Model Structure and Numerical Properties of Normal Equations

Brett Ninness* Håkan Hjalmarsson†

Abstract

There has been recent interest in using ortho-normalised forms of fixed denominator model structures for system identification. A key motivating factor in the employment of these forms is that of improved numerical properties. Namely, for white input, perfect conditioning of the least-squares normal equations is achieved by design. However, for the more usual case of coloured input spectrum, it is not clear what the numerical conditioning properties should be in relation to simpler and perhaps more natural model structures. This paper provides theoretical and empirical evidence to argue that in fact, even though the orthonormal structures are only designed to provide perfect numerical conditioning for white input, they still provide improved conditioning for a wide variety of coloured inputs.

Technical Report EE9801, Department of Electrical and Computer Engineering, University of Newcastle,AUSTRALIA

1 Introduction

The inspiration for this work is the recent and relatively intense activity [20, 19, 5, 8, 16, 6, 2, 12, 25, 23, 24, 14, 3, 15, 22] proposing the use of an orthonormal parameterisation of linear discrete time systems. In response, this paper poses the question: what is the benefit of forming model structures that are orthonormal with respect to white spectra, but not for the more common case of coloured spectra? The answer found here via theoretical and empirical argument is that the orthonormal model structures are, in numerical conditioning terms, particularly *robust* to spectral colouring while simpler more natural forms are particularly fragile. While such a principle has been implicit in the above-mentioned works, to the authors knowledge it has not previously been explicitly analysed as is done here.

To be more specific on these points, this paper focuses on estimation problems in which N point data records of an input sequence $\{u_t\}$ and output sequence $\{y_t\}$ of a linear and time-invariant system are available. It is assumed that this data is generated as follows

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

*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

†This author is with the Department of Sensors, Signals and Systems-Automatic Control, The Royal Institute of Technology, S-100 44 Stockholm, Sweden. This author can be contacted at email:hakan.hjalmarsson@s3.e.kth.se or FAX: +46 8 790 7329

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 $\{\nu_t\}$ is some sort of possible noise corruption. The input sequence $\{u_t\}$ is assumed to be quasi–stationary in the sense used by Ljung [10] so that the limit

$$R_u(\tau) \triangleq \lim_{N\to\infty} \frac{1}{N} \sum_{t=0}^{N-1} \mathbf{E}\{u_t u_{t+\tau}\}$$

exists such that $\sum_\tau |R_u(\tau)| < \infty$. In this case $\{u_t\}$ has an associated spectral density

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

and it is assumed in this paper that $\Phi_u(\omega) > 0$.

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

$$G(q,\beta) = \sum_{k=0}^{p-1} \beta_k \mathcal{F}_k(q). \tag{1}$$

Here the $\{\beta_k\}$ are real valued coefficients and the transfer functions $\{\mathcal{F}_k(q)\}$ may be chosen in various ways, but in every case the poles of the transfer functions $\{\mathcal{F}_k(q)\}$ are selected from the set $\{\xi_0, \xi_1, \cdots, \xi_{p-1}\} \subset \mathbf{D}$ where $\mathbf{D} \triangleq \{z \in \mathbf{C} : |z| < 1\}$ with $\mathbf{C}$ being the field of complex numbers. These 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 [20, 8, 25].

An advantage of this simple model structure is that it is linearly parameterised in $\{\beta_k\}$, so that with $\beta \triangleq [\beta_0, \beta_1, \cdots, \beta_{p-1}]^T$ then the least-squares estimate

$$\widehat{\beta} = \arg\min_{\beta\in\mathbf{R}^p} \left\{ \frac{1}{N} \sum_{t=0}^{N-1} \left(y_t - G(q,\beta)u_t\right)^2 \right\} \tag{2}$$

is easily computed. Specifically, the solution $\widehat{\beta}$ to (2) can be written in closed form once the model structure (1) is cast in familiar linear regressor form notation as $G(q,\beta)u_t = \psi_t^T \beta$ where

$$\psi_t = \Lambda_p(q)\, u_t, \qquad \Lambda_p(q) \triangleq [\mathcal{F}_0(q), \mathcal{F}_1(q), \cdots, \mathcal{F}_{p-1}(q)]^T \tag{3}$$

so that (2) is solved as

$$\widehat{\beta} = \left(\sum_{t=0}^{N-1} \psi_t \psi_t^T \right)^{-1} \sum_{t=0}^{N-1} \psi_t y_t \tag{4}$$

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

However, a large literature [20, 19, 5, 8, 16, 6, 2, 12, 25, 23, 24, 14, 3, 15, 22] has suggested that instead of using the model structure (1), the so-called 'orthonormal' form of it should be employed. That is, the model structure (1) should be re-parameterised as

$$G(q,\theta) = \sum_{k=0}^{p-1} \theta_k \mathcal{B}_k(q) \tag{5}$$

where again the coefficients $\{\theta_k\}$ are real valued, and the $\{\mathcal{B}_k(q)\}$ are transfer functions such that

$$\text{Span}\{\mathcal{F}_0, \mathcal{F}_1, \cdots, \mathcal{F}_{p-1}\} = \text{Span}\{\mathcal{B}_0, \mathcal{B}_1, \cdots, \mathcal{B}_{p-1}\} \tag{6}$$

with the further requirement that the $\{\mathcal{B}_k(q)\}$ are also orthonormal:

$$\langle \mathcal{B}_n, \mathcal{B}_m \rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathcal{B}_n(e^{j\omega})\overline{\mathcal{B}_m(e^{j\omega})}\,\mathrm{d}\omega = \frac{1}{2\pi j}\oint_{\mathbf{T}} \mathcal{B}_n(z)\overline{\mathcal{B}_m(z)}\,\frac{\mathrm{d}z}{z} = \begin{cases} 1 & ; m = n \\ 0 & ; m \neq n. \end{cases} \tag{7}$$

Here $\mathbf{T} \triangleq \{z \in \mathbf{C} : |z| = 1\}$ is the complex unit circle. There have been several orthonormal basis function formulations proposed in the literature [8, 20, 21, 2] but this paper focuses on the particular choice discussed in [12] of

$$\mathcal{B}_n(q) = \begin{cases} \dfrac{\sqrt{1-|\xi_n|^2}}{q-\xi_n}\displaystyle\prod_{k=0}^{n-1}\left(\frac{1-\overline{\xi_k}q}{q-\xi_k}\right) & ; n \geq 1 \\ \\ \dfrac{\sqrt{1-|\xi_0|^2}}{q-\xi_0} & ; n = 0. \end{cases} \tag{8}$$

In this case, defining in a manner analogous to the previous case

$$\phi_t = \Gamma_p(q)u_t, \qquad \Gamma_p(q) \triangleq [\mathcal{B}_0(q), \mathcal{B}_1(q), \cdots, \mathcal{B}_{p-1}(q)]^T \tag{9}$$

then the least squares estimate with respect to the model structure (5) is given as

$$\widehat{\theta} = \left(\sum_{t=0}^{N-1}\phi_t\phi_t^T\right)^{-1}\sum_{t=0}^{N-1}\phi_t y_t. \tag{10}$$

A key point is that since there is a linear relationship $\phi_t = J\psi_t$ for some non-singular J, then $\widehat{\beta} = J^T\widehat{\theta}$ and hence modulo numerical issues the least-squares frequency response estimate is invariant to the change in model structure between (1) and (5). Specifically:

$$\begin{aligned} G(e^{j\omega}, \widehat{\beta}) &= \Lambda_p^T(e^{j\omega})\widehat{\beta} \\ &= \Lambda_p^T(e^{j\omega})\left(\sum_{t=0}^{N-1}\psi_t\psi_t^T\right)^{-1}\sum_{t=0}^{N-1}\psi_t y_t \\ &= \Lambda_p^T(e^{j\omega})\left[J^{-1}\left(\sum_{t=0}^{N-1}\phi_t\phi_t^T\right)J^{-T}\right]^{-1}J^{-1}\sum_{t=0}^{N-1}\phi_t y_t \\ &= [J\Lambda_p(e^{j\omega})]^T\left(\sum_{t=0}^{N-1}\phi_t\phi_t^T\right)^{-1}\sum_{t=0}^{N-1}\phi_t y_t \\ &= \Gamma_p^T(e^{j\omega})\widehat{\theta} \\ &= G(e^{j\omega}, \widehat{\theta}). \end{aligned}$$

Given this exact equivalence of frequency response estimates, it is important to question the motivation for using the structure (8) (which is complicated by the precise definition of the orthonormal bases (8) or whichever other one is used [8, 2]) in place of some other one such as (1). In particular, depending

on the choice of the $\{\mathcal{F}_k(q)\}$, the structure (1) may be more natural and/or be more straightforward to implement, so it is important to examine the rationale for employing the equivalent ortho-normalised version (5).

To date, a major part of addressing this question has been to motivate the use of the orthonormal form (5) along numerical conditioning lines [20, 21, 8, 12]. To elaborate further on this point, it is well known [7] that the numerical properties of the solution of the normal equations arising in least squares estimation using the model structures (1) and (5) are governed by the condition numbers $\kappa(R_\psi(N))$ and $\kappa(R_\phi(N))$ of the matrices

$$R_\psi(N) \triangleq \frac{1}{N}\sum_{t=0}^{N-1}\psi_t\psi_t^T, \qquad R_\phi(N) \triangleq \frac{1}{N}\sum_{t=0}^{N-1}\phi_t\phi_t^T$$

where the vectors ψ_t and ϕ_t are defined in (3) and (9) respectively. However, by the quasi–stationarity assumption and by Parseval's Theorem, the following limits exist:

$$R_\psi \triangleq \lim_{N\to\infty} R_\psi(N) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\Lambda_p(e^{j\omega})\Lambda_p^\star(e^{j\omega})\Phi_u(\omega)\,\mathrm{d}\omega \tag{11}$$

$$R_\phi \triangleq \lim_{N\to\infty} R_\phi(N) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\Gamma_p(e^{j\omega})\Gamma_p^\star(e^{j\omega})\Phi_u(\omega)\,\mathrm{d}\omega, \tag{12}$$

(here $\cdot^\star$ denotes 'conjugate transpose') so that the numerical properties of least squares estimation using the model structures (1) and (5) should be closely related to the condition numbers $\kappa(R_\psi)$ and $\kappa(R_\phi)$. These condition number quantities, are defined for an arbitrary non-singular matrix R as [7]

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

which is clearly dependent on the matrix norm chosen. Most commonly, the matrix 2-norm is used [7], 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, and is a measure of the Euclidean norm sensitivity of the solution vector x of the equation $Rx = b$ to errors in the vector b. If not specified otherwise, it will be understood in this paper that this 2-norm defined condition number is being considered.

Now, for white input $\{u_t\}$, by definition its spectrum $\Phi_u(\omega)$ is a constant (say α) so that by orthonormality $R_\phi = \alpha I$ and hence the normal equations are perfectly numerically conditioned. However, an obvious question concerns how the condition numbers of R_ψ and R_ϕ compare for the more commonly encountered coloured input case. A key result in this context is that purely by virtue of the orthonormality in the structure (5), an upper bound on the conditioning of R_ϕ may be guaranteed for any Φ_u by virtue of the fact that [16, 13] ($\lambda(R)$ denotes the set of eigenvalues of the matrix R)

$$\min_{\omega\in[-\pi,\pi]}\Phi_u(\omega) \le \lambda(R_\phi) \le \max_{\omega\in[-\pi,\pi]}\Phi_u(\omega). \tag{13}$$

No such bounds are available for the matrix R_ψ corresponding to the general (non-orthonormal) structure (1). This suggests that the numerical conditioning associated with (5) might be superior to that of (1) across a range of coloured Φ_u, and not just the white Φ_u that the structure (5) 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 $R_\psi = I$ might equally well occur

for some particular coloured input. If so, then in this latter scenario the structure (5) would actually be inferior to (1) in numerical conditioning terms. Therefore, in spite of the guarantee (13), it is not clear when and why numerical considerations would lead to the structure (5) being preferred over the often-times simpler one (1).

The rest of this paper is devoted to examining these questions. In addressing them, the paper begins in §2 by establishing a general framework for studying the question of the existence of a spectrum Φ_u for which perfect numerical conditioning occurs. Using this framework, §3 and §4 establish first by a simple 2-dimensional motivating example, and then for the case of arbitrary dimension, that firstly it may easily be the case that R_ψ is never a perfectly conditioned diagonal matrix for any Φ_u and secondly, the manifolds of all possible R_ψ and R_ϕ are not the complete manifold of all possible symmetric $p \times p$ dimensional positive definite matrices. Instead, the respective manifolds of R_ψ and R_ϕ are of what may be much smaller dimension. Therefore, since a perfectly conditioned matrix is, by construction, in the manifold of possible R_ϕ, and since the possible manifolds of R_ψ and R_ϕ are restricted, this provides further evidence that parameterisation with respect to an orthonormal basis may provide improved numerical conditioning across a range of possible input spectra.

Further aspects of this conjecture are examined in greater detail in §5 and §6 by a strategy of deriving approximations for the eigenvalue locations of R_ϕ. These refine (13) in that they are expressed directly in terms of the Φ_u (actually, in terms of its positive real part) and the location of the fixed poles $\{\xi_k\}$ in such a way as to illustrate that the numerical conditioning of R_ϕ is (as is intuitively reasonable) closely related to the smoothness of Φ_u.

In §7, for the specific case of $p = 2$, and for specific examples of $\mathcal{F}_0, \mathcal{F}_1$, a class of Φ_u are derived for which R_ϕ is guaranteed to have smaller condition number than R_ψ. In §8, this is generalised by analysis that is asymptotic in p, and is such as to establish that for model structures (3) with the $\{\mathcal{F}_k(q)\}$ chosen so that essentially a numerator is being estimated and a denominator $D_p(q)$ is being fixed, then this leads to poorer numerical conditioning than if the equivalent orthonormal structure (5) is used provided that the variation (across $\omega \in [-\pi, \pi]$) of $\Phi_u(\omega)$ is smaller than that of $\Phi_u(\omega)/|D_p(e^{j\omega})|^2$. Finally, §9 provides some concluding perspectives on the work presented here.

Note that as previously mentioned, although there are a number of possible alternatives [8, 21, 2, 18] for the construction of orthonormal bases that satisfy the span condition (6), the particular choice (8) will be used here. The reason for this is that the formulation (8) offers an explicit formulation for the orthonormal bases, and this will prove to be essential for the precise characterisation of the spectral properties of R_ϕ. Note also, that under the span condition (6), all choices of orthonormal bases will lead to matrices R_ϕ that are unitarily congruent to one another, and which therefore possess precisely the same spectral properties. Therefore, any spectral conclusions made relative to the basis (8) will in fact apply to any orthonormal basis, such as the ones considered in [8, 2].

2 Existence of Spectra

This section addresses the issue of the existence of a particular coloured Φ_u for which the non-orthonormal model structure (1) leads to perfect conditioning ($R_\psi = I$) and would thus make it a superior choice on numerical grounds relative to the 'orthonormal' structure (5). This issue is, in fact, subsumed by that of designing a $\Phi_u(\omega)$ parameterised via real valued coefficients $\{c_k\}$ as

$$\Phi_u(\omega) = \sum_{k=-\infty}^{\infty} c_k e^{j\omega k} \tag{14}$$

and so as to achieve an arbitrary symmetric, positive definite R_ψ. 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

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

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

$$\Pi \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ \vdots \end{bmatrix} = \mathrm{vec}\{R_\psi\} \tag{15}$$

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

$$\Pi \triangleq \frac{1}{2\pi j} \oint_{\mathbf{T}} [\Lambda_p(z) \otimes I_p] \overline{\Lambda_p(z)} [1, z + z^{-1}, z^2 + z^{-2}, \cdots] \, \frac{\mathrm{d}z}{z}. \tag{16}$$

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

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

The solution of (15) must be performed subject to the constraint that the Toeplitz matrix

$$\begin{bmatrix} c_0 & c_1 & c_2 & \cdots \\ c_1 & c_0 & c_1 & \\ c_2 & & \ddots & \\ \vdots & & & \ddots \end{bmatrix}$$

is positive definite, which is a necessary and sufficient condition [17] for $\Phi_u(\omega) > 0$.

Now it might be supposed that since (15) is an equation involving $p(p+1)/2$ 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_ψ.

Perhaps surprisingly, this turns out not to be the case, the reason being that (as established in Theorem 4.1 following) the rank of Π in (16) is always only p. In fact therefore, the achievable R_ψ live only in a sub-manifold of the $p(p+1)/2$ dimensional manifold of $p \times p$ symmetric matrices, and this sub-manifold *may not contain a perfectly conditioned matrix*. Furthermore, as can be seen by (16), this sub-manifold that the possible R_ψ lie in will be completely determined by the choice of the functions $\mathcal{F}_k(z)$ in the model structure (1) and hence also in the definition for $\Lambda_p(z)$ in (3). These principles are most clearly exposed by considering some simple two dimensional examples.

3 Two Dimensional Example

Consider the simplest case of $p = 2$ wherein there are only 3 constraints inherent in (15), and one may as well neglect the third row of $[\Lambda_p(z) \otimes I_p]\overline{\Lambda_p(z)}$ (since it is equal, by symmetry, to the second row) and instead consider

$$\begin{bmatrix} \mathcal{F}_0(z)\mathcal{F}_0(1/z) \\ \mathcal{F}_0(z)\mathcal{F}_1(1/z) \\ \mathcal{F}_1(z)\mathcal{F}_1(1/z) \end{bmatrix} = \begin{bmatrix} \mathcal{F}_0(1/\xi_0)\mathcal{F}_0(z) + (1/z\xi_0)\mathcal{F}_0(1/\xi_0)\mathcal{F}_0(1/z) \\ \mathcal{F}_1(1/\xi_0)\mathcal{F}_0(z) + (1/z\xi_1)\mathcal{F}_0(1/\xi_1)\mathcal{F}_1(1/z) \\ \mathcal{F}_1(1/\xi_1)\mathcal{F}_1(z) + (1/z\xi_1)\mathcal{F}_1(1/\xi_1)\mathcal{F}_1(1/z) \end{bmatrix} \tag{17}$$

where in forming the right hand side of the above equation it has been assumed that $\mathcal{F}_0(z)$ has a pole at $z = \xi_0$, $\mathcal{F}_1(z)$ has a pole at $z = \xi_1$, that $\mathcal{F}_0(0) \neq 0$, $\mathcal{F}_1(0) \neq 0$ and that $\xi_0, \xi_1 \in \mathbf{R}$. That is $\mathcal{F}_0(z)$ and $\mathcal{F}_1(z)$ are of the simple form

$$\mathcal{F}_0(z) \triangleq \frac{1}{z - \xi_0}, \quad \mathcal{F}_1(z) \triangleq \frac{1}{z - \xi_1}, \quad \xi_0, \xi_1 \in \mathbf{R}. \tag{18}$$

The advantage of the re–parameterisation into causal and anti–causal components in (17) is that it is then straightforward to calculate Π from (16) as

$$\Pi = \begin{bmatrix} \mathcal{F}_0(1/\xi_0)(1/\xi_0) & 2\mathcal{F}_0(1/\xi_0) & 2\mathcal{F}_0(1/\xi_0)\xi_0 & \cdots \\ \mathcal{F}_0(1/\xi_1)(1/\xi_1) & \mathcal{F}_1(1/\xi_0) + \mathcal{F}_0(1/\xi_1) & \mathcal{F}_1(1/\xi_0)\xi_0 + \mathcal{F}_0(1/\xi_1)\xi_1 & \cdots \\ \mathcal{F}_1(1/\xi_1)(1/\xi_1) & 2\mathcal{F}_1(1/\xi_1) & 2\mathcal{F}_1(1/\xi_1)\xi_1 & \cdots \end{bmatrix}. \tag{19}$$

Given this formulation, it is then clear that

$$\left[\frac{\mathcal{F}_1(1/\xi_0)}{2\mathcal{F}_0(1/\xi_0)}, -1, \frac{\mathcal{F}_0(1/\xi_1)}{2\mathcal{F}_1(1/\xi_1)}, \right] \Pi = [0, 0, 0, \cdots] \tag{20}$$

provided that

$$\mathcal{F}_0(1/\xi_1)\xi_0 = \mathcal{F}_1(1/\xi_0)\xi_1 \tag{21}$$

which is certainly true for the first order $\mathcal{F}_0(z), \mathcal{F}_1(z)$ in (18). Therefore, Π is of row (and hence column) rank no more than two. Therefore, regardless of the choice of Φ_u, it is only possible to manipulate (via change of Φ_u) the corresponding R_ψ in a two dimensional sub-manifold of the full three dimensional manifold of symmetric two-by-two matrices.

Furthermore, the identity matrix is not part of the two-dimensional sub-manifold, since if it were to lie in the subspace spanned by the columns of Π, it would have to be orthogonal to the normal vector specifying the orientation of this subspace (the left hand row vector in (20)). But it isn't, since

$$\left[\frac{\mathcal{F}_1(1/\xi_0)}{2\mathcal{F}_0(1/\xi_0)}, -1, \frac{\mathcal{F}_0(1/\xi_1)}{2\mathcal{F}_1(1/\xi_1)}\right] \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \neq 0$$

provided $\mathcal{F}_0, \mathcal{F}_1$ are of the form shown in (18). In fact, by the same argument, no diagonal matrix with positive valued entries is part of the manifold of achievable covariance matrices.

Therefore, even though Φ_u can be viewed as an infinite dimensional quantity, its effect on R_ψ is not powerful enough to achieve an arbitrary positive definite symmetric matrix. In particular, there is no Φ_u for which the simple and natural fixed denominator basis (18) is perfectly conditioned.

However, if instead of (18) the alternative simple and natural choice

$$\mathcal{F}_0(z) \triangleq \frac{1}{(z-\xi_0)(z-\xi_1)}, \quad \mathcal{F}_1(z) \triangleq \frac{z}{(z-\xi_0)(z-\xi_1)} \tag{22}$$

for the fixed denominator basis functions are made, then again straightforward (but tedious) calculation provides

$$\Pi = C \begin{bmatrix} \frac{\xi_0}{1-\xi_0^2} - \frac{\xi_1}{1-\xi_1^2} & \frac{\xi_0^2}{1-\xi_0^2} - \frac{\xi_1^2}{1-\xi_1^2} & \frac{\xi_0^3}{1-\xi_0^2} - \frac{\xi_1^3}{1-\xi_1^2} & \cdots \\ \frac{1}{2}\left(\frac{1+\xi_0^2}{1-\xi_0^2}\right) - \frac{1}{2}\left(\frac{1+\xi_1^2}{1-\xi_1^2}\right) & \left(\frac{1+\xi_0^2}{1-\xi_0^2}\right)\xi_0 - \left(\frac{1+\xi_1^2}{1-\xi_1^2}\right)\xi_1 & \left(\frac{1+\xi_0^2}{1-\xi_0^2}\right)\xi_0^2 - \left(\frac{1+\xi_1^2}{1-\xi_1^2}\right)\xi_1^2 & \cdots \\ \frac{\xi_0}{1-\xi_0^2} - \frac{\xi_1}{1-\xi_1^2} & \frac{\xi_0^2}{1-\xi_0^2} - \frac{\xi_1^2}{1-\xi_1^2} & \frac{\xi_0^3}{1-\xi_0^2} - \frac{\xi_1^3}{1-\xi_1^2} & \cdots \end{bmatrix}$$

where $C \triangleq (\xi_0 - \xi_1)^{-1}(1-\xi_0\xi_1)^{-1}$ so that again Π is only of rank two, this time since

$$[1, 0, -1]\,\Pi = [0, 0, 0, \cdots].$$

However, the important difference in this case is that since (as shown above) the vector $[1, 0, -1]^T$ is orthogonal to the space spanned by the columns of Π, and since $[1, 0, 1]^T$ is also orthogonal to this vector, then the identity matrix does lie in the manifold of R_ψ that can be generated by the manipulation of Φ_u.

4 Higher Dimensions

Given these motivating arguments specific to a two-dimensional case, it is of interest to consider the case of arbitrary dimension. As the algebra considered in the previous section illustrated, such a study will become very tedious as the dimension is increased. To circumvent this difficulty, the key idea of this section is to replace the study of the rank of Π associated with an arbitrary basis $\{\mathcal{F}_n(q)\}$ (such as those in (18) or (22)) by its rank with respect to the orthonormal basis $\{\mathcal{B}_n(q)\}$ specified in (8). Fundamental to this strategy is that via the span equivalence condition (6) the rank is invariant to the change of basis, so that the most tractable one may as well be employed. The suitability of $\{\mathcal{B}_n\}$ in this context is embodied in the following lemma.

Lemma 4.1. *For $\{\mathcal{B}_n(z)\}$ defined by (8), the inner product defined by (7) and assuming all the $\{\xi_k\}$ are distinct*

$$\langle \mathcal{B}_m(z), \mathcal{B}_n(z) z^k \rangle = \begin{cases} \xi_n^k & ; m = n, k \geq 0, \\ \overline{\xi}_n^{|k|} & ; m = n, k < 0, \\ \displaystyle\sum_{i=m}^{n} A_{m,n}^i \overline{\xi}_i^k & ; n > m, k < 0, \\ 0 & ; n > m, k \leq 0 \end{cases}$$

where

$$A_{m,n}^i \triangleq \frac{\sqrt{(1-|\xi_m|^2)(1-|\xi_n|^2)}(1-|\xi_i|^2)}{(1-\xi_m\overline{\xi}_i)(1-\xi_n\overline{\xi}_i)} \prod_{\substack{k=m \\ k\neq i}}^{n} \left(\frac{1-\xi_k\overline{\xi}_i}{\overline{\xi}_i - \xi_k}\right) \tag{23}$$

and

$$\sum_{i=m}^{n} A_{m,n}^{i} = 0.$$

Proof. Suppose that $m = n$. Then using the formulation (8) and in the case of $k \geq 0$

$$\begin{aligned}\langle \mathcal{B}_m(z), \mathcal{B}_n(z)z^k \rangle &= \frac{1}{2\pi j}\oint_{\mathbf{T}} \frac{(1-|\xi_n|^2)z}{(1-\xi_n z)(z-\overline{\xi}_n)} z^{-k}\,\frac{\mathrm{d}z}{z} \\ &= \frac{1}{2\pi j}\oint_{\mathbf{T}} \frac{(1-|\xi_n|^2)z}{(z-\xi_n)(1-\overline{\xi}_n z)} z^{k}\,\frac{\mathrm{d}z}{z} \\ &= \xi_n^k\end{aligned}$$

where the change of variable $z \mapsto 1/z$ was employed in progressing to the last line. Similarly, for $k < 0$ the result

$$\langle \mathcal{B}_m(z), \mathcal{B}_n(z)z^k \rangle = \overline{\xi_n}^{|k|}$$

will clearly emerge. Now suppose (without loss of generality by symmetry) that $n > m$. In this case, for $k \leq 0$

$$\langle \mathcal{B}_m(z), \mathcal{B}_n(z)z^k \rangle = \frac{1}{2\pi j}\oint_{\mathbf{T}} \frac{\sqrt{(1-|\xi_m|^2)(1-|\xi_n|^2)}}{(z-\xi_m)(1-\overline{\xi}_n z)} \prod_{\ell=m}^{n-1}\left(\frac{z-\xi_\ell}{1-\overline{\xi}_\ell z}\right) z^{|k|}\,\mathrm{d}z = 0$$

with the result following by using Cauchy's integral formula after recognising that the integrand is analytic on the interior of $\mathbf{T}$. Now suppose that $k \geq 0$. Then again employing the change of variable $z \mapsto 1/z$ and Cauchy's Residue Theorem

$$\begin{aligned}\langle \mathcal{B}_m(z), \mathcal{B}_n(z)z^k \rangle &= \frac{1}{2\pi j}\oint_{\mathbf{T}} \frac{\sqrt{(1-|\xi_m|^2)(1-|\xi_n|^2)}}{(1-\xi_m z)(1-\xi_n z)} \prod_{\ell=m}^{n}\left(\frac{1-\xi_\ell z}{z-\overline{\xi}_\ell}\right) z^{k}\,\mathrm{d}z \\ &= \sum_{i=m}^{n} A_{m,n}^{i}\overline{\xi}_i^k\end{aligned}$$

where under the assumption that the $\{\xi_k\}$ are distinct, the $\{A_{m,n}^i\}$ terms are as given in (23). Finally, by setting $k = 0$ and using the orthonormality of the $\{\mathcal{B}_n\}$ and Cauchy's Residue Theorem again:

$$0 = \langle \mathcal{B}_m, \mathcal{B}_n \rangle = \sum_{i=m}^{n} A_{m,n}^{i}.$$

□

This lemma is the key to providing a more important result in Theorem 4.1 on the fundamental flexibility of manipulating R_ϕ or R_ψ by changing Φ_u. However, in order to develop this most clearly it is expedient to split Φ_u into 'causal' and 'anti–causal' components as

$$\Phi_u(\omega) = \varphi(e^{j\omega}) + \varphi(e^{-j\omega}) \tag{24}$$

where $\varphi(z)$ is known as the 'positive real' part of Φ_u and is given by the so-called Herglotz-Riesz transform [17] as

$$\varphi(z) = \frac{c_0}{2} + \sum_{k=1}^{\infty} c_k z^k = \frac{1}{4\pi}\int_{-\pi}^{\pi}\left(\frac{1+ze^{j\omega}}{1-ze^{j\omega}}\right)\Phi_u(\omega)\,\mathrm{d}\omega. \tag{25}$$

With this definition in hand, the following lemma is available which builds on the previous one.

Lemma 4.2. *The matrix R_ϕ defined via (12), (8) and (9) has entries given by*

$$[R_\phi]_{m,n} = \begin{cases} \varphi(\xi_n) + \varphi(\overline{\xi_n}) & ; n = m, \\ \displaystyle\sum_{i=m}^{n} A^i_{m,n}\varphi(\overline{\xi_i}) & ; n > m \end{cases}$$

where $A^i_{m,n}$ is defined in (23) and it is understood that the array indexing of R_ϕ begins at $m, n = 0$.

Proof. By the formulation (12)

$$[R_\phi]_{m,n} \;=\; \langle \mathcal{B}_m, \mathcal{B}_n\Phi_u\rangle = c_0\langle \mathcal{B}_m, \mathcal{B}_n\rangle + \sum_{k=1}^{\infty} c_k\langle \mathcal{B}_m, \mathcal{B}_n z^k\rangle + c_k\langle \mathcal{B}_m, \mathcal{B}_n z^{-k}\rangle.$$

Therefore, if $n = m$, then by Lemma 4.1

$$[R_\phi]_{n,n} = c_0 + \sum_{k=1}^{\infty} c_k(\xi_n^k + \overline{\xi}_n^k) = \varphi(\xi_n) + \varphi(\overline{\xi}_n).$$

On the other hand, if $n > m$ and again using Lemma 4.1

$$[R_\phi]_{m,n} \;=\; \sum_{k=1}^{\infty} c_k \sum_{i=m}^{n} A^i_{m,n}\overline{\xi}_i^k = \sum_{i=m}^{n} A^i_{m,n}\left[\varphi(\overline{\xi}_i) - \frac{c_0}{2}\right] = \sum_{i=m}^{n} A^i_{m,n}\varphi(\overline{\xi}_i)$$

where in progressing to the last line the fact that

$$\sum_{i=m}^{n} A^i_{m,n} = 0$$

has been used. □

Although this Lemma will be used later for further developments, its main purpose here is to settle the question raised earlier in §2, 3 as to just how much flexibility there is in the assignment of R_ϕ, R_ψ by manipulation of the spectral density Φ_u.

Theorem 4.1. *With Π defined as in (16), and for all bases that maintain the same span as in condition (6), then the rank of Π is given as*

$$\operatorname{Rank}\Pi = p.$$

Proof. The main idea of the proof is to recognise that the rank of Π defined in (16) is invariant to a change of the basis function $\{\mathcal{F}_k\}$ making up Λ_p involved in the definition of Π, and itself defined in (3). This statement must be made subject to the proviso that in making the change of basis, the underlying space being spanned remains the same, which is condition (6). This is because under this assumption, and denoting the two matrices resulting from two different bases as R_ψ, and R'_ψ, then a non-singular $p \times p$ matrix J will exist such that $R_\psi = JR'_\psi J^T$. Since the rank of Π is the number of degrees of freedom in the choice of the components of R_ψ by manipulation of the $\{c_k\}$ parameterising Φ_u via (14), then provided J is non-singular, these degrees of freedom are invariant to congruence transformations by J.

With this idea in hand, the proof proceeds by electing to make the span-preserving change of basis

$$\{\mathcal{F}_0, \mathcal{F}_1, \cdots, \mathcal{F}_{p-1}\} \mapsto \{\mathcal{B}_0, \mathcal{B}_1, \cdots, \mathcal{B}_{p-1}\}$$

with the $\{\mathcal{B}_n\}$ being as defined in (8). In this case the rank of Π is the number of effective degrees of freedom in the formation of the elements of R_ϕ by means of the choice of the $\{c_k\}$. But this is the same as the effective degrees of freedom in forming R_ϕ by the choice of $\varphi(z)$, and Lemma 4.2 makes it clear that since all the terms in the $p \times p$ matrix R_ϕ are linear combinations of $\{\varphi(\xi_0), \cdots, \varphi(\xi_{p-1})\}$, then in fact there are only p degrees of freedom in the formation of R_ϕ by the choice of Φ_u. □

This theorem exposes the key feature imbuing orthonormal parameterisations with numerical robustness beyond the white input case. Specifically, for white input, $R_\phi = I$ is perfectly numerically conditioned, while for this same white input $R_\psi \triangleq \Sigma \neq I$ which has inferior conditioning. As Φ_u is changed from the white case, both R_ϕ and R_ψ will change, but *but only in p-dimensional submanifolds.*

This feature of highly restricted mobility raises the possibility that since (by construction) I is in the manifold of possible R_ϕ, but may not (as the previous section illustrated) be in the manifold of possible R_ψ, then the orthonormal model structure (8) may impart a numerical robustness to the associated normal equations across a range of coloured Φ_u. Examining this issue consumes the remainder of the paper which is motivated, as previously, by a simple two-dimensional example.

5 Robustness in Two Dimensional Case

To further examine the issue of numerical conditioning being preserved robustly across a range of non-white input spectra, it is again expedient to return to the simple 2×2 case for illustrative purposes. In conjunction with this, assume that the simple fixed denominator basis (18) is again under consideration, and which has associated Π matrix given by (19).

It has just been established that the space of possible R_ψ depend on the column range-space of Π, and that this latter space is two-dimensional. In fact, if Π is restricted to have only three columns, then it is straightforward to verify from (19) that

$$\Pi \begin{bmatrix} 2\xi_0\xi_1 \\ -(\xi_0 + \xi_1) \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

provided again that (21) holds. In this case, the first two columns of Π in (19) completely determine the whole column range space of Π. Therefore, denoting by Σ the matrix R_ψ for white input ($\Phi_u = 1$),

then by (19)

$$\Sigma \triangleq \begin{bmatrix} \mathcal{F}_0(1/\xi_0)(1/\xi_0) & \mathcal{F}_0(1/\xi_1)(1/\xi_1) \\ \mathcal{F}_0(1/\xi_1)(1/\xi_1) & \mathcal{F}_1(1/\xi_1)(1/\xi_1) \end{bmatrix} = \begin{bmatrix} \dfrac{1}{1-\xi_0^2} & \dfrac{1}{1-\xi_0\xi_1} \\ \dfrac{1}{1-\xi_0\xi_1} & \dfrac{1}{1-\xi_1^2} \end{bmatrix} \tag{26}$$

which means that all possible R_ψ are expressible as a perturbation away from Σ as

$$R_\psi = (\alpha_0 + (\xi_0+\xi_1)\alpha_1)\Sigma + \alpha_1(\xi_0-\xi_1)\begin{bmatrix} \mathcal{F}_0(1/\xi_0)(1/\xi_0) & 0 \\ 0 & -\mathcal{F}_1(1/\xi_1)(1/\xi_1) \end{bmatrix}. \tag{27}$$

Here the choice of $\alpha_0, \alpha_1 \in \mathbf{R}$ embody the two-degrees of freedom in the range of possible R_ψ.

Using the same ideas, but instead employing the orthonormal basis (8), then using Lemma 4.1 it is straightforward to see using the same reasoning that all possible R_ϕ can be interpreted as a perturbation from the identity matrix

$$R_\phi = (\beta_0 + (\xi_0+\xi_1)\beta_1)I + \beta_1(\xi_0-\xi_1)\begin{bmatrix} 1 & K \\ K & -1 \end{bmatrix} \tag{28}$$

where

$$K \triangleq A^0_{0,1} = \frac{\sqrt{(1-\xi_0^2)(1-\xi_1^2)}}{\xi_0-\xi_1}. \tag{29}$$

Again, the choice of the real variables $\beta_0, \beta_1 \in \mathbf{R}$ provides the two degrees of freedom in the assignment of R_ϕ. Therefore, since by (28) the matrix R_ϕ starts, for white input, at a perfectly conditioned matrix and then (as Φ_u becomes coloured) moves in a sub-manifold of 2×2 symmetric matrices, while at the same time R_ψ starts at the imperfectly conditioned matrix Σ and also moves only in a sub-manifold, which by the argument in §3 does not contain a perfectly conditioned matrix, it seems reasonable to suspect that the matrix R_ϕ might be better conditioned than R_ψ for *any* coloured input.

In order to further investigate this, it is necessary to be more precise as to how the eigenvalues of R_ϕ and R_ψ depend on the choice of Φ_u, and for this purpose the following result will prove useful.

Theorem 5.1. *Let $A = [a_{ij}]$ be an $n \times n$ real symmetric matrix. For a fixed index i let α and β be positive numbers satisfying*

$$\alpha\beta \geq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|^2$$

Then the interval $[a_{ii}-\alpha, a_{ii}+\beta]$ contains at least one eigenvalue of A.

Proof. See [1]. □

This result is employed in this paper instead of the similar and more widely known Geršgorin disc Theorem [9] since the latter can only assert the existence of bounds lying in a region if that region is disjoint from certain others. Theorem 5.1 clearly avoids this restriction.

Application of Theorem 5.1 then allows the two eigenvalues λ_1 and λ_2 of R_ψ given by (27) and R_ϕ given by (28) to be bounded as

$$\lambda(R_\psi) \in \left(\frac{(\alpha_0+2\alpha_1\xi_0)}{1-\xi_0^2} \pm \Delta_\psi\right) \cup \left(\frac{(\alpha_0+2\alpha_1\xi_1)}{1-\xi_1^2} \pm \Delta_\psi\right), \qquad \Delta_\psi \triangleq \frac{\alpha_0+\alpha_1(\xi_0+\xi_1)}{1-\xi_0\xi_1} \tag{30}$$

and

$$\lambda(R_\phi) \in ((\beta_0 + 2\xi_0\beta_1) \pm \Delta_\phi) \cup ((\beta_0 + 2\xi_1\beta_1) \pm \Delta_\phi)\,, \qquad \Delta_\phi \triangleq \beta_1\sqrt{(1-\xi_0^2)(1-\xi_1^2)}. \quad (31)$$

where the notation $(x \pm y)$ is meant to denote the open interval $(x-y, x+y)$.

These bounds illustrate an inherent numerical robustness of the orthonormal form for *any* input spectral density. Specifically, (31) shows the eigenvalues of R_ϕ to be in regions centred at $\beta_0 + 2\xi_0\beta_1$ and $\beta_0 + 2\xi_1\beta_1$ and bounded from these centres by a distance Δ_ϕ. But these centres are of the same form as those pertaining to $\lambda(R_\psi)$ save that the centres pertaining to $\lambda(R_\psi)$ are divided by $1-\xi_0^2$ and $1-\xi_1^2$. This latter feature will, particularly if one of ξ_0 or ξ_1 are near 1 and the other isn't, tend to make the centres of the eigenvalue bound regions very different.

Furthermore, the bound $\Delta_\phi = \beta_1\sqrt{(1-\xi_0^2)(1-\xi_1^2)}$ is forced to be small (regardless of Φ_u) if any one of the poles ξ_0 or ξ_1 to be near 1, while the bound Δ_ψ cannot be forced (by choice of ξ_0 and ξ_1) to be small in a way that is insensitive to the Φ_u. Therefore, the numerical conditioning of R_ϕ shows an inherent robustness to the particular Φ_u defining it.

6 Higher Dimensions again

Having argued for the specific $p = 2$ dimensional case that the superior numerical conditioning advantage of the orthonormal model structure (5) is a property that is robust across a range of coloured spectral densities Φ_u, this section extends the argument to arbitrary dimension. Central to this is the following result.

Theorem 6.1. *The eigenvalues $\{\lambda_0, \lambda_1, \cdots, \lambda_{p-1}\}$ of R_ϕ are contained in regions $\Delta_0, \Delta_1, \cdots, \Delta_{p-1}$ defined by*

$$\Delta_m \triangleq \{x \in \mathbf{R} : |x - 2\mathrm{Re}\,\varphi(\xi_m)| \le \alpha_m\}$$

where

$$\alpha_m^2 \triangleq \sum_{\substack{n=0\\ n\neq m}}^{p-1} \left(\sum_{i=m}^{n-1} |A^i_{m,n}||\varphi(\overline{\xi}_i) - \varphi(\overline{\xi}_{i+1})| \right)^2 .$$

Proof. By Theorem 5.1 the regions $\{\Delta_m\}$ are provided as being

$$\Delta_m = \{x \in \mathbf{R} : |x - [R_\phi]_{m,m}| \le \alpha_m\}$$

where

$$\alpha_m^2 \ge \sum_{\substack{n=0\\ n\neq m}}^{p-1} |[R_\phi]_{m,n}|^2 .$$

But by Lemma 4.2

$$[R_\phi]_{m,m} = \varphi(\xi_m) + \varphi(\overline{\xi}_m) = 2\mathrm{Re}\,\varphi(\xi_m)$$

and also by the same lemma

$$\begin{aligned}\sum_{\substack{n=0\\n\neq m}}^{p-1}|[R_\phi]_{m,n}|^2 &= \sum_{\substack{n=0\\n\neq m}}^{p-1}\left|\sum_{i=m}^{n}A^i_{m,n}\varphi(\overline{\xi}_i)\right|^2\\ &\leq \sum_{\substack{n=0\\n\neq m}}^{p-1}\left(\sum_{i=m}^{n-1}|A^i_{m,n}||\varphi(\overline{\xi}_i)-\varphi(\overline{\xi}_{i+1})|\right)^2\end{aligned}$$

where in progressing to the last line, the fact that

$$\sum_{i=m}^{n}A^i_{m,n}=0$$

was employed. □

Note that this theorem provides a tight characterisation in the sense that for white input, $\varphi(\xi_k) = c_0/2$ a constant, in which case the theorem provides the eigenvalues as being all at $\lambda_k = c_0$ with tolerance $\alpha_k = 0$.

However, more generally the theorem provides further indication of the general robustness of the condition number of R_ϕ. Specifically, if φ is smooth and the pole locations $\{\xi_k\}$ are chosen to be relatively 'clustered' around a common point, then this will imply that the terms $|\varphi(\overline{\xi}_i) - \varphi(\overline{\xi}_{i+1})|$ will be small, and hence via Theorem 6.1 the bounds α_m on the eigenvalue locations $\{2\mathrm{Re}\,\varphi(\xi_m)\}$ will be tight, and so the true eigenvalues should be very near to the locations $\{2\mathrm{Re}\varphi(\xi_m)\}$ which again if $\varphi(z)$ is smooth, will be relatively tightly constrained.

7 Conditions for Numerical Superiority

The most desirable result that a study such as this could produce would be one that precisely formulated the necessary and sufficient conditions on Φ_u and $\{\xi_0, \cdots, \xi_{p-1}\}$ such that the numerical conditioning of R_ϕ was superior to that of R_ψ. Unfortunately, this appears to be an extremely difficult question, mainly due to the very complicated manner in which the condition number of a matrix depends on the elements of that matrix.

Nevertheless, the purpose of this section is to at least establish sufficient conditions for when superiority exists, although because of the involved nature of the question, it is only answered for the limited case of dimension $p = 2$.

In order to proceed with this analysis of the numerical superiority (or not) of one model structure over another, it turns out to be better to avoid consideration of the condition number $\kappa(R)$ of a matrix R directly, but instead to consider a new function $f(R)$ of a matrix R which is monotonic in condition number $\kappa(R)$ and which is defined as

$$f(R) \triangleq \left(\frac{\kappa(R)-1}{\kappa(R)+1}\right)^2 = \left(\frac{\lambda_{\max}(R)/\lambda_{\min}(R)-1}{\lambda_{\max}(R)/\lambda_{\min}(R)+1}\right)^2 = \left(\frac{\lambda_{\max}(R)-\lambda_{\min}(R)}{\lambda_{\max}(R)+\lambda_{\min}(R)}\right)^2.$$

Using this idea, it is possible to establish the following result on the general superiority of the orthonormal structure from a numerical conditioning perspective.

Theorem 7.1. *For the two dimensional case of $p = 2$, consider R_ϕ defined by (12) and associated with the orthonormal model structure (5) and R_ψ defined by (11) with the $\{\mathcal{F}_k(q)\}$ defined by (22). Then for $\xi_0, \xi_1 \in \mathbf{R}^+$*

$$\kappa(R_\phi) \leq \kappa(R_\psi)$$

provided that Φ_u is such that the associated $\varphi(z)$ satisfies

$$\frac{\varphi(\xi_0) - \varphi(\xi_1)}{\xi_0 - \xi_1} > 0, \qquad \frac{\xi_0\varphi(\xi_1) - \xi_1\varphi(\xi_0)}{\xi_0 - \xi_1} > 0. \tag{32}$$

Proof. With the definition $C \triangleq (\xi_0 - \xi_1)^{-1}(1 - \xi_0\xi_1)^{-1}$, then straightforward (but tedious) algebra provides that

$$R_\psi = C \begin{bmatrix} \dfrac{2\xi_0\varphi(\xi_0)}{1-\xi_0^2} - \dfrac{2\xi_1\varphi(\xi_1)}{1-\xi_1^2} & \left(\dfrac{1+\xi_0^2}{1-\xi_0^2}\right)\varphi(\xi_0) - \left(\dfrac{1+\xi_1^2}{1-\xi_1^2}\right)\varphi(\xi_1) \\ \left(\dfrac{1+\xi_0^2}{1-\xi_0^2}\right)\varphi(\xi_0) - \left(\dfrac{1+\xi_1^2}{1-\xi_1^2}\right)\varphi(\xi_1) & \dfrac{2\xi_0\varphi(\xi_0)}{1-\xi_0^2} - \dfrac{2\xi_1\varphi(\xi_1)}{1-\xi_1^2} \end{bmatrix}.$$

Also, using Lemma 4.2, and with K given by (29) then R_ϕ may be expressed as

$$R_\phi = \begin{bmatrix} 2\varphi(\xi_0) & K[\varphi(\xi_0) - \varphi(\xi_1)] \\ K[\varphi(\xi_0) - \varphi(\xi_1)] & 2\varphi(\xi_1) \end{bmatrix}.$$

As well, note that for a 2×2 symmetric matrix A of the form

$$A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$$

then the function $f(A)$ may be calculated as

$$f(A) = \frac{(a-c)^2 + 4b^2}{(a+c)^2}.$$

In this case the calculation of $f(R_\psi)$ and $f(R_\phi)$ become

$$f(R_\psi) = \frac{(1+\xi_0\xi_1)^2\left[(1-\xi_0\xi_1)\left(\dfrac{\varphi(\xi_0)-\varphi(\xi_1)}{\xi_0-\xi_1}\right) + \left(\dfrac{\xi_0+\xi_1}{1+\xi_0\xi_1}\right)[\varphi(\xi_0)+\varphi(\xi_1)]\right]^2}{\left[[\varphi(\xi_0)+\varphi(\xi_1)] - \left(\dfrac{\xi_0\varphi(\xi_1)-\xi_1\varphi(\xi_0)}{\xi_0-\xi_1}\right)(1-\xi_0\xi_1)\right]^2}$$

and

$$f(R_\phi) = \frac{\left(\dfrac{\varphi(\xi_1)-\varphi(\xi_1)}{\xi_0-\xi_1}\right)^2(1-\xi_0\xi_1)^2}{[\varphi(\xi_0)+\varphi(\xi_1)]^2}.$$

Now, by assumption $\xi_0, \xi_1 \in \mathbf{R}^+$, so the numerator term of the $f(R_\psi)$ term is clearly greater than that of the $f(R_\phi)$ term if the first condition in (32) is satisfied and the denominator term of the $f(R_\psi)$ term is clearly smaller than that of the $f(R_\phi)$ term if the second condition in (32) is satisfied so that in this case $\kappa(R_\phi) \leq \kappa(R_\psi)$ is guaranteed provided the conditions (32) are met. □

The most important question now is how large the class of possible Φ_u is that satisfy the sufficient conditions (32). For the purpose of analysing this, it is expedient to use the representation (25) in which case condition (32) becomes

$$0 < \frac{\varphi(\xi_0) - \varphi(\xi_1)}{\xi_0 - \xi_1} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{(1 + \xi_0\xi_1)\cos\omega - (\xi_0 + \xi_1)}{|1 - \xi_0 e^{j\omega}|^2 |1 - \xi_1 e^{j\omega}|^2} \Phi_u(\omega)\, \mathrm{d}\omega \tag{33}$$

and similarly, after some algebra

$$0 < \frac{\xi_0\varphi(\xi_1) - \xi_1\varphi(\xi_0)}{\xi_0 - \xi_1} = \frac{1}{4\pi} \int_{-\pi}^{\pi} \frac{(1 - \xi_0^2\xi_1^2) + (\xi_0 + \xi_1)[(\xi_0 + \xi_1) - 2\cos\omega]}{|1 - \xi_0 e^{j\omega}|^2 |1 - \xi_1 e^{j\omega}|^2} \Phi_u(\omega)\, \mathrm{d}\omega \tag{34}$$

The weight functions

$$\chi_1(\omega) \triangleq \frac{(1 + \xi_0\xi_1)\cos\omega - (\xi_0 + \xi_1)}{|1 - \xi_0 e^{j\omega}|^2 |1 - \xi_1 e^{j\omega}|^2} \tag{35}$$

$$\chi_2(\omega) \triangleq \frac{(1 - \xi_0^2\xi_1^2) + (\xi_0 + \xi_1)[(\xi_0 + \xi_1) - 2\cos\omega]}{|1 - \xi_0 e^{j\omega}|^2 |1 - \xi_1 e^{j\omega}|^2} \tag{36}$$

appearing in these integral characterisations of (32) are plotted for the case of $\xi_0 = 0.5, \xi_1 = 0.6$ in the left diagram of figure 1. The weight function χ_1 being the solid line, and the weight χ_2 being the dash-dot line. These weight functions χ_1 and χ_2 clearly concentrate attention around $\omega = 0$, and

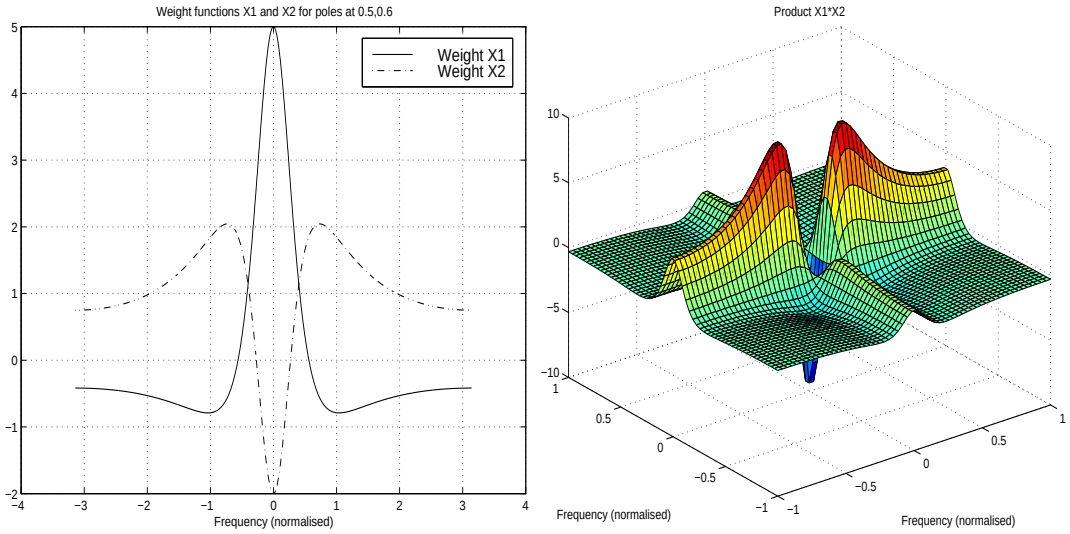

Figure 1: *The left figure shows the weight functions χ_1 and χ_2 defined in (35) and (36) for poles at $\xi_0 = 0.5, \xi_1 = 0.6$. The right figure shows the product of these weights (which is what is important in (37)) for poles at $\xi_0 = 0.8, \xi_1 = 0.95$.*

in such a way that any Φ_u of general low pass nature will, when weighted by them and integrated as in (33) and (34), generally produce a positive result, and hence satisfy the necessary conditions on Theorem 7.1.

To further emphasise this, it is at least clear from (33) and the plot of $\chi_1(\omega)$ that in general any $\Phi_u(\omega)$ that decays as ω tends to π will be such as to satisfy

$$\frac{\varphi(\xi_0) - \varphi(\xi_1)}{\xi_0 - \xi_1} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \chi_1(\omega)\Phi_u(\omega)\, \mathrm{d}\omega > 0.$$

What may not be so clear at first inspection is whether this same class of 'low-pass' $\Phi_u(\omega)$ also lead to the second necessary condition of Theorem 7.1 being satisfied. Namely $(\xi_0\varphi(\xi_1) - \xi_1\varphi(\xi_0))/(\xi_0 - \xi_1) > 0$. This can be clarified by examining the positive sign definiteness of the product

$$\left(\frac{\xi_0\varphi(\xi_1) - \xi_1\varphi(\xi_0)}{\xi_0 - \xi_1}\right)\left(\frac{\varphi(\xi_0) - \varphi(\xi_1)}{\xi_0 - \xi_1}\right) = \frac{1}{8\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\chi_1(\omega)\chi_2(\sigma)\Phi_u(\omega)\Phi_u(\sigma)\,\mathrm{d}\omega\,\mathrm{d}\sigma. \quad (37)$$

The two-dimensional 'kernel' $\chi_1(\omega)\chi_2(\sigma)$ is plotted, again for the case of $\xi_0 = 0.5, \xi_1 = 0.6$ in the right hand diagram of figure 1. Clearly, the bulk of it over all values of ω and σ is positive, and consideration of it indicates that the only way that the product (37) can be negative is if $\Phi_u(\omega)$ is very strongly concentrated around $\omega = 0$. Specifically, the low-pass nature of Φ_u would need to imply a roll-off at round 5% of the sampling frequency or, put another way, the sampling rate would need to be around ten times larger than the minimum Nyquist rate implied by the bandwidth of Φ_u.

The conclusion therefore is, at least in the specific p=2 dimensional case, that R_ϕ has smaller condition number than R_ψ associated with $\mathcal{F}_0, \mathcal{F}_1$ given by the simple form (22) over a very wide range of input spectra Φ_u.

To examine this even more closely, specific classes of parameterised $\Phi_u(\omega)$ may be considered. For example, for the particularly simple class of Φ_u that have a spectral factorisation $\Phi_u(z) = H(z)H(1/z)$ of the form

$$H(z) = 1 + az \quad (38)$$

then

$$\varphi(z) = \frac{c_0}{2} + c_1 z = \frac{1 + a^2}{2} + az$$

so that the two conditions in (32) becomes $c_1 = a > 0$ $c_0 = 1 + a^2 > 0$ respectively. Therefore, $\kappa(R_\psi) > \kappa(R_\phi)$ for any input with spectral factor of the form (38) with $a > 0$. This is not a heavy restriction, since if $a < 0$, this implies that $\{u_t\}$ is differenced white noise, and hence $\Phi_u(e^{j\omega})$ is of the form shown in the left hand diagram of figure 2 and increasing at the folding frequency $\omega = \pi$, which indicates the samples $\{u_t\}$ have been taken too slowly in that aliasing of the underlying continuous time signal is occurring.

Now consider the case of the spectral factor $H(z)$ being second order and of the form

$$H(z) = (1 + az)(1 + bz)$$

so that

$$\varphi(z) = \frac{c_0}{2} + c_1 z + c_2 z^2 = \frac{1 + (a + b)^2 + a^2b^2}{2} + (a + b)(1 + ab)z + abz^2.$$

It seems impossible to analytically derive conditions on $a, b \in \mathbf{R}^+$ such that the conditions (32) will be satisfied, but numerical experiment derives that a, b in the shaded region shown in the right hand diagram of figure 2 do the job so that, for example, $\kappa(R_\psi) \geq \kappa(R_\phi)$ for any $a \geq 0$ and $b \geq 0$.

8 Asymptotic Analysis

As mentioned in the introduction, a key feature of the orthonormal parameterisation (5) is that associated with it is a covariance matrix with numerical conditioning guaranteed by the bounds

$$\min_{\omega\in[-\pi,\pi]} \Phi_u(\omega) \leq \lambda(R_\phi) \leq \max_{\omega\in[-\pi,\pi]} \Phi_u(\omega). \quad (39)$$

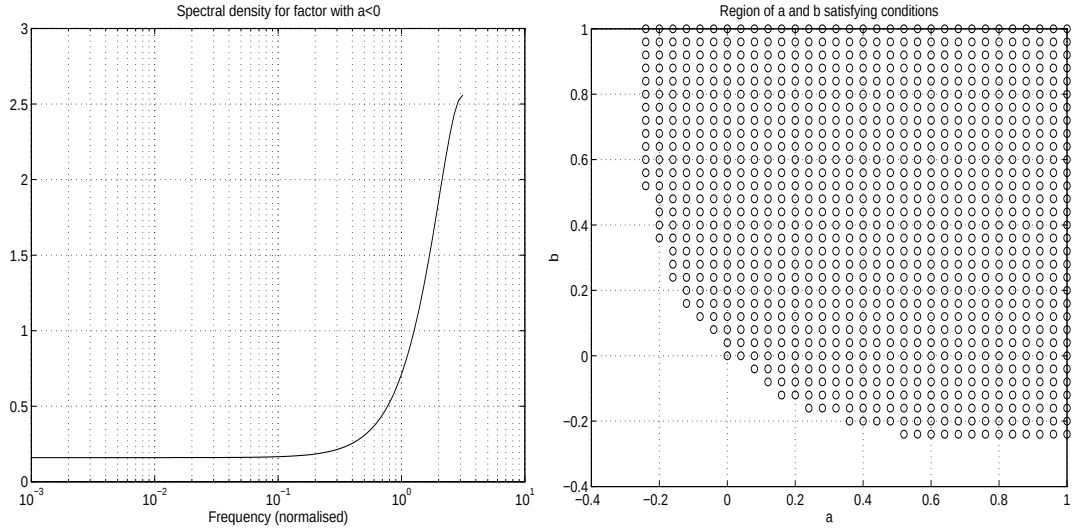

Figure 2: *(Left figure) Illustration of nature of spectral density $\Phi_u(\omega)$ for the case of its spectral factor being first order with a pole $a < 0$. (Right figure) Region of a and b in second order $H(z)$ guaranteeing $\kappa(R_\psi) > \kappa(R_\phi)$ for all $\xi_0, \xi_1 \in \mathbf{R}^+$*

A natural question to consider is how tight these bounds are. In [13], this was addressed by a strategy of analysis that is asymptotic in p. Specifically, define $M_\phi \triangleq \lim_{p\to\infty} R_\phi$. In this case, M_ϕ is an operator $\ell_2 \to \ell_2$, so that the eigenvalues of the finite dimensional matrix R_ϕ, generalise to the continuous spectrum $\lambda(M_\phi)$ which itself is defined as [4]

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

This spectrum can be characterised as follows.

Lemma 8.1. *Suppose that*

$$\sum_{k=0}^{\infty}(1 - |\xi_k|) = \infty.$$

Then

$$\lambda(M_\phi) = \text{Range}\{\Phi_u(\omega)\}.$$

Proof. See [13]. □

This provides evidence, that at least for large p (when the issue of numerical conditioning is most important), that the bounds (39) are in fact tight, and therefore

$$\kappa(R_\phi) \approx \frac{\max_\omega \Phi_u(\omega)}{\min_\omega \Phi_u(\omega)} \tag{40}$$

might be expected to be a reasonable approximation.

Of course, what would also be desirable is a similar approximation for R_ψ, and of course this will depend on the nature of the definition of the $\{\mathcal{F}_k(q)\}$. One particularly natural definition is that of (22) extended to arbitrary dimension p as

$$F_k(q) = \frac{z^k}{D_p(z)}, \qquad D_p(z) = \prod_{\ell=0}^{p-1}(z - \xi_\ell) \tag{41}$$

for $k = 0, 1, \cdots, p-1$ and $\{\xi_0, \cdots, \xi_{p-1}\} \in \mathbf{D}$ the fixed pole choices. This case is considered important, since possibly the most straightforward way of realising a fixed-pole estimate $G(q, \widehat{\beta})$ as originally defined in (2) of §1 would be to simply use pre-existing software for estimating FIR model structures, but after having pre-filtering the input sequence $\{u_t\}$ with the all-pole filter $1/D_p(q)$. This is identical to using the general model structure (1) with the $\{\mathcal{F}_k(q)\}$ choice of (41) above, with estimated FIR coefficients then simply being the numerator coefficient estimates $\{\beta_0, \cdots, \beta_{p-1}\}$.

Fortunately, for this common structure, it is also possible to develop an approximation of the condition number $\kappa(R_\psi)$ via the following asymptotic result which is a direct corollary of Theorem 8.1.

Corollary 8.1. *Consider the choice for the $\{\mathcal{F}_k(q)\}$ defining R_ψ via (3) and (11) given in (41). Suppose that only a finite number of the poles $\{\xi_k\}$ are chosen away from the origin so that*

$$D(\omega) \triangleq \lim_{p\to\infty} \prod_{\ell=0}^{p-1} |e^{j\omega} - \xi_\ell|^2 \tag{42}$$

exits. Define, in a manner analogous to that pertaining to Lemma 8.1, the operator $M_\psi : \ell_2 \to \ell_2$ as

$$M_\psi \triangleq \lim_{p\to\infty} R_\psi.$$

Then

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

Proof. Note that for the choice of $\{\mathcal{F}_k(q)\}$ formulated in (41), then by the definition (3), (9), (11) and (12) R_ψ is the same as a matrix R_ϕ where the orthonormal basis involves the choice of all the poles $\{\xi_k\}$ chosen at the origin, and the input spectrum $\Phi_u(\omega)$ is changed according to $\Phi_u(\omega) \mapsto \Phi_u(\omega)/|D_p(e^{j\omega})|^2$. By the assumptions on the pole locations, this latter quantity converges with increasing p to $\Phi_u(\omega)/|D(e^{j\omega})|^2$, so applying Lemma 8.1, which is invariant to the particular choice of the pole location, provides the result. □

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

$$\kappa(R_\psi) \approx \frac{\max_\omega \Phi_u(\omega)/|D_p(e^{j\omega})|^2}{\min_\omega \Phi_u(\omega)/|D_p(e^{j\omega})|^2}. \tag{43}$$

Now, considering that $|D_p(e^{j\omega})|^2 = \prod_{\ell=0}^{p-1}|e^{j\omega} - \xi_\ell|^2$ 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 $\Phi_u/|D_p|^2$ will be much more widely separated than those of Φ_u. The approximations (40) and (43) therefore indicate that estimation with respect to the orthonormal form (5) could be expected to be

much better conditioned than that with respect to the model structure (3) with the simple choice (41) for a very large class of Φ_u - an obvious exception here would be $\Phi_u = |D_p|^2$ for which $R_\psi = I$.

However, this conclusion depends on the accuracy of applying the asymptotically derived approximations (40) and (43) for finite p. In the absence of theoretical analysis, which appears intractable, simulation study can be pursued. Consider p 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_ψ and R_ϕ are shown as solid lines in the left and (respectively) right diagrams in figure (3). The dash-dot lines in these figures are the approximations (40) and (43). Clearly, in this case the approximations are quite accurate, even for what might be considered small p. Note that the minimum eigenvalue of R_ψ is shown only up until $p = 18$ since it was numerically impossible to calculate it for higher p.

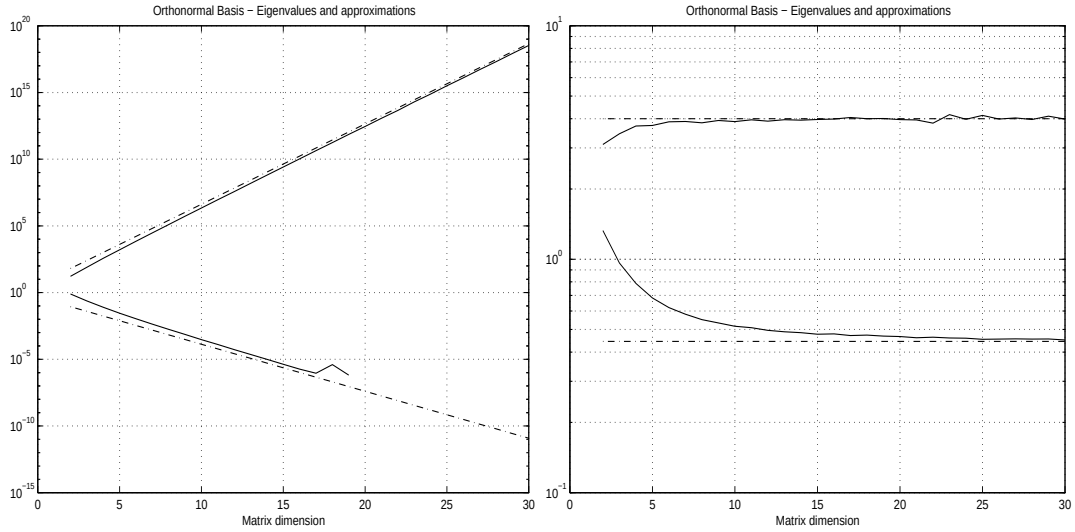

Figure 3: *Solid lines are maximum and minimum eigenvalues of (left figure) R_ψ and (right figure) R_ϕ for a range of dimensions p. The dash dot lines are the approximations (40) and (43).*

Again, this provides evidence that even though model structures (5) 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 improved conditioning over a very wide range of coloured inputs as well.

9 Conclusions

A variety of arguments have been presented to indicate that the condition numbers $\kappa(R_\psi)$ and $\kappa(R_\phi)$, which govern the numerical properties of least squares estimation associated with (respectively) simple 'fixed denominator' model structures and their ortho-normalised forms, are such that $\kappa(R_\psi) \geq \kappa(R_\phi)$ for a very wide class of input spectra Φ_u. While this might be considered somewhat surprising, since it is only designed to occur (by the construction of the 'orthonormal' model structure) for white Φ_u, it is also important since it provides a strong argument for why the extra programming effort should be expended to implement the various orthonormal model structures that have recently been examined in the literature. This analysis is made in counter-argument to the charge (as illustrated in the introduction), that a change of model structure is not the same as a change of estimation method - equivalent structures provide identical estimates, but modulo the numerical issues considered here.

References

[1] E. BARNES AND A. HOFFMAN, *On bounds for eigenvalues of real symmetric matrices,* Linear Algebra and its Applications, 40 (1981), pp. 217–223.

[2] P. BODIN, T. OLIVEIRA E SILVA, AND B. WAHLBERG, *On the construction of orthonormal basis functions for system identification,* in Proceedings of the 13'th IFAC World Congress, San Francisco, 1996, pp. 291–296.

[3] J. BOKOR AND F. SCHIPP, *Approximate identification in Laguerre and Kautz bases,* Automatica, 34 (1998), pp. 463–468.

[4] A. BÖTTCHER AND B. SILBERMANN, *Invertibility and Asymptotics of Toeplitz Matrices,* Akadamie–Verlag, Berlin, 1983.

[5] G. DAVIDSON AND D. FALCONER, *Reduced complexity echo cancellation using orthonormal functions,* IEEE Transactions on Circuits and Systems, 38 (1991), pp. 20–28.

[6] N.F. DUDLEY WARD AND J. PARTINGTON, *Robust identification in the disc algebra using rational wavelets and orthonormal basis functions,* International Journal of Control, 64 (1996), pp. 409–423.

[7] G. GOLUB AND C. V. LOAN, *Matrix Computations,* Johns Hopkins University Press, 1989.

[8] P. HEUBERGER, P.M.J. VAN DEN HOF, AND O. BOSGRA, *A generalized orthonormal basis for linear dynamical systems,* IEEE Transactions on Automatic Control, AC-40 (1995), pp. 451–465.

[9] R.A. HORN AND C.R. JOHNSON, *Matrix Analysis,* Cambridge University Press, Cambridge, 1985.

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

[11] L. LJUNG AND S. GUNNARSSON, *Adaptation and tracking in system identification–a survey,* Automatica, 26 (1990), pp. 7–21.

[12] B. NINNESS AND F. GUSTAFSSON, *A unifying construction of orthonormal bases for system identification,* IEEE Transactions on Automatic Control, 42 (1997), pp. 515–521.

[13] B. NINNESS, H. HJALMARSSON, AND F. GUSTAFSSON, *Generalised Fourier and Toeplitz results for rational orthonormal bases,* SIAM Journal on Control and Optimization, 37 (1999), pp. 429–460.

[14] T. OLIVEIRA E SILVA, *Optimality conditions for truncated laguerre networks,* IEEE Transactions on Signal Processing, 42 (1994), pp. 2528–2530.

[15] T. OLIVEIRA E SILVA, *Stationary conditions for the L^2 error surface of the generalized orthonormal basis functions lattice filter,* Signal Processing, 56 (1997), pp. 233–253.

[16] P.M.J. VAN DEN HOF, P.S.C. HEUBERGER, AND J. BOKOR, *System identification with generalized orthonormal basis functions,* Automatica, 31 (1995), pp. 1821–1834.

[17] S.PILLAI AND T.SHIM, *Spectrum Estimation and System Identification,* Springer–Verlag, 1993.

[18] Z. SZABO, J. BOKOR, AND F. SCHIPP, *Identification of rational approximate models in H_∞ using generalised orthonormal basis*, IEEE Transactions on Automatic Control, 44 (1999), pp. 153–158.

[19] B. WAHLBERG, *Identification of resonant systems using Kautz filters*, in Proceedings of the 30th Conference on Decision and Control, 1991, pp. 2005–2010.

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

[21] B. WAHLBERG, *System identification using Kautz models*, IEEE Transactions on Automatic Control, AC-39 (1994), pp. 1276–1282.

[22] B. WAHLBERG AND P. MÄKILÄ, *On approximation of stable linear dynamical systmes using Laguerre and Kautz functions*, Automatica, 32 (1996), pp. 693–708.

[23] G. A. WILLIAMSON, *Tracking random walk systems with vector space adaptive filters*, IEEE Transactions on Circuits and Systems-II: Analog and Digital Signal Processing, 42 (1995), pp. 543–547.

[24] G. A. WILLIAMSON AND C. RICHARD JOHNSON JR., *Some effects of parameterization change in system identification*, in Proceedings of the American Control Conference, 1992, pp. 1268–1269.

[25] G. A. WILLIAMSON AND S. ZIMMERMANN, *Globally convergent adaptive IIR filters based on fixed pole locations*, IEEE Trans. Signal Processing, 44 (1996), pp. 1418–1427.