OPTIMAL Q-MARKOV COVER FOR FINITE PRECISION IMPLEMENTATION

By

Darrell Williamson
Australian National University Canberra
ACT 2601, Australia

and

Robert E. Skelton
Purdue University
West Lafayette, Indiana

ABSTRACT

The existing q-Markov COVER realization theory does not take into account the problems of arithmetic errors due to both the quantization of states and coefficients of the reduced order model. All q-Markov COVERs allow some freedom in the choice of parameters. In this talk, we exploit this freedom in the existing theory to optimize the models with respect to these finite wordlength effects.
Optimal q-Markov Cover for
Finite Precision Implementation

Darrell Williamson* and Robert E. Skelton**

Abstract
The existing q-Markov COVER realization theory does not take into account
the problems of arithmetic errors due to both the quantization of states and
coefficients of the reduced order model. All q-Markov COVERs allow some
freedom in the choice of parameters. In this paper we exploit this freedom in the
existing theory to optimize the models with respect to these finite wordlength
effects.

*Dept. of Systems Engineering, Research School of Physical Sciences, Australian National
University Canberra, ACT 2601, Australia
**School of Aeronautics and Astronautics, Purdue University, West Lafayette, IN 47907, U.S.A.
Introduction

An asymptotically stable system can be characterized in terms of its impulse response sequence (Markov parameters) and its output covariance sequence (covariance parameters) due to a zero mean white noise input process. A general approach has been developed [3] for realizing a system which matches q Markov parameters and q covariance parameters. Such a system is referred to as a q-Markov COVER, and q-Markov COVERs may be generated from output data [3,4] or from higher order models [5,6]. The Markov and covariance parameters are not independent and consequently the q-Markov COVER is not unique. In particular, all q-Markov COVERs are not related by state space similarity transformations [4]. In this paper we shall exploit the remaining degrees of freedom to optimize the q-Markov COVER realization with respect to an aspect of its finite wordlength realization.

Specifically, when digital controllers are to be implemented, both the controller coefficients and the controller states must be represented in finite wordlength precision. This finite wordlength (FWL) representation (or quantization) causes inaccuracies in the response when compared to the ideal (i.e. infinite precision) behaviour. Effects of quantization on the controller are increased noise at the output due to internal state quantization, and errors in time and frequency response characteristics due to coefficient errors.

In digital filter design, the FWL effects are known to be most significant when the poles of the filter are very close to the unit circle [12]. In particular, narrow band filters have all these poles near $z = 1 \pm jo$. For digital control, the zero-order-hold equivalent of a continuous time model (or controller) with a pole at $\lambda$ will have a discrete pole at $\exp(\lambda T)$. Hence for fast sampling and/or low damping of the continuous models, the discrete model will behave like a narrow band filter. The synthesis of optimal digital controllers with respect to arithmetic quantization noise is an important consideration in design especially for continuous time systems operating under a fast sampling rate [9,10]. The effects of quantization depend highly on the structure of the controller. This paper seeks to reduce these errors in the synthesis of q-Markov COVERs.

1. Discrete q-Markov COVER

Consider the asymptotically stable nominal discrete system
\[
x(k+1) = Ax(k) + Bu(k) \quad x(k) \in \mathbb{R}^n, u(k) \in \mathbb{R}^m \\
y(k) = Cx(k) \quad y(k) \in \mathbb{R}^n
\]

(1.1)

where \{u(k)\} is a zero mean process with unit intensity \( E\{u(k)u^*(j)\} = \delta_{ij} \) and \( E\{x(k)u^*(j)\} = 0 \) for \( k \geq j \). The Markov parameters \( M_i \) and covariance parameters \( R_j \) of (1.1) are defined by

\[
M_i \triangleq C A^i B \quad R_j \triangleq C A^j X C^* \quad j \geq 0, R_j \triangleq C A^j X C^* \quad j \leq 0
\]

(1.2)

where the state covariance matrix \( X \) satisfies the Lyapunov Equation

\[
X = AXA^* + BB^*
\]

(1.3)

These parameters \( M_i \) and \( R_j \) appear as coefficients in the expansion of the transfer function \( H(z) \) and power spectral density \( H(z)H^*(z) \); that is

\[
H(z) = C(zI-A)^{-1}B = \sum_{i=0}^{\infty} M_i z^{-i-1}; \quad H(z)H^*(z) = \sum_{j=-\infty}^{\infty} R_j z^{-j}
\]

We suppose that as data we are given the first \( q \)-Markov and first \( q \)-covariance parameters \( \{M_i, R_j; i = 0, 1, ..., q-1\} \) of an asymptotically stable system from which we construct the two data matrices

\[
D_q \triangleq R_q - M_q M^*_q \in \mathbb{R}^{n_q \times n_q}
\]

\[
\overline{D}_q \triangleq R_q - \overline{M}_q \overline{M}^*_q \in \mathbb{R}^{n_q \times n_q}
\]

(1.4a)

where \( R_q, M_q \) and \( \overline{M}_q \) are the Toeplitz matrices of the data as defined by

\[
R_q \triangleq \begin{bmatrix}
R_0 & R_1^* & ... & R_{q-1}^* \\
R_1 & R_0 & ... & R_{q-2}^* \\
\vdots & \vdots & \ddots & \vdots \\
R_{q-2} & \cdots & \cdots & R_0
\end{bmatrix}
\]

(1.4b)

\[
M_q \triangleq \begin{bmatrix}
0 & 0 & ... & 0 & 0 \\
M_0 & 0 & ... & 0 & 0 \\
M_1 & M_0 & ... & \cdot & \cdot \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
M_{q-2} & M_{q-3} & \cdots & M_0 & 0
\end{bmatrix}, \quad \overline{M}_q \triangleq \begin{bmatrix}
M_0 & 0 & ... & 0 \\
M_1 & M_3 & ... & 0 \\
\vdots & \vdots & \ddots & \vdots \\
M_{q-2} & \cdot & \cdots & 0 \\
M_{q-1} & M_{q-2} & \cdots & M_0
\end{bmatrix}
\]

The first data matrix \( D_q \) in (1.4a) is Hermitian and it is shown in [3-4] to be
positive semidefinite. Hence we can obtain a (nonunique) full rank factorization

\[ D_q = P_q P_q^*; \quad P_q e^{R_{n_q \times n_q}}, \]  

(1.5a)

where

\[ r_q = \text{rank} (D_q) = \text{rank}(P_q) \leq n_q \]  

(1.5b)

If we partition \( P_q \) according to

\[ P_q = [E_q F_q]; \quad E_q e^{R_{n_q \times n_q}}, \quad F_q e^{R_{(q-1)n_q \times n_q}}, \]  

(1.6)

then it follows that the second data matrix \( \tilde{D}_q \) can be factored as

\[ \tilde{D}_q = \tilde{P}_q \tilde{P}_q^*; \quad \tilde{P}_q e^{\tilde{R}_{n_q \times n_q}} \]  

(1.7)

where

\[ \tilde{P}_q = [F_q G_q^*]; \quad G_q e^{R_{n_q \times n_q}} \]  

(1.8)

for some \( G_q \) (to be determined). The following result has been established.

Theorem 1.1 [3]

Given the \( q \) Markov parameters \( \{M_i; \ i = 0,1,...,q-1\} \) and the \( q \) covariance parameters \( \{R_i; \ i = 0,1,...,q-1\} \) and a matrix \( G_q \) in (1.8) such that (1.7) is satisfied, then the realization \( \{A_q, B_q, C_q\} \) of order \( r_q \) defined by

\[ A_q = P_q^* \tilde{P}_q; \quad B_q = P_q^* [M_q^* \cdots M_{q-1}^*]; \quad C_q = E_q \]  

(1.9)

where \( P_q^* \) denotes the Moore-Penrose inverse of \( P \) is a \( q \)-Markov COVER. The corresponding controllability gramian \( X_q \) is given by

\[ X_q = I \]  

(1.10)

Furthermore

\[ P_q = [C_q^* A_q^* C_q^* \cdots (A_q^*)^* C_q^*]^* \]  

(1.11)

This theorem describes a large but not complete class \( C_q \) of \( q \)-Markov COVERs parameterized by \( \{G_q\} \) such that for some \( E_q, F_q \) the data matrices \( D_q, \tilde{D}_q \) satisfy (1.5)-(1.8). Each matrix \( G_q \) will (generally) result in a \( q \)-Markov COVER having a different transfer function. In order to compute the set of all such \( G_q \), observe in (1.5)-(1.8) that

858
Then

$$D_q = \begin{bmatrix} E_q \\ F_q \end{bmatrix} \begin{bmatrix} E_q^* F_q^* \end{bmatrix}.$$  \hfill (1.12a)

Then

$$\tilde{D}_q = \begin{bmatrix} \tilde{D}_{q-1} \\ \tilde{d}_q \\ \tilde{d}_{qq} \end{bmatrix} = \begin{bmatrix} F_q \\ G_q \end{bmatrix} \begin{bmatrix} F_q^* G_q^* \end{bmatrix}$$  \hfill (1.12b)

implies

$$E_q E_q^* = R_q, \quad F_q F_q^* = \tilde{D}_{q-1}, \quad F_q G_q^* = \tilde{d}_q, \quad G_q G_q^* = \tilde{d}_{qq}$$  \hfill (1.13)

Now expand $D_q$ in terms of its singular value decomposition

$$D_q = \left( U_1 U_2 \right) \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U_1^* \\ U_2^* \end{bmatrix}; \quad \Sigma_1 \in R^{n_y \times n_y}.$$  \hfill (1.14)

Then from (1.12a)

$$(E_q^* F_q^*) = \Sigma_1^{1/2} U_1^*$$  \hfill (1.15)

so that $E_q = C_q$ is defined by the first $n_y$ rows and $F_q$ by the last $(q-1)n_y$ rows of $U_1 \Sigma_1^{1/2}$. Define

$$\rho_q \triangleq \text{rank} (F_q).$$  \hfill (1.16a)

Then from (1.15)

$$\rho_q \leq \min (r_q, (q-1)n_y).$$  \hfill (1.16b)

Next, expand $F_q$ in (1.13) in terms of its singular value decomposition. If strict inequality occurs in (1.16b) we have

$$F_q = [U_\alpha \, U_\beta] \begin{bmatrix} \Sigma_q & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_\alpha \\ V_\beta \end{bmatrix}; \quad \Sigma_q \in R^{\rho_q \times \rho_q}.$$  \hfill (1.17)

The Moore-Penrose inverse $F_q^+$ of $F_q$ is then given by

$$F_q^+ = V_\alpha \Sigma_q^{-1} U_\alpha^*$$  \hfill (1.18)

Corollary 1.1
Define

(i) \( G_{q1} \triangleq (F_q^* d_q)^* \in \mathbb{R}^{n \times r_q} \)  

(ii) \( G_{q2} \in \mathbb{R}^{n \times r_q} \) such that \( G_{q2} G_{q2}^* \triangleq d_{q2}^* d_{q2} - d_{q2}^* D_{q2}^+ d_{q2} \)

where

\[
s_q \triangleq \text{rank} \left[ d_{q2} - d_{q2}^* D_{q2}^+ d_{q2} \right]
\]

and

(iii) \( G_{q3} \triangleq V^*_\beta \in \mathbb{R}^{(r_q - \rho_q) \times n_q} \).

Then if strict inequality occurs in (1.16b) the set of all \( G_q \) which satisfy (1.13) are given by

\[
G_q = G_{q1} + G_{q2} U_q G_{q3}
\]

where

\[
U_q \in \mathbb{R}^{r_q \times (r_q - \rho_q)}; \quad s_q \leq r_q - \rho_q \leq n_q
\]

is an arbitrary row unitary matrix (i.e. \( U_q U_q^* = I \)). Furthermore, if the Moore-Penrose \( P_q^+ \) of

\[
P_q = [E_q^* F_q^*]^*
\]

is expressed as

\[
P_q^+ = [\tilde{L}_{11} \quad L_{12}]; \quad \tilde{L}_{11} \in \mathbb{R}^{r_q \times (q - 1)n_q}, \quad L_{12} \in \mathbb{R}^{r_q \times n_q}
\]

then the corresponding state space representation \( \{A_q, B_q, C_q\} \) of the q-Markov COVER is given by

\[
A_q = L_{11} + L_{12} G_q; \quad L_{11} = \tilde{L}_{11} F_q \in \mathbb{R}^{r_q \times r_q}
\]

\[
B_q = P_q^*[M_0^* M_1^* \cdots M_{q-1}^*]; \quad C_q = E_q.
\]

If \( r_q = \rho_q \), then \( G_q = G_{q1} \) is unique.

**Proof:** The expression for \( F_q G_q^* \) in (1.13) implies \( G_q^* \) is of the form

\[
G_q^* = F_q^* d_q + G_{q3} M^*; \quad M \in \mathbb{R}^{n \times (r_q - \rho_q)}
\]

for some \( M \). Then expanding \( G_q G_q^* \) using (1.13) we have
\[
\bar{d}_{qq} = \bar{d}_{q}(F_q^*)^*F_q^*\bar{d}_q + \bar{d}_{q}(F_q^*)^*G_{q3}M^* + MG_{q3}F_q^*\bar{d}_q = MG_{q3}G_{q3}M^*
\]

Also from (1.13) and (1.21)

\[
(F_q^*)^*F_q^* = \bar{D}_{q-1}^+, \quad G_{q3}G_{q3}^* = I; \quad (F_q^*)^*G_{q1}^* = 0
\]

so that

\[
MM^* = \bar{d}_{qq} - \bar{d}_{q}(F_q^*)^*F_q^*\bar{d}_q
\]

Since \(MM^*\) has rank \(s_q\),

\[
s_q = \text{rank}(G_{q2}G_{q2}^*) = r_q - \rho_q
\]

2. Optimal Finite Wordlength q-Markov COVER

A fixed point finite wordlength realization of the ideal (i.e. infinite precision) q-Markov COVER (1.1) shall be referred to as a q-FWL Markov COVER and is described by

\[
\hat{x}(k+1) = \hat{A}Q[\hat{x}(k)] + \hat{B}\hat{u}(k)
\]

\[
\hat{y}(k) = \hat{C}Q[\hat{x}(k)]
\]

(2.1)

\[
Q[\hat{x}(k)] = \hat{x}(k) - e(k)
\]

where \(e(k)\) is the error in computing \(\hat{x}(k)\). The components of the matrices \(\hat{A}, \hat{B}, \hat{C}\) are assumed to have a \(W_0\) bit fractional representation obtained by quantization of the components of \(A, B, C\) in (1.1). The components of \(\hat{x}(k)\) have a \(W+W_0\) bit fractional part while components of \(Q[\hat{x}(k)]\) and \(\hat{u}(k)\) all have a \(W\) bit fractional part. The components of the state residue vector \(e(k)\) has a \(W+W_0\) bit fractional representation in which the most significant \(W\) bits are zero. The LHS and RHS of (2.1) are therefore consistent with respect to their fractional wordlength representation. The number of bits required to represent the integer parts of \(\hat{A}, \hat{B}\) and \(\hat{C}\) depend on the dynamic range of the coefficients. State space structures in which all coefficients are less than unity are therefore advantageous in this regard. The required integer representation of \(Q[\hat{x}(k)]\) will depend on the dynamic range of the input signal \(\hat{u}(k)\). Inadequate dynamic range will result in arithmetic overflow. The accuracy in the computation of \(\hat{x}(k)\) is determined by its fractional wordlength \(W\).

Define the state error vector \(e_x(k)\) and output error vector \(e_y(k)\) by
\[
\varepsilon_x(k) \triangleq \hat{x}(k) - x(k); \quad \varepsilon_y(k) \triangleq \hat{y}(k) - y(k) \tag{2.2}
\]
Then from (1.1), (2.1) and (2.2)
\[
\varepsilon_x(k+1) = A\varepsilon_x(k) - Ae(k) + \Delta AQ[\hat{x}(k)] + \Delta Bu(k) + B\Delta u(k) \tag{2.3}
\]
\[
\varepsilon_y(k) = C\varepsilon_x(k) - Ce(k) + \Delta CQ[\hat{x}(k)]
\]
where

\[
\Delta A = \hat{A} - A; \quad \Delta B = \hat{B} - B; \quad \Delta C = \hat{C} - C
\]
\[
\Delta u(k) = \hat{u}(k) - u(k)
\]

There are five terms which contribute to the output error (i) internal arithmetic errors \(e(k)\) due to state quantization (ii) coefficient errors due to errors \(\Delta A\) in \(A\) (iii) \(\Delta B\) in \(B\) (iv) \(\Delta C\) in \(C\), and (v) input quantization errors \(\Delta u(k)\). Under weak 'sufficiently exciting' conditions on the input \(\{u(k)\}\) it can be shown [6] that if \(Q[\cdot]\) in (2.1) denotes 'roundoff' quantization, then \(\{e(k)\}\) is a zero mean uniform white process with covariance
\[
E\{e(k)e^*(k)\} = \gamma^2 I; \quad \gamma^2 = \frac{1}{12} 2^{-2w}. \tag{2.4}
\]
Similarly \(\{\Delta u(k)\}\) is assumed to be a zero mean white uniform process with
\[
E\{\Delta u(k)\Delta^* u(k)\} = \gamma^2 I \tag{2.5}
\]
We assume that the quantized coefficients \(\hat{A}, \hat{B}, \hat{C}\) are obtained by rounding \(A, B, C\) to \(w_o\) bit fractions. Consequently, all components \(\Delta p\) of the error matrices \(\Delta A, \Delta B, \Delta C\) satisfy
\[
|\Delta p| < \gamma_o; \quad \gamma_o = \frac{1}{2} 2^{-w_o}. \tag{2.6}
\]
For simplicity we normalize the error matrices and define \(\delta A, \delta B, \delta C\) by
\[
\delta A \triangleq \frac{1}{\gamma_o} \Delta A; \quad \delta B \triangleq \frac{1}{\gamma_o} \Delta B; \quad \delta C \triangleq \frac{1}{\gamma_o} \Delta C \tag{2.7}
\]
so that all components \(\delta p\) of \(\delta A, \delta B\) and \(\delta C\) satisfy
\[
|\delta p| < 1. \tag{2.8}
\]
The steady state output error covariance \(Y\) of \(\{\varepsilon_y(k)\}\) is then given by (we assume independence of \(e(k), e(k)\) and \(\hat{x}(k)\)).
where
\[ Y = 
\begin{align*}
\text{CPC}^* + \gamma_2^2 CC^* + \gamma_0^2 (\delta C)(\hat{X} + \gamma^2 I)(\delta C)^* + \gamma_0 \gamma^2 (C(\delta C)^* + (\delta C)C^*),
\end{align*}
\]
(2.9)

and
\[ P = 
\begin{align*}
E \{ e_x(k)e_x^*(k) \} 
= APA^* + \gamma^2 AA^* + \gamma_0^2 (\delta A)(\hat{X} + \gamma^2 I)(\delta A)^* + \gamma_0^2 (\delta B)(\delta B)^* + \gamma^2 BB^*
\end{align*}
\]

For the remainder of this section we assume no coefficient errors (i.e. \( \gamma_0 = 0 \) in (2.9)) and consider only the effects due to finite state wordlength (FSWL). The issue of coefficient error shall be resumed in Section 4.

**Theorem 2.1**

Define the output noise measure
\[ J \triangleq \text{tr}[Y]. \]

Then for \( \gamma_0 = 0 \)
\[ J = \gamma^2 \{ \text{tr}[K] + \text{tr}[B^* KB] \} \]
(2.10)

where
\[ K = A^* KA + C^* C. \]
(2.11)

**Proof:** From (2.9)
\[ Y = \text{CPC}^*; \quad \tilde{P} = \text{APA}^* + \gamma^2 Z = P + \gamma^2 I \]
where
\[ Z = I + BB^*. \]

Now
\[ \tilde{P} = \gamma^2 \sum_{k=0}^{\infty} A^k Z(A^k)^* \]
and
\[ K = \sum_{k=0}^{\infty} (A^k)^* C^* CA^k \]

so that

863
A fixed point q-FSWL Markov COVER corresponding to the (ideal) q-
Markov COVER (1.1) is therefore described by
\[
\hat{x}(k+1) = AQ[\hat{x}(k)] + Bu(k)
\]
\[
\hat{y}(k) = CQ[\hat{x}(k)]
\]
\[
Q[\hat{x}(k)] = \hat{x}(k) - e(k)
\]
The output noise gain (\(\eta_x\)) due to state quantization and the output noise gain
(\(\eta_u\)) due to input quantization are defined by
\[
\eta_x \triangleq \text{tr}[K]; \quad \eta_u \triangleq \text{tr}[B^*KB]
\]
The noise gain \(\eta_x\) generally varies with state space representation whereas \(\eta_u\) is
independent of the coordinate basis. Specifically, consider the q-FSWL Markov
COVER
\[
\hat{x}(k+1) = AQ[\hat{x}(k)] + Bu(k)
\]
\[
\hat{y}(k) = CQ[\hat{x}(k)]
\]
\[
Q[\hat{x}(k)] = \hat{x}(k) - f(k)
\]
where
\[
A = T^{-1}AT, \quad B = T^{-1}B, \quad C = CT
\]
\[ \eta_z = \text{tr}[T^*K]\; \text{and} \; \tilde{\eta}_u = \text{tr}[B^*KB] \] (2.17)

Notice from (2.13) that the noise gain \( \eta_u \) due to input quantization errors is unaffected by a similarity transformation. Conversely the noise gain \( \eta_z \) due to state quantization generally changes with co-ordinate bases. There is no change if \( T \) is unitary. The q-FSWL Markov COVER (2.14) is superior to the q-FSWL Markov COVER (2.12) if

\[ \eta_z < \eta_x. \] (2.18)

However the comparison in (2.18) must be made under the assumption of identical scaling of the states \( \hat{x}(k) \) and \( \hat{z}(k) \). Specifically, equal l2-scaling of gain \( \alpha \) from a zero mean unit intensity white noise input \( \hat{u}(k) \) to the state components \( \hat{x}_j(k) \) of \( \hat{x}(k) \) requires

\[ X_{jj} = \alpha \quad \text{for all } j \] (2.19)

where \( X_{jj} \) denotes the jth diagonal component of the state covariance matrix \( X \) given by (1.3). Equal l2-scaling of gain \( \alpha \) of components of \( \hat{z}(k) \) in (2.14) requires

\[ Z_{jj} = \alpha; \quad Z = AAZ^* + BB^* \] (2.20)

Equality in l2-scaling of representations (2.12) and (2.14) is equivalent to equality in the state dynamic range (i.e. number of bits in the integer representation of states) for a given probability of overflow. We now state a result which is important for establishing l2-scaling.

**Lemma 2.1** [8,9] Suppose \( M = M^* > 0 \) is an nxn matrix. Then a necessary and sufficient condition for the existence of a unitary matrix \( V \) such that

\[ VMV^*_{jj} = \alpha \quad \text{for all } j \]

is

\[ \text{tr}[M] = n\alpha \]

We have shown in Lemma 1.1 that different similarity transformations of an ideal q-Markov COVER corresponds to different factorization of the first data matrix \( D_q \) in (1.5a). Our aim is to optimize this factorization.

**Definition 2.1**
The *Optimal q-FSWL Markov COVER* minimizes the output quantization noise gain \( \eta \) due to state quantization errors; that is

\[
\eta_{\text{opt}} = \min_{T, G_q} \text{tr}[T^* K_q T]; \quad T^* T = \Lambda^{-1}
\]  

subject to the \( l_2 \)-scaling constraint:

\[
\Lambda_{jj} = \alpha \text{ for all } j
\]  

where the observability gramian \( K_q \) satisfies

\[
K_q = A_q^* K_q A_q + C_q^* C_q
\]  

with \( \{A_q, B_q, C_q\} \) defined by (1.22)-(1.25).

\[
\Box
\]

In corollary 1.1 we have shown that all the degrees of freedom available to select \( G_q \) are confined to an arbitrary row unitary matrix \( U_q \). We now show how to optimize \( U_q \).

**Theorem 2.1**

a. The optimal q-FSWL Markov COVER (1.25) is defined by

\[
\eta_{\text{opt}} = r_q^{-1} \min_{U_q} (\text{tr}[K_q^{1/2}])^2
\]  

where \( U_q \in \mathbb{R}^{s \times (r_q - p_q)} \) is an arbitrary row unitary matrix and \( K_q \) satisfies (2.23).

b. The transfer function of the optimal q-FSWL Markov COVER has Hankel singular values given by the eigenvalues of \( K_q \) defined by the minimizing \( U_q \).

c. Suppose \( U_q = U_{q_0} \) is the minimizing solution corresponding to the optimal \( G_q = G_{q_0} \) in (1.22a). Let \( \{A_{q_0}, B_{q_0}, C_{q_0}\} \) be the corresponding state space realization in (1.24). Then the optimal q-FSWL Markov COVER has a (nonunique) state space representation \( \{T_0^{-1} A_{q_0} T_0, T_0^{-1} B_{q_0}, C_{q_0} T_0\} \) where

\[
T_0 = U_{o} \pi_{o} V_{o}^* = U_{o} \pi_{o} V_{o}^*
\]

such that

(i) the unitary matrix \( U_o \) is defined by

\[
865
\]
where
\[ K_{qO} = A_{qO}K_{qO}A_{qO}^* + C_{qO}C_{qO} \quad \text{and} \quad \Sigma_0^2 = \text{diag}\{\sigma_1^2, \sigma_2^2, ..., \sigma_{r_q}^2\} \]
in which \( \{\sigma_j^2\} \) are the optimal Hankel singular values (eigenvalues of \( K_{qO} \)).

(ii)
\[ \pi_0^2 = \frac{1}{\alpha^2 r_q} \left( \sum_{k=1}^{r_q} \sigma_{k_0} \right) \Sigma_0^{-1} \]
and (iii) \( V_0 \) is unitary such that
\[ (V_0 \Sigma_0 V_0^*)_j = \frac{\sum_{k=1}^{r_q} \sigma_{k_0}}{\nu_q} \quad \text{for all } j \]
\[ \eta_{\text{opt}} = \eta_q (\text{optimal}) = \frac{1}{\alpha^2 r_q} \left( \sum_{k=1}^{r_q} \sigma_{k_0} \right)^2 \]

**Proof:** By corollary 1.1 we have for \( G_q \) defined by (1.22) for any row unitary matrix \( U_q \) (of appropriately specified dimensions) that \( G_q \) defines a q-Markov COVER. The corresponding realization \( \{A_q, B_q, C_q\} \) for each such \( U_q \) has identity controllability grammian and observability grammian \( K_q \) defined by (2.23). Now given a particular \( U_q \), apply a similarity transformation
\[ T = U_q \pi_0 V_0^* \]
to the given q-Markov COVER. Then
\[ \text{tr}(T^* K_q T) = \text{tr}(\pi_0^2 U_q^* K_q U_q) \]
and
\[ (T^* T)^{-1} = V_0 \pi_0^{-2} V_0^* \]

By lemma 2.1, the \( l_2 \)-scaling constant can be satisfied for some \( V_0 \) provided \( \text{tr}(\pi_0^{-2}) = n\alpha \). Following Williamson [1, Theorem 4.1] (with a minor modification of the \( l_2 \)-scaling constraint), the optimal performance is given by
\[ \eta_{qopt} = \frac{\left( \sum_{j=1}^{r_q} \sigma_j \right)^2}{\alpha^2 r_q} \]

where \( \{\sigma_j^2\} \) are the eigenvalues of \( K_q \). That is,
\[ \text{tr}(K_q^{1/2}) = \sum_{j=1}^{r_q} \sigma_j \]

The optimal \( q \)-FSWL Markov COVER therefore achieves the minimum in (2.24). The structure of \( U_o, \pi_o, V_o \) in (2.25)-(2.29) follow directly from Williamson [1] (see proof of Theorem 4.1 with \( U = I \)).

3. Computation of the Optimal FSWL Markov COVER

Necessary conditions for the optimal solution in Theorem 2.1 can be obtained using the method of Lagrange multipliers. Specifically, let
\[ J = (\text{tr}(K_q^{1/2}))^2 + \text{tr}[\Lambda(-K_q + A_q^* K_q A_q + C_q^* C_q)] + \text{tr}[\Omega(I - U_q U_q^*)] \]  

where
\[ K_q = K_q^{1/2} K_q^{1/2}, \quad \Lambda = \Lambda^* \in R^{r \times r}; \quad \Omega = \Omega^* \in R^{s \times s} \]  

are symmetric Lagrange multipliers. After taking derivatives of \( J \) using (1.22) and (1.25)
\[ \frac{\partial J}{\partial \Lambda} = -K_q + A_q^* K_q A_q + C_q^* C_q \]
\[ \frac{\partial J}{\partial \Omega} = I - U_q U_q^* \]
\[ \frac{\partial J}{\partial K_q^{1/2}} = 2I - 2\Lambda K_q^{1/2} + 2A_q^* \Lambda A_q K_q^{1/2} \]
\[ \frac{\partial J}{\partial U_q} = 2G_{q2} L_{12} K_q A_q \Lambda G_{q3}^* - 2\Omega U_q \]

By setting these derivatives to zero we obtain the following result.

**Lemma 3.1** Necessary conditions for the derivation of the optimal \( q \)-FSWL Markov COVER are given by
\[ K_q = A_q^* K_q A_q + C_q^* C_q \]
\[ \Lambda = A_q \Lambda^* \Lambda_q^{-1/2}; \quad \Lambda = \Lambda^* \in R^{t \times t} \]
\[ U_q U_q^* = I \quad ; \quad U_q \in R^{t \times (t - p_q)} \]  
\[ \Omega U_q - P_q U_q Q_q = R_q \quad ; \quad \Omega = \Omega^* \in R^{t \times t} \]

where

\[ P_q = P_q^* = G_{q2} L_{12}^* K_q L_{12} G_{q2} \in R^{t \times t} \]
\[ Q_q = Q_q^* = G_{q3} \Lambda G_{q3}^* \in R^{(t_q - \rho_q) \times (t_q - \rho_q)} \]  
\[ R_q = G_{q2} L_{12}^* K_q (L_{11} + L_{12} G_{q1}) \Lambda G_{q3}^* \in R^{t \times (t_q - \rho_q)} \]

and \( A_q, G_{qj}, L_{ij} \) are defined by (1.20)-(1.24)

These necessary conditions cannot be solved explicitly for the optimal row unitary matrix \( U_q \) and so an iterative procedure is required. One possible algorithm is now described.

**Recursive Algorithm for Optimal q-FSWL Markov COVER:**

1. Set \( j = 0 \) and choose any row unitary \( U_q(0) \) in (1.21a)

2. Form \( A_q(j) \) from

\[ A_q(j) = (L_{11} + L_{12} G_{q1}) + L_{12} G_{q2} U_q(j) G_{q3} \]  

3. Compute \( K_q(j) \): \( K_1(j) = A_q^*(j) K_q(j) A_q(j) + C_q^* C_q \)  

4. Compute \( \Lambda(j) \): \( \Lambda(j) = A_q(j) \Lambda(j) A_q^*(j) + K_q^{-1/2}(j) \); \( \Lambda(j) = \Lambda^*(j) \)  

5. Compute \( P_q(j), Q_q(j), R_q(j) \):

\[ P_q(j) = G_{q2} L_{12}^* K_q(j) L_{12} G_{q2}; \quad Q_q(j) = G_{q3} \Lambda(j) G_{q3}^* \]
\[ R_q(j) = G_{q2} L_{12}^* K_q(j) (L_{11} + L_{12} G_{q1}) \Lambda(j) G_{q3}^* \]  

---

869
Update $U_q(j)$ by solving the nonlinear algebra problem:

$$
\Omega(j)U_q(j+1) - P_q(j)U_q(j+1)Q_q(j) = R_q(j); \quad \Omega(j) = \Omega^*(j) \tag{3.5e}
$$

$$
U_q(j+1)U_q^*(j+1) = I
$$

The most difficult step at each stage of the algorithm is to solve (3.5e) for a row unitary $U_q(j+1)$ and symmetric $\Omega(j)$. There is generally no explicit solution except for the following special cases.

**Lemma 3.2** Consider the equation

$$
\Omega U_q - P_q U_q Q_q = R_q; \quad \Omega \in R^{n \times n} \tag{3.6}
$$

where

$$
P_q = P_q^* \in R^{n \times n}; \quad Q_q = Q_q^* \in R^{(\tau_q - \rho_q) \times (\tau_q - \rho_q)}; \quad R_q \in R^{n \times (\tau_q - \rho_q)} \tag{3.7}
$$

are given. Then there exists an analytical solution $(\Omega, U_q)$ with $\Omega$ symmetric and $U_q$ row unitary when $s_q = 1$ or $Q_q = \beta I$. ($\beta$ scalar)

a. When $s_q = 1$, $\Omega$ and $P_q$ are scalars and $R_q$ is a row vector. Then $U_q$ is arbitrary for $R_q = 0$ while for $R_q \neq 0$

$$
U_q = R_q (\Omega I - P_q Q_q)^{-1}; \quad \|U_q\| = 1 \tag{3.8}
$$

b. When $Q_q = \beta I$, let $R_q R_q^*$ have the singular value decomposition

$$
R_q R_q^* = (V_1 \quad V_2) \begin{bmatrix}
\Sigma_{q_1} & 0 \\
0 & 0
\end{bmatrix} \begin{bmatrix}
V_1^* \\
V_2^*
\end{bmatrix}
$$

where $\Sigma_{q_1}$ is invertible. Then

$$
U_q = (V_1^*)^+ \Sigma_{q_1}^{-1/2} V_1^* R_q; \quad \Omega = \beta P_q + V_1 \Sigma_{q_1}^{1/2} V_1^* \tag{3.9}
$$

In particular, when $R_q R_q^*$ has full rank,

$$
U_q = (R_q R_q^*)^{-1/2} R_q \tag{3.10}
$$

**Proof:** For case (a)

$$
U_q(\Omega I - P_q Q_q) = R_q; \quad \Omega \text{ scalar}
$$

so that (3.8) follows by the row unitary property $U_q U_q^* = I$. In case (b)
and using the row unitary property of \( U_q \)

\[
(\Omega - \beta P_q) U_q = R_q
\]

Hence using the SVD of \( R_q R_q^* \)

\[
V_1 U_q^* R_q = U_q
\]

But \( V_1^* V_1 = I \) and \( V_1^* \) has full row rank which gives (3.9).

Strictly speaking, (3.8) is not an analytical solution since the scalar \( \Omega \) must still be chosen so that \( ||U_q|| = 1 \). Note that by Corollary 1.1, \( G_q 3G_q^* = I \) so that \( Q_q(j) = I \) in (3.5b) if \( A(j) = I \). The necessary condition (3.5e) is equivalent to assuming \( K_q(j), A(j), P_q(j), Q_q(j) \) and \( R_q(j) \) are known and optimizing over row unitary \( U_q(j+1) \). That is, after dropping the index \( j \) and \( j+1 \) in (3.5e) we have the following result.

**Lemma 3.3** Suppose \( P_q, Q_q \) and \( R_q \) in (3.7) are known. Then a necessary condition for a row unitary matrix \( U_q \) to achieve optimally for the problem:

\[
\min_{U_q} \text{tr}[Q_q U_q^* P_q U_q + 2R_q U_q]; \quad U_q \in \mathbb{R}^{r_q \times (r_q - p_q)}
\]  

(3.11)

is that there exists a symmetric matrix \( \Omega \) such that (3.6) is satisfied.

Furthermore, the optimization in (3.11) is equivalent to

\[
\min_{U} J(U); \quad U \in \mathbb{R}^{r_q \times (r_q - p_q)}
\]  

(3.12a)

where

\[
J(U) = \text{tr}[QU^* PU + 2RU]
\]  

(3.12b)

over unitary matrices \( U^* = [U_q^* V_q^*] \) where \( Q = Q_q \) and

\[
P = P^* = \begin{bmatrix} P_q & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{(r_q - p_q) \times (r_q - p_q)}
\]

\[
R = [R_q 0] \in \mathbb{R}^{(r_q - p_q) \times (r_q - p_q)}
\]

The advantage of the point of view (3.12) is that \( U \) can be treated as a **square**
The solution to (3.12) when \( U \) is a 2x2 unitary matrix is provided in the following lemma. The result can be derived by directly substituting into (3.12).

**Lemma 3.4** Suppose \( P = P^* = [p_{ij}] \), \( Q = Q^* = [q_{ij}] \) and \( R = [r_{ij}] \) are 2x2 matrices. Then the minimum in (3.12) over 2x2 unitary matrices \( U \) is achieved by either

(i) \( U = \text{diag}(u_1,u_2) \) where \( u_1^2 = 1 \), \( u_2^2 = 1 \) minimize

\[
J_1 = r_{11}u_1 + r_{22}u_2 + 2q_{12}p_{12}u_1u_2
\]  

or (ii)

\[
U = \begin{bmatrix}
x & \sqrt{1-x^2} \\
-\sqrt{1-x^2} & x
\end{bmatrix}
\]

where \( |x| \leq 1 \) minimizes

\[
J_2(x) = ax^2 + 2bx + 2(cx+d)\sqrt{1-x^2}
\]

\[
a = (p_{11} - p_{22})(q_{11} - q_{22}), \quad b = r_{11} + r_{22}
\]

\[
c = q_{12}(p_{11} - p_{22}) + p_{12}(q_{22} - q_{11}), \quad d = r_{21} - r_{12}
\]

\( \Box \Box \Box \)

Note that we must optimize over the disjoint sets of 2x2 unitary matrices consisting of *signature matrices* (as in (3.13)) and *rotations* (as in (3.14)). The optimal solution of (3.13) can be obtained by inspection of the magnitudes of the coefficients in \( u_j \). For example, suppose

\[
|r_{11}| \geq |q_{12}p_{12}| \geq |r_{22}|
\]

Then

\[
u_1 = -\text{sgn}(r_{11}); \quad u_1u_2 = -\text{sgn}(q_{12}p_{12})
\]

However the optimization in (3.14) requires numerical solution.

A general \( n \times n \) unitary matrix \( U \) is either a *signature matrix* (i.e. a diagonal matrix \( \Sigma \) such that \( \Sigma^2 = I \)) or a product of \( \frac{1}{2} n(n-1) \) *rotations* \( U_{ij} \) where the components of \( U_{ij}(k,l) \) \( U_{ij} \) are defined by

\[
U_{ij}(i,i) = U_{ij}(j,j) = \cos \theta_{ij}
\]  

(3.15a)
A particular signature matrix is also defined by (3.15b) where

\[
U_{ij}(k,k) = \pm 1 \quad \text{for} \quad k = i,j
\]

By letting

\[
U = \prod_{ij} U_{ij}
\]

The optimization in (3.12) can be reduced to a sequence of one dimensional optimizations over the angles \(\theta_{ij}\). To be complete, \(J(U)\) should also be evaluated separately for all \(2^n (n = r_q - p_q)\) signature matrices. A compromise during the iterative procedure is to include the possibility of components \(U_{ij}\) being defined by (3.16) as well as (3.15a). Rather than present the general result we only illustrate by means of an example.

Specifically, suppose we express a 3x3 unitary matrix \(U\) as

\[
U = U_{12}U_{13}U_{23}
\]

Then by invoking the trace property, \(J\) in (3.12b) can equivalently be expressed as

\[
J(U_{ij}) = \text{tr}[Q_{ij}U_{ij}^*P_{ij}U_{ij} + 2R_{ij}U_{ij}] \quad (3.18a)
\]

where

\[
Q_{12} = U_{12}U_{23}QU_{23}^*U_{13}^*; \quad P_{12} = P; \quad R_{12} = U_{13}U_{23}R
\]

\[
Q_{13} = U_{13}QU_{23}^*; \quad P_{13} = U_{12}^*PU_{12}; \quad R_{13} = U_{13}RU_{12}
\]

\[
Q_{23} = Q; \quad P_{23} = U_{13}^*U_{12}^*PU_{12}U_{13}; \quad R_{23} = RU_{12}U_{13}
\]

With \(i = i_0\), and \(j = j_0\) fixed in (3.18a), \(J\) can be optimization over \(U_{i_0j_0}\). The procedure is recursive. That is, first assume \(i = 1, j = 2\) with \(U_{13}\) and \(U_{23}\) both initialized to (say) the identity. After optimizing over \(U_{12}\), fix \(U_{12}\) and \(U_{13}\) and optimize over \(U_{23}\), etc. Many cycles may be necessary for convergence.

In order to explicitly demonstrate the formulation for each of the 2x2 optimizations consider the case \(i = 1, j = 2\), and express

873
Q_{12} = \begin{bmatrix} 1 & 2 \\ Q_{12} & Q_{12} \end{bmatrix}, \quad P_{12} = \begin{bmatrix} 1 & 2 \\ P_{12} & P_{12} \end{bmatrix}, \quad R_{12} = \begin{bmatrix} 1 & 2 \\ R_{12} & R_{12} \end{bmatrix}

\begin{bmatrix} 2 & 3 \\ Q_{12} & Q_{12} \end{bmatrix}

\begin{bmatrix} 2 & 3 \\ P_{12} & P_{22} \end{bmatrix}

\begin{bmatrix} 4 & 3 \\ R_{12} & R_{22} \end{bmatrix}

where Q_{12}, P_{12}, R_{12} \in \mathbb{R}^{2x2}. Then from (3.15), (3.16) the optimal \( \theta_{12} \) which minimizes \( J_{12}(U_{12}) \) also minimizes

\[ J_{12}(\theta_{12}) = \text{tr}[Q_{12}U^*_0P_{12}U_0 + 2(R_{11} + Q_{12}P_{12}^*)U_0] \]

where components of the 2x2 unitary matrix \( U^*_0 \) is defined by (3.15a) or (3.16) for \( i, j, \in \{1,2\} \) The 2x2 optimization of \( J_{12}(\theta_{12}) \) over \( \theta_{12} \) is partially solved in lemma 3.4.

Before concluding this section it is important to reiterate that the dimension of the problem for optimizing over the row unitary matrices \( U_q \) is generally low. In particular from (1.21b) both the number of rows and columns of \( U_q \) is not greater than the number of outputs. For a single output systems, \( U_q \) is a scalar and so there are at most two possibilities, and no optimization is necessary. That is, for \( \rho_q < r_q \) we merely evaluate the cost in (2.24) for two values of \( G_q \) in (1.21a) corresponding to \( U_q = \pm 1 \), while if \( \rho_q = r_q \), then \( G_q = G_{q1} \) is unique.

4. Coefficient Errors

Recall that \( Y \) in (2.9) is the error in the covariance of the output \( \{\gamma(k)\} \) due to finite precision implementation of both states and coefficients of the q-Markov COVER. The optimal q-FSWL Markov COVER minimizes the trace of \( Y \) when there are no coefficient errors (corresponding to \( \gamma_0 = 0 \)). Furthermore, when there are no coefficient errors, there are no errors in either the Markov parameters \( M_i \) or covariance parameters \( R_j \) in (1.2). Once coefficient errors are introduced and all finite wordlength (FWL) errors are considered, there is no longer a clear interpretation of what should constitute the optimal q-FSWL Markov COVER. One possibility is to again attempt to minimize the trace of \( Y \). Alternative performance criteria could be based on the errors \( \Delta M_i \) and \( \Delta R_j \) in the Markov and covariance parameters as given by

\[ M_i + \Delta M_i = (C+\Delta C)(A+\Delta A)(B+\Delta B); \]

\[ R + \Delta R_j = (C+\Delta C)(A+\Delta A)^\dagger X(C+\Delta C)^* \]

where \( \overline{X} \) satisfies \( \overline{X} = AA^* + BB^* \). For example, one could attempt to minimize
However there are no results which directly connect $C_M$ or $C_R$ with errors in time or frequency response of the q-Markov COVERs. Furthermore, the analytical and computational aspects involved in the resulting optimization would be very difficult if not practically impossible.

A convenient approach to parameter optimization is to assume a statistical model for parameter errors. A statistical design can be justified along the following lines. Suppose (as is the case in practice) that both the Markov parameters $M_i$ and covariance parameters $R_j$ are known only to be accurate up to a specified wordlength, and any higher precisional representation is regarded as uncorrelated random noise. Then the calculation of all q-Markov COVERs (for a particular row unitary matrix $U_q$) will also only be accurate to a finite precision beyond which the parameter representation contains uncorrelated random noise.

Lemma 4.1 Suppose $M = M^* > 0$ and $K = K^* > 0$ are given nxn matrices. Let $v_j \in \mathbb{R}^n$ be a zero mean random variable uniformly distributed between ±1 with uncorrelated components which are also uncorrelated with components of $v_i$. Then we have

$$E \{ v_j^* M v_j \} = \frac{1}{3} \text{tr}[M].$$

Furthermore

$$E \{ \text{tr}[V^* M V K] \} = \frac{1}{3} \text{tr}[MK]$$

where

$$V = [v_1^2 \cdots v_n^2] \in \mathbb{R}^{n \times n}.$$

Unfortunately these results cannot be applied directly to (2.9) since $X$ itself is a random variable. However if we approximate $\bar{X}$ by $X$ we can deduce the following result.

**Theorem 4.1**

Suppose the components of $\delta A$, $\delta B$ and $\delta C$ are zero mean uncorrelated random variables uniformly distributed between ±1. Then $E(J)$ where $J = \text{tr}[Y]$ is
approximated by $E\{\hat{J}\}$ where

$$E\{\hat{J}\} = \gamma^2 \text{tr}[B^*KB] + (\gamma^2 + \frac{\gamma_0^2}{3})\text{tr}[K] + \frac{\gamma_0^2}{3}(\text{tr}[XX] + \text{tr}[X]) \quad (4.5)$$

where $K, X$ are defined by (2.11) and (1.3).

\[\square\square\square\]

Proof: From (2.9) ignoring the linear term in $\delta C$

$$J = \gamma^2 (\text{tr}[K] + \text{tr}[B^*KB]) +$$

$$\gamma_0^2 (\text{tr}[(\delta A)^*X(\delta A)K] + \text{tr}[(\delta B)^*K(\delta B)] + \text{tr}[(\delta C)^*X\delta C])$$

The result then follows using Theorem 2.1.

\[\square\square\square\]

Under a similarity transformation $T$, the performance measure (4.5) becomes

$$E\{\hat{J}_T\} \triangleq \gamma^2 \text{tr}[B^*KB] + (\gamma^2 + \frac{\gamma_0^2}{3})\text{tr}[T^*KT] + \frac{\gamma_0^2}{3}(\text{tr}[XX] + \text{tr}[T^{-1}X(T^{-1})^*]) \quad (4.6)$$

Note that both $\text{tr}[B^*KB]$ and $\text{tr}[XX]$ are invariant. In fact the invariant eigenvalues $\{\sigma_k^2\}$ of $XX$ are the squares of the Hankel singular values of the system defined by $\{A,B,C\}$. Consequently we need only consider the minimization of

$$(\gamma^2 + \frac{\gamma_0^2}{3})\text{tr}[T^*KT] + \frac{\gamma_0^2}{3} \text{tr}[T^{-1}X(T^{-1})^*] \quad (4.7)$$

over similarity transformations $T$. We make use of an earlier result [8] to provide the minimum in (4.7).

Theorem 4.2 [8]

Consider a minimal asymptotically stable order system $\{A,B,C\}$ with controllability gramman $X$ and observability gramman $K$. Let $\hat{X}$ and $\hat{K}$ be the transformed grammans as a result of applying a similarity transformation $T$; that is

$$\hat{X} = T^{-1}X(T^{-1})^*; \quad \hat{K} = T^*KT \quad (4.8)$$

Then
where \( \{ \sigma_k^2 \} \) are the Hankel singular values. Moreover equality is achieved in (4.9) if and only if

\[
\tilde{K} = \alpha^2 \tilde{X}
\]  

(4.10)

In particular, in (4.7)

\[
\min_T J_T = \gamma^2 \text{tr}[B^*KB] + \frac{\gamma_0^2}{3} \left( \sum_{k=1}^{r_q} \sigma_k^2 + 2\alpha \sum_{k=1}^{r_q} \sigma_k \right)
\]  

(4.11a)

where

\[
\alpha = \sqrt{1 + 3(\gamma/\gamma_0)^2}
\]  

(4.11b)

The minimum value is achieved in (4.11a) when \( \tilde{K}, \tilde{X} \) satisfy (4.10) with \( \alpha \) given by (4.11b)

One optimal realization (4.10) is a scaled internally balanced structure; that is

\[
\tilde{X}_1 = \alpha^{-1} \text{diag} (\sigma_1, \sigma_2, ..., \sigma_{r_q}); \quad \tilde{K}_1 = \alpha \text{diag} (\sigma_1, \sigma_2, \cdots, \sigma_{r_q})
\]  

(4.12)

From the point of view of \( l_2 \)-scaling, equal diagonal components of \( \tilde{X} \) guarantee equal dynamic range of the state components. It is evident from (4.10) that any unitary transformation \( \tilde{U} \) applied to the coordinate basis having \( \tilde{X} \) and \( \tilde{K} \) as the respective controllability and observability grammians will not alter the optimal performance. Consequently an optimal realization in which all diagonal components of the controllability grammian are equal exists with controllability grammian \( \tilde{U}^*\tilde{X}_1\tilde{U} \) and observability grammian \( \tilde{U}^*\tilde{K}_1\tilde{U} \) such that

\[
\tilde{U}^*\tilde{X}_1\tilde{U}_{jj} = \frac{1}{\alpha r_q} \sum_{k=1}^{r_q} \sigma_k \quad \text{for all } j
\]  

(4.13)

where \( \tilde{X}_1, \tilde{K}_1 \) are defined by (4.12) and \( \tilde{U} \) unitary. The existence of \( \tilde{U} \) is guaranteed by lemma 2.1 and an explicit algorithm for constructing a (nonunique) \( \tilde{U} \) is available in [9, Appendix A].

Corollary 4.1

The optimal q-FSWL COVER which minimizes (2.21) subject to the \( l_2 \)-scaling constraint
\[ \Lambda_{ij} = \frac{1}{\alpha r_q} \sum_{k=1}^{r_q} \sigma_k \quad \text{for all } j \] (4.14)

also minimizes \( E\{\tilde{J}_T\} \) in (4.6).

This result provides a connection between the optimal q-FSWL COVER structure which minimizes only the effects due to state quantization noise, and the suboptimal q-FWL Markov COVER structure which minimizes \( E\{\tilde{J}_T\} \) subject to the assumed random parameter error model stated in Theorem 4.1. Once again we note that the result is suboptimal in the sense that \( \hat{X} \) and \( X \) in (2.9) and (4.5) are only approximately equal. The result of Corollary 4.1 is also only of academic value since the \( l_2 \)-constraint (4.14) is not known until the design is complete since the Hankel singular values \( \{\sigma_j\} \) depend on the optimal row unitary matrix \( U_p \) as provided in Theorem 2.1. However a more explicit result can be stated.

**Corollary 4.2**

The optimal q-FSWL cover subject to the \( l_2 \)-scaling constraint (2.22) also minimizes \( E\{\tilde{J}_T\} \) in (4.12) subject to (2.22).

---

5. **An Example**

Consider a 5 mode simply supported beam of length \( \pi \) having 2 inputs \( u_1, u_2 \) and 2 outputs \( y_1, y_2 \)

\[ u_1 = F(0.2\pi, t), \quad u_2 = T(\pi, t) \]
\[ y_1 = \theta(0, t), \quad y_2 = \mu(0.6\pi, t) \]

where \( F(0.2\pi, t) \) denotes a force applied at .2\( \pi \) units from the left end of the beam, \( T(\pi, t) \) denotes a torque at the right end of the beam, \( \theta(0, t) \) denotes angular deflection at the left end, and \( \mu(0.6\pi, t) \) denotes rectilinear deflection at 0.6\( \pi \) from the left end of the beam. The equations of motion are assumed to be described by

\[ \ddot{\eta}_k + 2\xi_k \omega_k \dot{\eta}_k + \omega_k^2 \eta_k = [\sin(0.2\pi k) \text{kcos}(\pi k)] \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \]
\[
\begin{bmatrix}
Y_1 \\
Y_2
\end{bmatrix} = \sum_{k=1}^{5} \begin{bmatrix}
k \cos(0.6\pi k) \\
\sin(0.6\pi k)
\end{bmatrix} \eta_k
\]

(5.1)

where \(\omega_k = k^2\) rads/sec. and \(\xi = 0.005\). A continuous time 10th order state space model is defined by

\[
\dot{x} = Fx + Gu, \quad y = Cx
\]

where

\[
x = (\eta_1 \dot{\eta}_1 \eta_2 \dot{\eta}_2 \cdots \eta_5 \dot{\eta}_5)^T
\]

(5.2)

A zero order hold equivalent 10th order discrete model (1.1) is defined by

\[
A = e^{FT}; \quad B = \int_0^T e^{FS}d\sigma G
\]

For the numerical work, a sampling period \(T = 0.025\) sec. was selected which corresponded to approximately 10 samples in the shortest period. The eigenvalues of \(A\) are at

\[
0.996 \pm j0.0250, \quad 0.9985 \pm j0.0500, \quad 0.9968 \pm j0.0750, \quad 0.9945 \pm j0.0998, \quad 0.9916 \pm j0.1246.
\]

Using the algorithm described in Corollary 1.1 the following results were obtained.

<table>
<thead>
<tr>
<th>(q)</th>
<th>(s_q)</th>
<th>(r_q)</th>
<th>(\rho_q)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>(\geq 7)</td>
<td>2</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>

Hence for \(q = 2, 3, 4\), \(U_q\) in (1.22b) can be an arbitrary 2x2 unitary matrix, while for \(q \geq 5\) there is no remaining freedom in the \(q\)-COVER.

Optimal \(q\)-FSWL COVER designs:

\[
U_q = \begin{bmatrix}
\cos\theta_q & \sin\theta_q \\
-sin\theta_q & \cos\theta_q
\end{bmatrix}; \quad \theta_2 = 40^\circ \quad \theta_3 = 0^\circ \quad \theta_4 = 65^\circ
\]

(other cases \(\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}\) and \(\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}\) were also checked and neither was optimal).
The cost ranges from (2.29) for $\alpha = 1$ were
\[
\begin{align*}
\eta_{2_{\text{opt}}} &= 0.3143 \times 10^6 \leq \eta_2 \leq 0.8478 \times 10^6 \\
\eta_{3_{\text{opt}}} &= 0.2570 \times 10^6 \leq \eta_3 \leq 0.4764 \times 10^6 \\
\eta_{4_{\text{opt}}} &= 0.0019 \times 10^8 \leq \eta_4 \leq 0.1308 \times 10^8
\end{align*}
\]

The actual FWL output roundoff noise is given by
\[
\gamma^2 \eta_q; \quad \gamma^2 = \frac{1}{12} 2^{-2^W}
\]
where $W$ bits are assigned to the fractional wordlength of the state. Hence a factor of 4 improvement in $\eta_q$ corresponds to a wordlength saving of 1 bit. There is little savings in this example when $q = 2, 3$. However for $q = 4$ we have a saving of 4 bits. In practice, for fast sampling and low structural damping, the savings would increase as the dimension of the model increases (e.g. a simply supported beam of 50 modes with $q = 8$).
References


