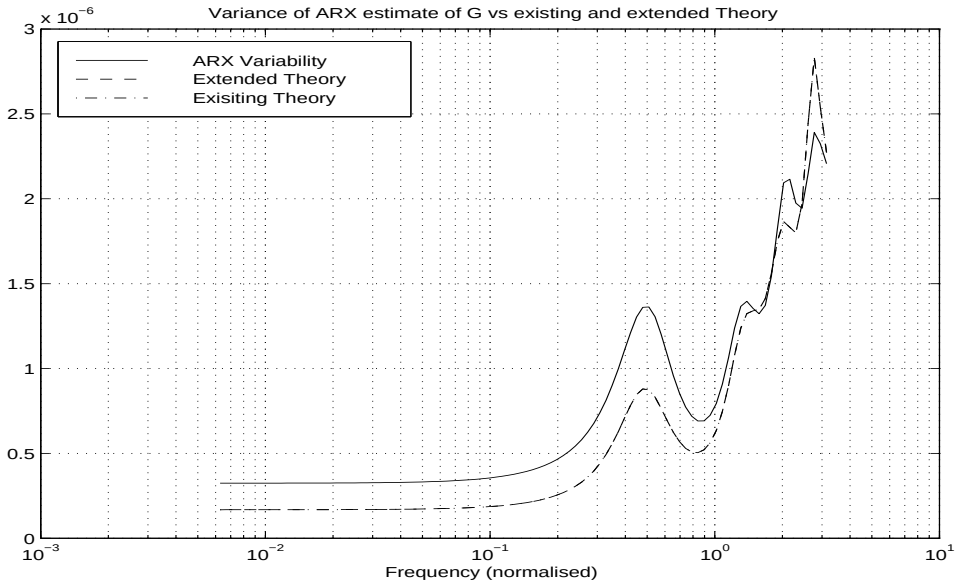

Figure 9: *Conventional ARX with all noise model zeros at the origin. This is a comparison of Monte–Carlo estimate of sample variability (solid line) with (dash-dot line) the approximate expression (75). Note that this last line obscures a dashed line which is the new approximation (56) because for the case of all poles at the origin the pre-existing approximation (75) and the new one (56) are identical.*

This perhaps unexpected phenomenon can be illustrated in a fashion similar to that of previous sections by considering estimation of the 'Åström system'

$$G(q) = \frac{q + 0.5}{q^2 - 1.5q + 0.7}, \quad H(q) = 1$$

using the least squares method (34) and an n=8'th order ARX-like structure (74) on the basis of $N = 10000$ observed open loop input-output measurements, the former being white Gaussian noise with spectral density $\Phi_u(\omega) = 0.25/(1.25 - \cos\omega)$ and the latter being corrupted by white Gaussian noise of variance $\sigma^2 = 0.001$. Suppose also, that all fixed noise model zeroes $\{\xi_k\}$ in $D_n(q)$ are chosen at the origin, so that a true ARX structure is employed. Note that in this example, and all the rest following in this section, the bias error in the estimation process is negligible, and hence the variance error will represent the total estimation error. In any event, since both $N = 10000$ and $n = 8$ can reasonably be considered large [45], then the approximation (75) for the variance error could be expected to be accurate, and indeed it appears to be so when shown as the dash-dot line in figure 9, with the sample average (over 500 Monte–Carlo simulations) estimate of the true variability being shown as the solid line.

However, if three noise model zeros are moved away from the origin to be at

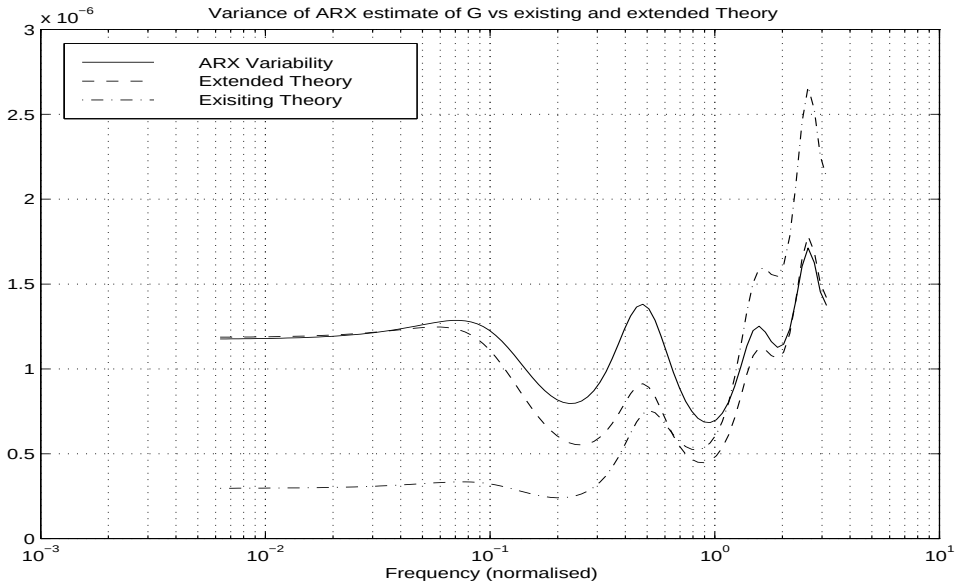

Figure 10: *ARX-type structure with three noise model zeros not at the origin:comparison of Monte–Carlo estimate of sample variability (solid line) with (dash-dot line) the approximate expression (75) and (dashed line) the new approximation (56).*

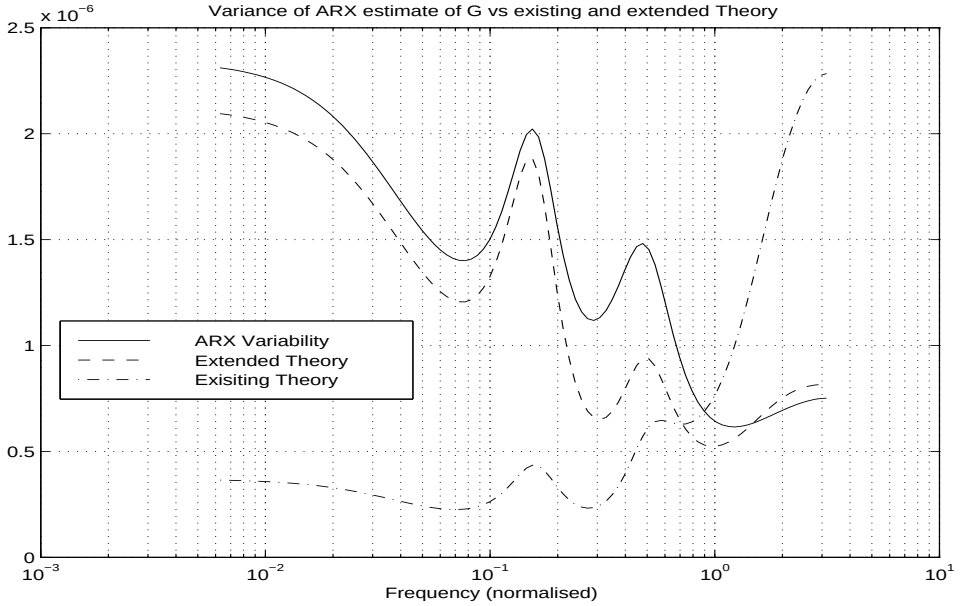

Figure 11: *ARX-type structure with six noise model zeros not at the origin.*

$\xi_k = \{0.8, 0.8, 0.8\}$, then the ensuing comparison of the theoretical (dash-dot line) approximation (75) and the Monte–Carlo estimate (solid line) of true variability shown in figure 10 shows much less agreement. Continuing, by choosing six noise model zeros away from the origin at $\{\xi_k\} = \{0.8, 0.8, 0.8, 0.7, 0.7, 0.7\}$, the results of this choice shown in figure 11 indicate that now the approximation between (dash-dot line) the theoretical approximation (75) and (solid line) the estimated true variability is so poor as to be considered very un-informative.

In contrast, the dashed line in figures 9–11 (which, in figure 9) is equal to and hence obscured by the dash-dot line) remains a good approximation regardless of the fixed zero position.

This line is in fact the approximation (56) applied again, this time to a setting when zeroes of the noise model are fixed rather than poles of dynamic model.

In accordance with the previous fixed denominator analysis, the improved approximation shown in figures 9–11 is obtained by re-parameterising the model structure into an equivalent orthonormal form which is tractable to analysis using the new generalised Fourier and Toeplitz results of [59] presented earlier.

The unifying feature between the fixed denominator and fixed noise model zero cases are that they can both be considered cases of data pre-filtering with the fixed all-pole filter $F(q) = 1/D_n(q)$. In fact, we would suggest that regardless of what model structure (Box-Jenkins, output error) is used in the data-filtered formulation of (31) as

$$F(q)y_t = G(q,\theta)F(q)u_t + H(q,\theta)e_t$$

then whenever such data pre-filtering is used which is substantially all-pole, then the expression (56) should be used to provide the most accurate variance error quantification. More detailed comment on this point is available in [58].

4 Historical comments

Having provided some discussion as to how the ideas of rational orthonormal bases may be applied to certain current system identification problems, we close with a brief survey of the pedigree of these ideas by tracing their genealogy.

1800's Although rational functions are considered in these notes, the consideration of polynomials seems to be involved in the genesis of the ideas wherein during the late 19'th century it was noted that differential equations defined operators, and that these operators often have an integral representation that is 'self-adjoint' and hence defines an orthonormal basis that the solution is most simply expressed with respect to. Nowadays such differential equations are termed 'Sturm-Liouville' systems, well

known cases being Schrödingers equation and the so-called Wave equation [9]. In these instances, the orthonormal bases are defined in terms of Laguerre and Legendre orthonormal functions that themselves are defined in terms of Laguerre and Legendre orthogonal polynomials.

These are polynomials that are orthogonal on the real line with respect to certain weights, and they have been studied since the time of Gauss in relation to efficient methods of numerically evaluating integrals via so-called 'quadrature' formulae. Nowadays they are still being evaluated for this purpose, and others related to studying problems of approximation theory.

The relationship between these Laguerre and Legendre polynomials and the Laguerre and Legendre bases mentioned in modern day signal processing and control literature is via Fourier and Bilinear Transformation of the so-called Laguerre and Legendre functions, which are the corresponding polynomials multiplied by exponential functions. See [55] for details on these points, together with discussion of the relationship of other classical 19'th and early 20'th century orthogonal polynomials (Chebychev polynomials for example) to rational orthonormal bases that are relevant to modern day signal and control theory problems.

1925-1928 The first mention of rational orthonormal bases seems to have occurred in this period with the independent work of Takenaka [67] and Malmquist [51]. The particular bases considered were those formulated here in (14). The context of this early work was application to approximation via interpolation, with the ensuing implications for generalised quadrature formula's considered.

1930's–1940's Nowadays, the work of Malmquist is more remembered than that of Takenaka because Walsh credits Malmquist with first presenting the formulation (14) that formed the genesis of Walsh's work wide ranging work that further studied the application of these bases for approximation on the unit disk (discrete time analysis in modern day engineering parlance) and on the half plane (continuous time). Walsh wrote dozens of papers on this topic, but the major work [75] collects most of his results over the years.

At the same time, Wiener began examining the particular case of continuous time Laguerre networks for the purpose of building the optimal predictors that he began thinking about at MIT, and later developed during the war for anti-aircraft applications. Early work due to his student Lee appeared as [43] which was later presented more fully in [42]. Wiener's own work was finally declassified in 1949 and appeared as [78], although there are mentions of his thoughts on the utility of orthonormal parameterisations in [77].

Also of fundamental importance in this period was that Szegö provided a unified discussion and analysis of polynomials orthogonal on various domains, including the unit circle. At the same time in Russia, Geronimus was also working on these ideas, but they were not to appear until 1961 [30].

1950's This was a very active period. For example, the classic work [32] appeared in which the earlier orthogonal polynomial ideas of Szegö were applied via the consideration of the associated Toeplitz matrices, to various problems in function approximation, probability theory and stochastic processes. In particular, the famous Levinson recursions for the solution of Wiener's optimal predictors were cast in terms of orthonormal bases and the associated 'three term recurrence' formulas that may be used to calculate the bases.

Also in this period Kautz [37, 38] revisited the formulation (14) and its continuous time version for the purposes of network synthesis, as did the work [33]. From a mathematical point of view, a standard reference on orthogonal polynomials was produced [66], and the genesis of examining orthonormally parameterised approximation problems (which is system identification) via the ideas of 'reproducing kernels' was formulated [2].

1960's In this period, work on orthonormal parameterisations for systems and control applications includes that of [8, 52, 65], which again revisited the bases (14), (although using a different formulation, that of [12]) which examined issues of completeness, that of [14] that was specific to Laguerre bases, and the famous book [35] where chapter 12 provides discussion of orthonormal parameterisations for control theory applications. The classic engineering monograph [21] also provided an overview of the use of certain orthonormal parameterisations for system analysis purposes.

From a mathematical perspective, the book [1] appeared which explored for the first time in monograph form the application of Szegö's orthonormal polynomial bases to various approximation problems. As well, the seminal works [11, 18, 26] appeared, which contain superb expositions of the role of orthogonality and reproducing kernels in an approximation theory context.

1970's This period saw the genesis of the application of orthonormal parameterisations for the purpose of VLSI implementation of digital filter structures. Fettweiss's so-called 'wave digital filter' formulation [28, 29, 27] seems to have begun the area, with the monograph [64] and paper [53] being the most readable account of the ideas. Allied to these ideas was the work [24] examining the use of the general rational orthonormal basis (14) for the purposes of optimal prediction of stationary processes and

that of [36] providing further examination of orthogonal polynomial bases. Laguerre bases were examined for the purposes of system identification in [40, 39].

1980's-1990's This period has seen, at least in the engineering literature, and explosion of interest in the use of orthonormal parameterisations. In a signal processing setting see [19, 79, 17, 62, 54], in a control theory oriented system identification setting see [13, 60, 70, 74, 34, 63, 56, 6, 15, 16, 4, 5, 3], in a model approximation setting see [49, 50, 61, 73], for applications to adaptive control problems see [81, 80, 25], for applications to VLSI implementation of digital filter structures see [20, 69], and for applications to modelling and predicting stationary stochastic processes see [71, 24, 23, 22]

References

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

[2] N. Aronszajn. Theory of reproducing kernels. *Acta Mathematica*, pages 337–404, May 1950.

[3] Hüseyin Açkay and Brett Ninness. Rational basis functions for robust identification from frequency and time domain measurements. *Technical Report EE9718, Department of Electrical and Computer Engineering, University of Newcastle, Australia. Submitted to Automatica*, 1997. `http://www.ee.newcastle.edu.au/users/staff/brett`

[4] Per Bodin, Tomas Oliveira e Silva, and Bo Wahlberg. On the construction of orthonormal basis functions for system identification. In *Proceedings of the 13'th IFAC World Congress, San Francisco*, pages 291–296, 1996.

[5] Per Bodin and Bo Wahlberg. Thresholding in higher order transfer function estimation. In *Proceedings of the 33rd IEEE Conference on Decision and Control*, pages 3400–3405, 1994.

[6] J. Bokor, L. Gianone, and Z. Szabo. Construction of generalised orthonormal bases in $\mathcal{H}_2$. Technical report, Computer and Automation Institute, Hungarian Academy of Sciences, 1995.

[7] A. Böttcher and B. Silbermann. *Invertibility and Asymptotics of Toeplitz Matrices*. Akadamie–Verlag, Berlin, 1983.

[8] Paul W. Broome. Discrete orthonormal sequences. *Journal of the Association for Computing Machinery*, 12(2):151–168, April 1965.

[9] Fredrick Byron and Robert Fuller. *Mathematics of Classical and Quantum Physics*, volume 1 and 2 of *Series in Advanced Physics*. Addison-Wesley, 1969.

[10] P.E. Caines. *Linear Stochastic Systems*. John Wiley and Sons, New York, 1988.

[11] E.W. Cheney. *Introduction to Approximation Theory*. McGraw Hill, New York, 1966.

[12] Preston R. Clement. On completeness of basis functions used for signal analysis. *SIAM Review*, 5(2):131–139, 1963.

[13] Preston R. Clement. Laguerre functions in signal analysis and parameter identification. *Journal of the Franklin Institute*, 313(2):85–95, 1982.

[14] G.J. Clowes. Choice of time-scaling facotr for linear systems approximations using orthonormal laguerre functions. *IEEE Transactions on Automatic Control*, 10(4):487–489, October 1965.

[15] W.R. Cluett and L. Wang. Some asymptotic results in recursive identification using Laguerre models. *International Journal of Adatptive Control and Signal Processing*, 5:313–333, 1991.

[16] W.R. Cluett and L. Wang. Frequency smoothing using laguerre model. *Proceedings of the IEE-D*, 139:88–96, 1992.

[17] G.W. Davidson and D.D. Falconer. Reduced complexity echo cancellation using orthonormal functions. *IEEE Transactions on Circuits and Systems*, 38(1):20–28, January 1991.

[18] P.J. Davis. *Interpolation and Approximation*. Blaisdell Publishing Company, 1963.

[19] Albertus C. den Brinker. Laguerre-domain adaptive filters. *IEEE Transactions on Signal Processing*, 42(4):953–956, April 1994.

[20] E. Deprette and P. DeWilde. Orthogonal cascade realization of real multiport digital filters. *Circuit Theory and Applications*, 8:245–272, 1980.

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

[22] P. Dewilde and H. Dym. Lossless inverse scattering, digital filters and estimation theory. *IEEE Transactions on Information Theory*, 30(4):664–662, July 1984.

[23] Patrick DeWilde and Harry Dym. Schur recursions, error formulas, and convergence of rational estimators for stationary stochastic sequences. *IEEE Transactions on Information Theory*, IT-27(4):446–461, July 1981.

[24] Patrick DeWilde, Augusto Vieira, and Thomas Kailath. On a generalised Szegö–Levinson realization algorithm for optimal linear predictors based on a network synthesis approach. *IEEE Transactions on Circuits and Systems*, CAS-25(9):663–675, September 1978.

[25] G.A. Dumont, Y. Fu, and A.L. Elshafi. Orthonormal functions in identification and adaptive control. *Proceedings of IFAC International Symposium on Intelligent Tuning and Adaptive Control, Singapore*, 1990.

[26] Bernard Epstein. *Orthogonal Families of Analytic Functions*. Macmillan, 1965.

[27] A. Fettweiss. Factorisation of transfer matrices of lossless two-ports. *IEEE Transactions on Circuit Theory*, 17:86–94, 1970.

[28] A. Fettweiss. Digital filter structures related to classical filter networks. *Archive für Elektronik und Übertragung*, 25:79–89, 1971.

[29] A. Fettweiss and K. Meerkötter. Suppression of parasitic oscillations in wave digital filters. *IEEE Trans. Circuits and Systems*, 22:239–246, 1975.

[30] L. Ya Geronimus. *Orthogonal polynomials: Estimates, asymptotic formulas, and series of polynomials orthogonal on the unit circle and on an interval*. Consultants Bureau, New York, 1961. Authorized translation from the Russian.

[31] Gene Golub and Charles Van Loan. *Matrix Computations*. Johns Hopkins University Press, 1989.

[32] U. Grenander and G. Szegö. *Toeplitz Forms and their Applications*. University of California Press, Berkeley, 1958.

[33] J.W. Head. Approximation to transient by means of laguerre series. *Proceedings of the Cambridge Philosophical Society*, 52, 1956.

[34] P.S.C. Heuberger, P.M.J. Van den Hof, and O.H. Bosgra. A generalized orthonormal basis for linear dynamical systems. *IEEE Transactions on Automatic Control*, AC-40(3):451–465, March 1995.

[35] I. M. Horowitz. *Synthesis of Feedback Systems*. Academic Press, New York, 1963.

[36] T. Kailath, A. Vieira, and M. Morf. Inverses of Toeplitz operators, Innovations, and Orthogonal Polynomials. *SIAM Review*, 20(1):106–119, January 1978.

[37] William H. Kautz. Network synthesis for specified transient response. Technical Report 209, Massachusetts Institute of Technology, Research Laboratory of Electronics, 1952.

[38] William H. Kautz. Transient synthesis in the time domain. *IRE Transactions on Circuit Theory*, 1:29–39, 1954.

[39] R.E. King and P.N. Paraskevopoulos. Digital Laguerre filters. *Circuit Theory and Applications*, 5:81–91, 1977.

[40] R.E. King and P.N. Paraskevopoulos. Parametric identification of discrete time SISO systems. *International Journal of Control*, 30(6):1023–1029, 1979.

[41] T.W. Körner. *Fourier Analysis*. Cambridge University Press, 1988.

[42] Yuk Wing Lee. *Statistical Theory of Communication*. Wiley, New York, 1960.

[43] Y.W. Lee. Synthesis of electric networks by means of the fourier transforms of laguerre's functions. *Journal of Mathematics and Physics*, XI:83–113, 1933.

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

[45] L.Ljung. Asymptotic variance expressions for identified black-box transfer function models. *IEEE Transactions on Automatic Control*, AC-30(9):834–844, 1985.

[46] L.Ljung and P.E.Caines. Asymptotic normality of prediction error estimators for approximate system models. *Stochastics*, 3:29–46, 1979.

[47] L.Ljung and B. Wahlberg. Asymptotic properties of the least squares method for estimating transfer functions and disturbance spectra. *Advances in Applied Probability*, 24:412–440, 1992.

[48] L.Ljung and Z.D.Yuan. Asymptotic properties of black-box identification of transfer functions. *IEEE Transactions on Automatic Control*, 30(6):514–530, 1985.

[49] P.M. Mäkila. Approximation of stable systems by Laguerre filters. *Automatica*, 26:333–345, 1990.

[50] P.M. Mäkilä. Laguerre series approximation of infinite dimensional systems. *Automatica*, 26:985–995, 1990.

[51] F. Malmquist. Sur la détermination d'une classe de fonctions analytiques par leurs valeurs dans un ensemble donné de points. *Comptes Rendus du Sixième Congrès des mathématiciens scandinaves (Kopenhagen)*, pages 253–259, 1925.

[52] J.M. Mendel. A unified approach to the synthesis of orthonormal exponential functions useful in systems analysis. *IEEE Transactions on Systems Science and Cybernetics*, 2(1):54–62, August 1966.

[53] C.T. Mullis and R.A. Roberts. Roundoff noise in digital filters: Frequency transformations and invariants. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(6):538–550, 1976.

[54] Brett Ninness and Juan Carlos Gómez. Frequency domain analysis of tracking and noise properties of adaptive algorithms. *To appear, IEEE Transactions on Signal Processing*, 1996.

[55] Brett Ninness and Fredrik Gustafsson. A unifying construction of orthonormal bases for system identification. Technical Report EE9432, Department of Electrical and Computer Engineering at the University of Newcastle, August 1994. `http://www.ee.newcastle.edu.au/users/staff/brett`

[56] Brett Ninness and Fredrik Gustafsson. A unifying construction of orthonormal bases for system identification. *IEEE Transactions on Automatic Control*, 42(4):515–521, April 1997.

[57] Brett Ninness and Håkan Hjalmarsson. Model structure and numerical properties of normal equations. *Technical Report EE9801, Department of Electrical and Computer Engineering, University of Newcastle, Australia.*, 1998. `http://www.ee.newcastle.edu.au/users/staff/brett`

[58] Brett Ninness, Håkan Hjalmarsson, and Fredrik Gustafsson. The fundamental role of general orthonormal bases in system identification. *Technical Report EE9739, Department of Electrical and Computer Engineering, University of Newcastle, Australia. Submitted to IEEE Transactions on Automatic Control.*, 1997. `http://www.ee.newcastle.edu.au/users/staff/brett`

[59] Brett Ninness, Håkan Hjalmarsson, and Fredrik Gustafsson. Generalised Fourier and Toeplitz results for rational orthonormal bases. *Technical Report EE9740, Department of Electrical and Computer Engineering, University of Newcastle, Australia.*

Submitted to SIAM Journal on Control and Optimization, 1997. http://www.ee.newcastle.edu.au/users/staff/brett

[60] Ü. Nurges. Laguerre models in problems of approximation and identification. *Adaptive Systems*, pages 346–352, 1987. Pulished by Plenum Publishing, Translated from Avtomatica i Telemekhanika (3) pp 88-96, March 1987.

[61] Jonathan R. Partington. Approximation of delay systems by Fourier-Laguerre series. *Automatica*, 27(3):569–572, 1991.

[62] Hector Perez and Shingeo Tsujii. A system identification algorithm using orthogonal functions. *IEEE Transactions on Signal Processing*, 38(3):752–755, March 1991.

[63] P.M.J. Van den Hof, P.S.C. Heuberger, and J. Bokor. System identification with generalized orthonormal basis functions. *Automatica*, 31(12):1821–1834, December 1995.

[64] R.A. Roberts and C.T. Mullis. *Digital Signal Processing*. Addison-Wesley, 1987.

[65] D.C. Ross. Orthonormal exponentials. *IEEE Transactions on Communication and Electronics*, 71(12):173–176, March 1964.

[66] G. Sansone. *Orthogonal Functions*. Interscience Publishers, New York, 1959.

[67] S. Takenaka. On the orthogonal functions and a new formula of interpolation. *Japanese Journal of Mathematics*, II:129–145, 1925.

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

[69] P.P. Vaidyanathan. A unified approach to orthogonal digital filters and wave digital filters, based on lbr two-pair extraction. *IEEE Transactions on Circuits and Systems*, CAS-32(7):673–686, July 1985.

[70] B. Wahlberg. System identification using Laguerre models. *IEEE Transactions on Automatic Control*, AC-36(5):551–562, 1991.

[71] B. Wahlberg and E.J. Hannan. Parameteric signal modelling using Laguerre filters. *The Annals of Applied Probabililty*, 3(2):467–496, 1993.

[72] B. Wahlberg and L.Ljung. Design variables for bias distribution in transfer function estimation. *IEEE Transactions on Automatic Control*, AC-31:134–144, 1986.

[73] B. Wahlberg and P.M. Mäkilä. On approximation of stable linear dynamical systmes using laguerre and kautz functions. *Automatica*, 32(5):693–708, May 1996.

[74] Bo Wahlberg. System identification using Kautz models. *IEEE Transactions on Automatic Control*, AC-39(6):1276–1282, June 1994.

[75] J.L. Walsh. *Interpolation and Approximation by Rational Functions in the Complex Domain*, volume XX of *Colloquium Publications*. American Mathematical Society, 1935.

[76] H. Widom. *Studies in Real and Complex Analysis*, chapter Toeplitz Matrices. MAA Studies in Mathematics. Prentice Hall, Englewood Cliffs, NJ, 1965.

[77] Norbert Wiener. *The Fourier Integral and Certain of its Applications*. Cambridge University Press, 1933.

[78] Norbert Wiener. *Extrapolation, Interpolation and Smoothing of Stationary Time Series*. M.I.T. Press, 1949.

[79] Geoffrey A. Williamson. Tracking random walk systems with vector space adaptive filters. *IEEE Transactions on Circuits and Systems-II: Analog and Digital Signal Processing*, 42(8):543–547, August 1995.

[80] C.C. Zervos, P.R. Belanger, and G.A. Dumont. Controller tuning using orthonormal series identification. *Automatica*, 24:165–175, 1988.

[81] C.C. Zervos and G.A. Dumont. Deterministic adaptive control based on Laguerre series representation. *International Journal of Control*, 48:2333–2359, 1988.

[82] Y.C. Zhu. Black box identification of MIMO transfer functions:Asymptotic properties of prediction error models. *International Journal of Adaptive Control and Signal Processing*, 3:357–373, 1989.

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