An Analytical Approach for the Synthesis of Two-Dimensional State-Space Filter Structures with Minimum Weighted Sensitivity

Takao Hinamoto, *Senior Member, IEEE*, Yoshitaka Zempo, Yoshio Nishino, and Wu-Sheng Lu, *Fellow, IEEE*

***Abstract*— This paper considers the problem of synthesizing the finite-word-length (FWL) two-dimensional (2-D) state-space filter structures with minimum weighted sensitivity. Two kinds of frequency-weighted sensitivity measures, one based on a mixture of L_1/L_2 norms and the other a pure L_2 norm, are defined in place of the usual sensitivity measure and an upper bound expressed in terms of 2-D weighted Gramians is used to evaluate the weighted L_1/L_2 mixed sensitivity. A simple technique is then developed for obtaining a set of filter structures with very low weighted L_1/L_2-sensitivity. In this connection, the optimal coordinate transformation is characterized in a closed form. Next, an iterative procedure is proposed to obtain the optimal coordinate transformation that minimizes the weighted L_2-sensitivity measure. Once the initial value is given, the estimate at each iteration can be calculated analytically. Finally, two numerical examples are given to illustrate the utility of the proposed technique.**

***Index Terms*—Finite word length, optimal realization, Roesser model, two-dimensional IIR digital filter, weighted coefficient sensitivity.**

I. Introduction

UNDESIRABLE finite-word-length (FWL) effects arise in the fixed-point implementation of recursive digital filters. One of them is the deviation of the actual transfer function from the ideal transfer function, which is caused by the truncation or rounding of the filter coefficients. As is well known, the state-space approach allows such an effect to be minimized by appropriately choosing a filter structure that minimizes a well-defined FWL effect. Several techniques have been reported to synthesize linear state-space systems that minimize the coefficient sensitivity [1]–[7]. A similar technique for multi-input–multi-output continuous-time systems has also been presented [8]. In addition, the problem of minimizing the coefficient sensitivity of two-dimensional (2-D) state-space digital filters has been studied [9]–[15]. Based on the Roesser local state-space (LSS) model [16], Zilouchian and Carroll have investigated a coefficient sensitivity bound in 2-D state-space digital filters [9]. Effective methods for synthesizing 2-D filter structures with minimum coefficient sensitivity have been investigated [10], [13]. In [10], all the frequency regions are treated uniformly. The method reported in [13] studies the sensitivity behavior of a transfer function within a specified frequency range. Based on the Fornasini-Marchesini second LSS model [17], similar techniques have been explored [11], [14]. The frequency-weighted sensitivity measures have been introduced in [13], [14], where a constraint on the weights of the various terms of the measure is imposed. More recently, the frequency-weighted L_2-sensitivity problem has been considered by [15] via a 2-D gradient-flow-based optimization technique that was initiated in [7] for the one-dimensional (1-D) case. It was argued in [7] and [15] that the L_2 sensitivity minimization, although technically more challenging, is more natural and reasonable than the conventional L_1/L_2 mixed sensitivity minimization.

This paper treats the problem of reducing the coefficient sensitivity of 2-D state-space digital filters within a specified frequency range. Here, the Roesser LSS model is employed to describe 2-D state-space digital filters. From a practical viewpoint, we are interested in the sensitivity performance of the transfer function within a specified frequency range. This is achieved by defining a weighted sensitivity function. One contribution of our paper is to address the case of general unconstrained frequency weights for 2-D state-space digital filters. Another, is to solve the corresponding problem of synthesizing the filter structures with minimum weighted sensitivity. First, the sensitivities of a 2-D transfer function with respect to state-space parameters are analyzed in conjunction with frequency weighted functions. The overall frequency-weighted sensitivity measure is then evaluated, using a mixture of L_1/L_2 norms, as well as a pure L_2 norm. Second, a simple technique is developed for synthesizing the 2-D filter structures with very low frequency-weighted L_1/L_2-sensitivity. A closed-form solution that is optimal in a certain sense is obtained. The 1-D version of this closed-form solution turns out to be more efficient than the one proposed in [5]. Notice that the closed-form solution reported in [5] is restrictive and only exists under a certain constraint. Third, an iterative procedure is presented to find the optimal coordinate transformation that minimizes the weighted L_2-sensitivity measure. This

Manuscript received October 15, 1997; revised May 26, 1998. This paper was recommended by Associate Editor G. Martinelli.

T. Hinamoto, Y. Zempo and Y. Nishino are with the Faculty of Engineering, Hiroshima University, Higashi-Hiroshima, Japan 739-8527.

W.-S. Lu is with the Department of Electrial and Computer Engineering, University of Victoria, Victoria, BC, V8W 3P6, Canada.

Publisher Item Identifier S 1057-7122(99)08102-7.

procedure is advantageous since the estimate is calculated analytically at each iteration. Finally, two numerical examples are presented to demonstrate the validity of the proposed technique.

Throughout this paper, the n-dimensional identity matrix is denoted by $\boldsymbol{I}_n$. The transpose (conjugate transpose) of any matrix $\boldsymbol{A}$ is indicated by $\boldsymbol{A}^t$ ($\boldsymbol{A}^*$) and tr$\boldsymbol{A}$ and $\oplus$ are used to denote the trace of a square matrix $\boldsymbol{A}$ and the direct sum of matrices, respectively.

II. Weighted L_1/L_2 Mixed Sensitivity Analysis

Consider the following LSS model $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)_{m,n}$ for 2-D digital filters which was originally proposed by Roesser [16]:

$$\begin{bmatrix}\boldsymbol{x}_{11}(i,j)\\ y(i,j)\end{bmatrix} = \begin{bmatrix}\boldsymbol{A} & \boldsymbol{b}\\ \boldsymbol{c} & d\end{bmatrix}\begin{bmatrix}\boldsymbol{x}(i,j)\\ u(i,j)\end{bmatrix} \tag{1}$$

where

$$\boldsymbol{x}_{11}(i,j) = \begin{bmatrix}\boldsymbol{x}^h(i+1,j)\\ \boldsymbol{x}^v(i,j+1)\end{bmatrix}, \qquad \boldsymbol{x}(i,j) = \begin{bmatrix}\boldsymbol{x}^h(i,j)\\ \boldsymbol{x}^v(i,j)\end{bmatrix}$$
$$\boldsymbol{A} = \begin{bmatrix}\boldsymbol{A}_1 & \boldsymbol{A}_2\\ \boldsymbol{A}_3 & \boldsymbol{A}_4\end{bmatrix}, \qquad \boldsymbol{b} = \begin{bmatrix}\boldsymbol{b}_1\\ \boldsymbol{b}_2\end{bmatrix}, \qquad \boldsymbol{c} = [\boldsymbol{c}_1 \quad \boldsymbol{c}_2].$$

Here $\boldsymbol{x}^h(i,j)$ is an $m \times 1$ horizontal state vector, $\boldsymbol{x}^v(i,j)$ is an $n \times 1$ vertical state vector, $u(i,j)$ is a scalar input, $y(i,j)$ is a scalar output, and $\boldsymbol{A}_1$, $\boldsymbol{A}_2$, $\boldsymbol{A}_3$, $\boldsymbol{A}_4$, $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, $\boldsymbol{c}_1$, $\boldsymbol{c}_2$, d are real constant matrices of appropriate dimensions. The LSS model (1) is assumed to be BIBO stable, separately locally controllable, and separately locally observable [18]. Define

$$\begin{aligned}\boldsymbol{F}(z_1,z_2) &= (\boldsymbol{S}-\boldsymbol{A})^{-1}\boldsymbol{b}, \quad \boldsymbol{G}(z_1,z_2) = \boldsymbol{c}(\boldsymbol{S}-\boldsymbol{A})^{-1}\\ H(z_1,z_2) &= \boldsymbol{c}(\boldsymbol{S}-\boldsymbol{A})^{-1}\boldsymbol{b} + d, \quad \boldsymbol{S} = z_1\boldsymbol{I}_m \oplus z_2\boldsymbol{I}_n\end{aligned} \tag{2}$$

where $\boldsymbol{F}(z_1,z_2)$ consists of the transfer functions from the filter input to the filter states, $\boldsymbol{G}(z_1,z_2)$ is defined as the set of transfer functions from the input of each of delay operators to the output, and $H(z_1,z_2)$ is the transfer function from the filter input to the output.

Let the coordinate transformation be specified as

$$\overline{\boldsymbol{x}}(i,j) = \boldsymbol{T}^{-1}\boldsymbol{x}(i,j) \tag{3}$$

where $\boldsymbol{T} = \boldsymbol{T}_1 \oplus \boldsymbol{T}_4$ and $\boldsymbol{T}_1$ ($\boldsymbol{T}_4$) is an $m \times m$ ($n \times n$) nonsingular matrix. Then an algebraically equivalent realization $(\overline{\boldsymbol{A}}, \overline{\boldsymbol{b}}, \overline{\boldsymbol{c}}, d)_{m,n}$ given by

$$\overline{\boldsymbol{A}} = \boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T}, \overline{\boldsymbol{b}} = \boldsymbol{T}^{-1}\boldsymbol{b}, \overline{\boldsymbol{c}} = \boldsymbol{c}\boldsymbol{T} \tag{4}$$

is obtained. From (2) and (4) it is clear that the transfer function $H(z_1,z_2)$ is invariant under such a transformation.

Definition 1: Let $\boldsymbol{X}$ be an $m \times n$ real matrix and let $f(\boldsymbol{X})$ be a scalar complex function of $\boldsymbol{X}$, differentiable with respect to all the entries of $\boldsymbol{X}$. The sensitivity function of f with respect to $\boldsymbol{X}$ is then defined as

$$\boldsymbol{S}_X = \frac{\partial f}{\partial \boldsymbol{X}} \quad \text{with} \quad (\boldsymbol{S}_X)_{ij} = \frac{\partial f}{\partial x_{ij}} \tag{5}$$

where x_{ij} denotes the (i,j)th entry of the matrix $\boldsymbol{X}$.

With these notations, it can easily be shown that

$$\begin{aligned}\frac{\partial H(z_1,z_2)}{\partial \boldsymbol{A}} &= \boldsymbol{G}^t(z_1,z_2)\boldsymbol{F}^t(z_1,z_2)\\ \frac{\partial H(z_1,z_2)}{\partial \boldsymbol{b}} &= \boldsymbol{G}^t(z_1,z_2)\\ \frac{\partial H(z_1,z_2)}{\partial \boldsymbol{c}^t} &= \boldsymbol{F}(z_1,z_2).\end{aligned} \tag{6}$$

The term d and the sensitivity with respect to it are coordinate independent and therefore they are neglected here.

To consider the sensitivity behavior of the transfer function in a specified frequency band, or even at some discrete frequency points, the weighted sensitivity functions are defined as

$$\begin{aligned}\frac{\delta H(z_1,z_2)}{\delta \boldsymbol{A}} &= W_A(z_1,z_2)\frac{\partial H(z_1,z_2)}{\partial \boldsymbol{A}}\\ \frac{\delta H(z_1,z_2)}{\delta \boldsymbol{b}} &= W_B(z_1,z_2)\frac{\partial H(z_1,z_2)}{\partial \boldsymbol{b}}\\ \frac{\delta H(z_1,z_2)}{\delta \boldsymbol{c}^t} &= W_C(z_1,z_2)\frac{\partial H(z_1,z_2)}{\partial \boldsymbol{c}^t}\end{aligned} \tag{7}$$

where $W_A(z_1,z_2)$, $W_B(z_1,z_2)$, and $W_C(z_1,z_2)$ are three stable, causal scalar rational functions of the complex variables z_1 and z_2. It should be noted that δ in (7) is not meant to be a derivative operator, but rather a notation for defining the weighted parameter sensitivity as seen in (7). Let

$$W_A(z_1,z_2) = W_1(z_1,z_2)W_2(z_1,z_2) \tag{8}$$

be a factorization of $W_A(z_1,z_2)$. Note that, unlike the system considered in [13], there is no assumption such that

$$W_1(z_1,z_2) = W_B(z_1,z_2), \quad W_2(z_1,z_2) = W_C(z_1,z_2)$$

for the system considered here.

Definition 2: Let $\boldsymbol{X}(z_1,z_2)$ be an $m \times n$ complex matrix valued function of the complex variables z_1 and z_2. The L_p norm of $\boldsymbol{X}(z_1,z_2)$ is then defined as

$$||\boldsymbol{X}||_p = \left[\frac{1}{(2\pi j)^2}\oint\oint_{\Gamma^2} ||\boldsymbol{X}(z_1,z_2)||_F^p \frac{dz_1\,dz_2}{z_1 z_2}\right]^{1/p} \tag{9}$$

where $\Gamma^2 = \{(z_1,z_2)\colon |z_1| = 1, |z_2| = 1\}$ and $||\boldsymbol{X}(z_1,z_2)||_F$ is the Frobenius norm of the matrix $\boldsymbol{X}(z_1,z_2)$ defined by

$$||\boldsymbol{X}(z_1,z_2)||_F = \left[\sum_{p=1}^{m}\sum_{q=1}^{n}|x_{pq}(z_1,z_2)|^2\right]^{1/2}.$$

The overall weighted L_1/L_2 mixed sensitivity measure is now defined as

$$\begin{aligned}m_{1/2} = \left|\left|\frac{\delta H(z_1,z_2)}{\delta \boldsymbol{A}}\right|\right|_1^2 + \left|\left|\frac{\delta H(z_1,z_2)}{\delta \boldsymbol{b}}\right|\right|_2^2\\ + \left|\left|\frac{\delta H(z_1,z_2)}{\delta \boldsymbol{c}^t}\right|\right|_2^2.\end{aligned} \tag{10}$$

From (6)—(8), we can write (10) as

$$\begin{aligned}m_{1/2} = &||W_1(z_1,z_2)\boldsymbol{G}^t(z_1,z_2)W_2(z_1,z_2)\boldsymbol{F}^t(z_1,z_2)||_1^2\\ &+ ||W_B(z_1,z_2)\boldsymbol{G}^t(z_1,z_2)||_2^2\\ &+ ||W_C(z_1,z_2)\boldsymbol{F}(z_1,z_2)||_2^2.\end{aligned} \tag{11}$$

By the Cauchy–Schwartz inequality, we have

$$\begin{aligned}||W_1(z_1,z_2)\boldsymbol{G}^t(z_1,z_2)W_2(z_1,z_2)\boldsymbol{F}^t(z_1,z_2)||_1^2 \\ \le ||W_1(z_1,z_2)\boldsymbol{G}^t(z_1,z_2)||_2^2||W_2(z_1,z_2)\boldsymbol{F}(z_1,z_2)||_2^2\end{aligned} \tag{12a}$$

where the equality sign holds if and only if

$$\begin{aligned}|W_1(z_1,z_2)|^2\boldsymbol{G}(z_1,z_2)\boldsymbol{G}^*(z_1,z_2) \\ = \rho^2|W_2(z_1,z_2)|^2\boldsymbol{F}^*(z_1,z_2)\boldsymbol{F}(z_1,z_2)\end{aligned} \tag{12b}$$

for some nonzero real number ρ. To facilitate the mathematical treatment, an upper bound of $m_{1/2}$ is employed as follows:

$$\begin{aligned}M_{1/2} = {}&||W_1(z_1,z_2)\boldsymbol{G}^t(z_1,z_2)||_2^2||W_2(z_1,z_2)\boldsymbol{F}(z_1,z_2)||_2^2 \\ &+ ||W_B(z_1,z_2)\boldsymbol{G}^t(z_1,z_2)||_2^2 \\ &+ ||W_C(z_1,z_2)\boldsymbol{F}(z_1,z_2)||_2^2\end{aligned} \tag{13}$$

where $m_{1/2} \le M_{1/2}$. This upper bound can be viewed as a 2-D extension of the upper bound $m_{1/2} \le M_{1/2}$ for the 1-D case introduced by Thiele [2]. From (9) it is easy to show that

$$M_{1/2} = \operatorname{tr}\boldsymbol{K}_{o1}\operatorname{tr}\boldsymbol{K}_{c2} + \operatorname{tr}\boldsymbol{K}_{oB} + \operatorname{tr}\boldsymbol{K}_{cC} \tag{14}$$

where $\boldsymbol{K}_{o1}$, $\boldsymbol{K}_{c2}$, $\boldsymbol{K}_{oB}$, and $\boldsymbol{K}_{cC}$ are often referred to as weighted observability (for those with subindex o) and controllability (for those with subindex c) Gramians, and can be obtained by the following general expression:

$$\boldsymbol{K} = \frac{1}{(2\pi j)^2}\oint\oint_{\Gamma^2}\boldsymbol{Y}(z_1,z_2)\boldsymbol{Y}^*(z_1,z_2)\frac{dz_1\,dz_2}{z_1z_2} \tag{15}$$

with $\boldsymbol{Y}(z_1,z_2) = W_1^*(z_1,z_2)\boldsymbol{G}^*(z_1,z_2)$, $W_2(z_1,z_2)\boldsymbol{F}(z_1,z_2)$, $W_B^*(z_1,z_2)\boldsymbol{G}^*(z_1,z_2)$, and $W_C(z_1,z_2)\boldsymbol{F}(z_1,z_2)$, respectively.

The coordinate transformation defined by (3) transforms the weighted Gramians ($\boldsymbol{K}_{o1}, \boldsymbol{K}_{c2}, \boldsymbol{K}_{oB}, \boldsymbol{K}_{cC}$) into ($\overline{\boldsymbol{K}}_{o1}, \overline{\boldsymbol{K}}_{c2}, \overline{\boldsymbol{K}}_{oB}, \overline{\boldsymbol{K}}_{cC}$). Then (14) is changed to

$$\overline{M}_{1/2} = \operatorname{tr}\overline{\boldsymbol{K}}_{o1}\operatorname{tr}\overline{\boldsymbol{K}}_{c2} + \operatorname{tr}\overline{\boldsymbol{K}}_{oB} + \operatorname{tr}\overline{\boldsymbol{K}}_{cC} \tag{16}$$

where

$$\begin{aligned}\overline{\boldsymbol{K}}_{o1} &= \boldsymbol{T}^t\boldsymbol{K}_{o1}\boldsymbol{T}, \quad \overline{\boldsymbol{K}}_{c2} = \boldsymbol{T}^{-1}\boldsymbol{K}_{c2}\boldsymbol{T}^{-t} \\ \overline{\boldsymbol{K}}_{oB} &= \boldsymbol{T}^t\boldsymbol{K}_{oB}\boldsymbol{T}, \quad \overline{\boldsymbol{K}}_{cC} = \boldsymbol{T}^{-1}\boldsymbol{K}_{cC}\boldsymbol{T}^{-t}.\end{aligned}$$

Moreover, (16) is written as

$$\overline{M}_{1/2}(\boldsymbol{P}) = J(\boldsymbol{P}) + L(\boldsymbol{P}) \tag{17}$$

where

$$\begin{aligned}\boldsymbol{P} &= \boldsymbol{T}\boldsymbol{T}^t, \quad \boldsymbol{T} = \begin{bmatrix}\boldsymbol{T}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{T}_4\end{bmatrix} \\ J(\boldsymbol{P}) &= \operatorname{tr}[\boldsymbol{K}_{o1}\boldsymbol{P}]\operatorname{tr}[\boldsymbol{K}_{c2}\boldsymbol{P}^{-1}], \\ L(\boldsymbol{P}) &= \operatorname{tr}[\boldsymbol{K}_{oB}\boldsymbol{P}] + \operatorname{tr}[\boldsymbol{K}_{cC}\boldsymbol{P}^{-1}].\end{aligned}$$

The problem being considered here is to obtain the symmetric and positive-definite matrix $\boldsymbol{P}$ that minimizes (17), subject to the minimization of $J(\boldsymbol{P})$.

Remark 1: In order to effectively control the upper bound of the L_1-norm term in (11), while minimizing (17), we seek to find a 2-D coordinate transformation $\boldsymbol{T}$ (3) that minimizes (17) subject to the minimization of $J(\boldsymbol{P})$.

III. Filter Synthesis with Very Low Weighted L_1/L_2 Mixed Sensitivity

In this section, we consider the problem of obtaining the matrix $\boldsymbol{P} = \boldsymbol{T}\boldsymbol{T}^t$ that minimizes (17), subject to the minimization of $J(\boldsymbol{P})$, where $\boldsymbol{T}$ is block diagonal. An analytical method will be developed for obtaining such a matrix $\boldsymbol{P}$. The problem of iteratively minimizing (17) with respect to $\boldsymbol{P} = \boldsymbol{T}\boldsymbol{T}^t$ for any nonsingular $\boldsymbol{T}$ that is not block diagonal has been solved in [5]. However, apart from whether the $\boldsymbol{T}$ matrix is block diagonal or not, the two problems mentioned above are similar, yet different.

According to the partition

$$\boldsymbol{P} = \begin{bmatrix}\boldsymbol{P}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{P}_4\end{bmatrix} \quad \boldsymbol{P}_i = \boldsymbol{T}_i\boldsymbol{T}_i^t, \qquad i = 1, 4 \tag{18}$$

the weighted Gramians $\boldsymbol{K}_{o1}$, $\boldsymbol{K}_{c2}$, $\boldsymbol{K}_{oB}$, and $\boldsymbol{K}_{cC}$ can now be represented as

$$\begin{aligned}\boldsymbol{K}_{o1} &= \begin{bmatrix}\boldsymbol{K}_{o1}^{(1)} & \boldsymbol{K}_{o1}^{(2)} \\ \boldsymbol{K}_{o1}^{(3)} & \boldsymbol{K}_{o1}^{(4)}\end{bmatrix} \quad \boldsymbol{K}_{c2} = \begin{bmatrix}\boldsymbol{K}_{c2}^{(1)} & \boldsymbol{K}_{c2}^{(2)} \\ \boldsymbol{K}_{c2}^{(3)} & \boldsymbol{K}_{c2}^{(4)}\end{bmatrix} \\ \boldsymbol{K}_{oB} &= \begin{bmatrix}\boldsymbol{K}_{oB}^{(1)} & \boldsymbol{K}_{oB}^{(2)} \\ \boldsymbol{K}_{oB}^{(3)} & \boldsymbol{K}_{oB}^{(4)}\end{bmatrix} \quad \boldsymbol{K}_{cC} = \begin{bmatrix}\boldsymbol{K}_{cC}^{(1)} & \boldsymbol{K}_{cC}^{(2)} \\ \boldsymbol{K}_{cC}^{(3)} & \boldsymbol{K}_{cC}^{(4)}\end{bmatrix}.\end{aligned} \tag{19a}$$

If we introduce

$$\begin{aligned}\hat{\boldsymbol{K}}_{o1} &= \begin{bmatrix}\boldsymbol{K}_{o1}^{(1)} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{K}_{o1}^{(4)}\end{bmatrix} \quad \hat{\boldsymbol{K}}_{c2} = \begin{bmatrix}\boldsymbol{K}_{c2}^{(1)} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{K}_{c2}^{(4)}\end{bmatrix} \\ \hat{\boldsymbol{K}}_{oB} &= \begin{bmatrix}\boldsymbol{K}_{oB}^{(1)} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{K}_{oB}^{(4)}\end{bmatrix} \quad \hat{\boldsymbol{K}}_{cC} = \begin{bmatrix}\boldsymbol{K}_{cC}^{(1)} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{K}_{cC}^{(4)}\end{bmatrix}\end{aligned} \tag{19b}$$

$J(\boldsymbol{P})$ and $L(\boldsymbol{P})$ in (17) can be expressed as

$$\begin{aligned}J(\boldsymbol{P}) &= \operatorname{tr}[\hat{\boldsymbol{K}}_{o1}\boldsymbol{P}]\operatorname{tr}[\hat{\boldsymbol{K}}_{c2}\boldsymbol{P}^{-1}], \\ L(\boldsymbol{P}) &= \operatorname{tr}[\hat{\boldsymbol{K}}_{oB}\boldsymbol{P}] + \operatorname{tr}[\hat{\boldsymbol{K}}_{cC}\boldsymbol{P}^{-1}].\end{aligned} \tag{20}$$

Hence it suffices to deal with the matrices $\hat{\boldsymbol{K}}$ instead of $\boldsymbol{K}$. To make the exposition simple, we omit the hat and write $\boldsymbol{K}$ for $\hat{\boldsymbol{K}}$ in the following.

First, we minimize $J(\boldsymbol{P})$ and then minimize (17) subject to the minimization of $J(\boldsymbol{P})$. Using the formula for evaluating the matrix gradient [19, p. 275]

$$\begin{aligned}\frac{\partial[\operatorname{tr}(\boldsymbol{M}\boldsymbol{X})]}{\partial\boldsymbol{X}} &= \boldsymbol{M}^t \\ \frac{\partial[\operatorname{tr}(\boldsymbol{M}\boldsymbol{X}^{-1})]}{\partial\boldsymbol{X}} &= -(\boldsymbol{X}^{-1}\boldsymbol{M}\boldsymbol{X}^{-1})^t\end{aligned} \tag{21}$$

we obtain the equation for extrema of $J(\boldsymbol{P})$

$$\begin{aligned}\frac{\partial J(\boldsymbol{P})}{\partial\boldsymbol{P}} = {}&\operatorname{tr}[\boldsymbol{K}_{c2}\boldsymbol{P}^{-1}]\boldsymbol{K}_{o1} \\ &- \operatorname{tr}[\boldsymbol{K}_{o1}\boldsymbol{P}]\boldsymbol{P}^{-1}\boldsymbol{K}_{c2}\boldsymbol{P}^{-1} = \boldsymbol{0}.\end{aligned} \tag{22}$$

All the solutions of this equation take the form

$$\boldsymbol{P} = \rho\boldsymbol{P}_b \tag{23}$$

where $\boldsymbol{P}_b$ is the unique solution of the equation

$$\boldsymbol{P}\boldsymbol{K}_{o1}\boldsymbol{P} = \boldsymbol{K}_{c2}$$

and ρ is an arbitrary positive number. Moreover, $J(\boldsymbol{P})$ has the single extremum

$$J^o = J(\rho\boldsymbol{P}_b) = (\mathrm{tr}[\boldsymbol{K}_{o1}\boldsymbol{P}_b])^2 = (\mathrm{tr}[\boldsymbol{K}_{c2}\boldsymbol{P}_b^{-1}])^2 \tag{24}$$

which is independent of ρ.[1] Noting that $\boldsymbol{PWP} = \boldsymbol{M}$ has the unique solution [5]

$$\boldsymbol{P} = \boldsymbol{W}^{-(1/2)}[\boldsymbol{W}^{1/2}\boldsymbol{M}\boldsymbol{W}^{1/2}]^{1/2}\boldsymbol{W}^{-(1/2)} \tag{25}$$

where $\boldsymbol{W} > 0$ and $\boldsymbol{M} \geq 0$ are symmetric, the $\boldsymbol{P}_b$ matrix is given by

$$\boldsymbol{P}_b = (\boldsymbol{K}_{o1})^{-(1/2)}[(\boldsymbol{K}_{o1})^{1/2}\boldsymbol{K}_{c2}(\boldsymbol{K}_{o1})^{1/2}]^{1/2}(\boldsymbol{K}_{o1})^{-(1/2)} \tag{26}$$

and J^o is described by

$$J^o = (\mathrm{tr}[\boldsymbol{K}_{c2}\boldsymbol{K}_{o1}]^{1/2})^2. \tag{27}$$

Theorem 1: If $\boldsymbol{K}_{o1} = \boldsymbol{K}_{o1}^{(1)} \oplus \boldsymbol{K}_{o1}^{(4)}$ and $\boldsymbol{K}_{c2} = \boldsymbol{K}_{c2}^{(1)} \oplus \boldsymbol{K}_{c2}^{(4)}$ are $(m+n) \times (m+n)$ real symmetric positive-definite matrices, then the extremum J^o in (27) is really the minimum of $J(\boldsymbol{P})$. Furthermore, J^o can be expressed in terms of the square roots of the eigenvalues $\{\sigma_1^2, \sigma_2^2, \cdots, \sigma_{m+n}^2\}$ of $\boldsymbol{K}_{c2}\boldsymbol{K}_{o1}$ as follows:

$$J^o = J_{\min} = \left(\sum_{i=1}^{m+n} \sigma_i\right)^2. \tag{28}$$

Proof: The proof relies on the following inequality [20, p. 556]. If $\boldsymbol{D}$ is a real symmetric positive-definite matrix and if $\boldsymbol{Q}$ is any nonsingular real matrix, then

$$\mathrm{tr}[\boldsymbol{QDQ}^t]\,\mathrm{tr}[\boldsymbol{Q}^{-t}\boldsymbol{DQ}^{-1}] \geq (\mathrm{tr}\,\boldsymbol{D})^2 \tag{29}$$

where the equality sign holds if and only if

$$\rho\boldsymbol{QQ}^t = \boldsymbol{I} \tag{30}$$

for some positive real number ρ.

Choosing the above $\boldsymbol{D}$ and $\boldsymbol{Q}$ matrices as

$$\begin{aligned}\boldsymbol{D} &= [(\boldsymbol{K}_{o1})^{1/2}\boldsymbol{K}_{c2}(\boldsymbol{K}_{o1})^{1/2}]^{1/2}\\ \boldsymbol{Q} &= \boldsymbol{T}^t(\boldsymbol{K}_{o1})^{1/2}[(\boldsymbol{K}_{o1})^{1/2}\boldsymbol{K}_{c2}(\boldsymbol{K}_{o1})^{1/2}]^{-(1/4)}\end{aligned} \tag{31}$$

inequality (29) can be written as

$$\mathrm{tr}[\boldsymbol{K}_{o1}\boldsymbol{P}]\mathrm{tr}[\boldsymbol{K}_{c2}\boldsymbol{P}^{-1}] \geq (\mathrm{tr}[\boldsymbol{K}_{c2}\boldsymbol{K}_{o1}]^{1/2})^2 \tag{32}$$

where $\boldsymbol{P} = \boldsymbol{TT}^t$. On the other hand, taking $\boldsymbol{P} = \boldsymbol{TT}^t$ and $\boldsymbol{P}_b = \boldsymbol{T}_b\boldsymbol{T}_b^t$ into account and using (26), it is clear that (23) is equivalent to

$$\boldsymbol{T} = \sqrt{\rho}\boldsymbol{T}_b \tag{33}$$

where

$$\begin{aligned}\boldsymbol{T}_b &= (\boldsymbol{K}_{o1})^{-(1/2)}[(\boldsymbol{K}_{o1})^{1/2}\boldsymbol{K}_{c2}(\boldsymbol{K}_{o1})^{1/2}]^{1/4}\boldsymbol{U}\\ \boldsymbol{U} &= \boldsymbol{U}_1 \oplus \boldsymbol{U}_4\end{aligned}$$

and $\boldsymbol{U}_1(\boldsymbol{U}_4)$ is an arbitrary $m \times m$ $(n \times n)$ orthogonal matrix. Substituting (33) into (31) gives

$$\boldsymbol{QQ}^t = \rho\boldsymbol{I}_{m+n}. \tag{34}$$

This implies that equality in (32) holds, that is, the extremum J^o in (27) is actually the minimum of $J(\boldsymbol{P})$.

Let $\sigma_1^2, \sigma_2^2, \cdots, \sigma_{m+n}^2$ be the eigenvalues of $\boldsymbol{K}_{c2}\boldsymbol{K}_{o1}$. Then there exists a nonsingular matrix $\boldsymbol{R}$ such that $\boldsymbol{K}_{c2}\boldsymbol{K}_{o1} = \boldsymbol{R}^{-1}\,\mathrm{diag}(\sigma_1^2, \sigma_2^2, \cdots, \sigma_{m+n}^2)\,\boldsymbol{R}$. Hence

$$\begin{aligned}J^o &= J_{\min}\\ &= (\mathrm{tr}[\boldsymbol{K}_{c2}\boldsymbol{K}_{o1}]^{1/2})^2\\ &= (\mathrm{tr}[\mathrm{diag}(\sigma_1, \sigma_2, \cdots, \sigma_{m+n})])^2\\ &= \left(\sum_{i=0}^{m+n} \sigma_i\right)^2.\end{aligned} \tag{35}$$

This completes the proof of Theorem 1.

From (16), (19), and (33) we obtain

$$\rho^2\overline{\boldsymbol{K}}_{c2}^{(i)} = \overline{\boldsymbol{K}}_{o1}^{(i)}, \qquad i = 1, 4. \tag{36}$$

This shows that the weighted Gramians $\overline{\boldsymbol{K}}_{o1}$ and $\overline{\boldsymbol{K}}_{c2}$ are block balanced [21] when $\rho = 1$.

Remark 2: If (12b) can be derived from (36), then (12a) becomes an equality. However, unlike the 1-D case [2], the derivation is impossible in the 2-D case.

It turns out that the minimization of $J(\boldsymbol{P})$ forms a family of a matrix $\boldsymbol{P}$ parameterized by $\rho > 0$. We now proceed to determine ρ that minimizes $\overline{M}_{1/2}(\boldsymbol{P})$ in (17).

Theorem 2: The optimal solution $\boldsymbol{P} = \boldsymbol{P}_1 \oplus \boldsymbol{P}_4$ that minimizes the weighted sensitivity measure $\overline{M}_{1/2}(\boldsymbol{P})$ in (17) subject to the minimization of $J(\boldsymbol{P})$ is given by

$$\boldsymbol{P} = \sqrt{\frac{\mathrm{tr}[\boldsymbol{K}_{cC}\boldsymbol{P}_b^{-1}]}{\mathrm{tr}[\boldsymbol{K}_{oB}\boldsymbol{P}_b]}}\boldsymbol{P}_b \tag{37a}$$

or equivalently

$$\boldsymbol{T} = \left(\frac{\mathrm{tr}[\boldsymbol{K}_{cC}\boldsymbol{P}_b^{-1}]}{\mathrm{tr}[\boldsymbol{K}_{oB}\boldsymbol{P}_b]}\right)^{1/4}\boldsymbol{T}_b\boldsymbol{U} \tag{37b}$$

where $\boldsymbol{U} = \boldsymbol{U}_1 \oplus \boldsymbol{U}_2$ and $\boldsymbol{U}_1$ $(\boldsymbol{U}_4)$ is an arbitrary $m \times m$ $(n \times n)$ orthogonal matrix. The minimum of $\overline{M}_{1/2}(\boldsymbol{P})$ is

$$\overline{M}_{1/2}(\boldsymbol{P}) = \left(\sum_{i=1}^{m+n} \sigma_i\right)^2 + 2\sqrt{\mathrm{tr}[\boldsymbol{K}_{oB}\boldsymbol{P}_b]\,\mathrm{tr}[\boldsymbol{K}_{cC}\boldsymbol{P}_b^{-1}]}. \tag{38}$$

[1] Suppose $\boldsymbol{P}$ is a solution of (22). Since $\mathrm{tr}[K_{c2}\boldsymbol{P}^{-1}]$ and $\mathrm{tr}[K_{o1}\boldsymbol{P}]$ are positive, we can take $\rho > 0$ such that $\rho^2 = \mathrm{tr}[K_{o1}\boldsymbol{P}]/\mathrm{tr}[K_{c2}\boldsymbol{P}^{-1}]$. Then $\boldsymbol{P}_b = (1/\rho)P$ satisfies $\boldsymbol{P}_b K_{o1}\boldsymbol{P}_b = K_{c2}$. Indeed, $\boldsymbol{P}_b^{-1} = \rho\boldsymbol{P}^{-1}$ and

$$\begin{aligned}\mathbf{0} &= \mathrm{tr}[K_{c2}\boldsymbol{P}^{-1}]K_{o1} - \mathrm{tr}[K_{o1}\boldsymbol{P}]\boldsymbol{P}^{-1}K_{c2}\boldsymbol{P}^{-1}\\ &= \mathrm{tr}[K_{c2}\boldsymbol{P}^{-1}](K_{o1} - \rho^2\boldsymbol{P}^{-1}K_{c2}\boldsymbol{P}^{-1})\\ &= \mathrm{tr}[K_{c2}\boldsymbol{P}^{-1}](K_{o1} - \boldsymbol{P}_b^{-1}K_{c2}\boldsymbol{P}_b^{-1}).\end{aligned}$$

Thus, the solutions of (22) are exhausted by the solution of the form (23).

Proof: Substituting (23) into (17) gives

$$\overline{M}_{1/2}(\rho \boldsymbol{P}_b) = J_{\min} + \rho \operatorname{tr}[\boldsymbol{K}_{oB}\boldsymbol{P}_b] + \rho^{-1} \operatorname{tr}[\boldsymbol{K}_{cC}\boldsymbol{P}_b^{-1}]. \tag{39}$$

Here, the arithmetic-geometric inequality says that

$$\begin{aligned}\rho \operatorname{tr}[\boldsymbol{K}_{oB}\boldsymbol{P}_b] + \rho^{-1} \operatorname{tr}[\boldsymbol{K}_{cC}\boldsymbol{P}_b^{-1}] \\ \geq 2\sqrt{\operatorname{tr}[\boldsymbol{K}_{oB}\boldsymbol{P}_b] \operatorname{tr}[\boldsymbol{K}_{cC}\boldsymbol{P}_b^{-1}]}\end{aligned} \tag{40a}$$

where the equality is valid if and only if

$$\rho \operatorname{tr}[\boldsymbol{K}_{oB}\boldsymbol{P}_b] = \rho^{-1} \operatorname{tr}[\boldsymbol{K}_{cC}\boldsymbol{P}_b^{-1}]$$

or equivalently

$$\rho = \sqrt{\frac{\operatorname{tr}[\boldsymbol{K}_{cC}\boldsymbol{P}_b^{-1}]}{\operatorname{tr}[\boldsymbol{K}_{oB}\boldsymbol{P}_b]}}. \tag{40b}$$

Substituting (40b) into (23) yields (37a). Substituting (28) into (39) with (40a) produces (38).

This completes the proof of Theorem 2.

The next theorem describes the relation between the second term in (38) and the minimum of $L(\boldsymbol{P})$.

Theorem 3: If $\boldsymbol{K}_{oB} = \boldsymbol{K}_{oB}^{(1)} \oplus \boldsymbol{K}_{oB}^{(4)}$ and $\boldsymbol{K}_{cC} = \boldsymbol{K}_{cC}^{(1)} \oplus \boldsymbol{K}_{cC}^{(4)}$ are $(m+n) \times (m+n)$ real symmetric positive-definite matrices, then

$$\sqrt{\operatorname{tr}[\boldsymbol{K}_{oB}\boldsymbol{P}_b] \operatorname{tr}[\boldsymbol{K}_{cC}\boldsymbol{P}_b^{-1}]} \geq \sum_{i=1}^{m+n} \lambda_i \tag{41}$$

where $\lambda_1, \lambda_2, \cdots, \lambda_{m+n}$ are the square roots of the eigenvalues of $\boldsymbol{K}_{cC}\boldsymbol{K}_{oB}$. The equality in (41) holds if and only if the system satisfies

$$\boldsymbol{K}_{oB} = \alpha \boldsymbol{K}_{o1}, \quad \boldsymbol{K}_{cC} = \beta \boldsymbol{K}_{c2} \tag{42}$$

where α and β are some positive real numbers, $\boldsymbol{K}_{o1} = \boldsymbol{K}_{o1}^{(1)} \oplus \boldsymbol{K}_{o1}^{(4)}$, and $\boldsymbol{K}_{c2} = \boldsymbol{K}_{c2}^{(1)} \oplus \boldsymbol{K}_{c2}^{(4)}$.

Proof: To minimize $L(\boldsymbol{P})$ in (20), we carry out computations similar to those done in (22)–(27). The result is that $L(\boldsymbol{P})$ has the extremum

$$L^o = L(\boldsymbol{P}_c) = 2 \operatorname{tr}[\boldsymbol{K}_{oB}\boldsymbol{P}_c] = 2 \operatorname{tr}[\boldsymbol{K}_{cC}\boldsymbol{P}_c^{-1}] \tag{43}$$

at the matrix $\boldsymbol{P}_c$, which is the unique solution of the equation

$$\boldsymbol{P}\boldsymbol{K}_{oB}\boldsymbol{P} = \boldsymbol{K}_{cC}.$$

Since $\boldsymbol{P}_c$ is solved as

$$\begin{aligned}\boldsymbol{P}_c = &(\boldsymbol{K}_{oB})^{-(1/2)}[(\boldsymbol{K}_{oB})^{1/2}\boldsymbol{K}_{cC}(\boldsymbol{K}_{oB})^{1/2}]^{1/2} \\ &\cdot (\boldsymbol{K}_{oB})^{-(1/2)}\end{aligned} \tag{44}$$

we obtain

$$L^o = 2 \operatorname{tr}[\boldsymbol{K}_{cC}\boldsymbol{K}_{oB}]^{1/2}. \tag{45}$$

By an argument similar to those in Theorem 1, it can be shown that the extremum L^o in (45) is really the minimum of $L(\boldsymbol{P})$ and is expressed in the form

$$L^o = L_{\min} = 2\left(\sum_{i=1}^{m+n} \lambda_i\right). \tag{46}$$

Hence, the inequality (41) is proved.

(Necessity) Assume that the equality in (41) holds. Then, for a positive number ρ

$$\rho \boldsymbol{P}_b = \boldsymbol{P}_c \tag{47a}$$

must hold where $\boldsymbol{P}_b$ and $\boldsymbol{P}_c$ satisfy

$$\alpha \boldsymbol{P}_b \boldsymbol{K}_{o1} \boldsymbol{P}_b = \alpha \boldsymbol{K}_{c2}, \qquad \boldsymbol{P}_c \boldsymbol{K}_{oB} \boldsymbol{P}_c = \boldsymbol{K}_{cC} \tag{47b}$$

and α is any positive real number. Using (47a) enables one to change (47b) to

$$\alpha \boldsymbol{P}_c \boldsymbol{K}_{o1} \boldsymbol{P}_c = \alpha \rho^2 \boldsymbol{K}_{c2}, \qquad \boldsymbol{P}_b \boldsymbol{K}_{oB} \boldsymbol{P}_b = \frac{1}{\rho^2} \boldsymbol{K}_{cC} \tag{48}$$

Since $\boldsymbol{P}_b$ and $\boldsymbol{P}_c$ are the unique solutions, comparing (47b) with (48) concludes that

$$\boldsymbol{K}_{oB} = \alpha \boldsymbol{K}_{o1}, \qquad \boldsymbol{K}_{cC} = \beta \boldsymbol{K}_{c2} \tag{49}$$

where $\beta = \alpha \rho^2$.

(Sufficiency) Assume that (42) holds. Substituting (42) into (26), we obtain

$$\boldsymbol{P}_b = \sqrt{\frac{\alpha}{\beta}} \boldsymbol{P}_c \tag{50}$$

where $\boldsymbol{P}_c$ is given by (44). It is obvious that the equality in (41) holds.

This completes the proof of Theorem 3.

It should be noted that

$$\begin{aligned}\min_{\boldsymbol{P}} \overline{M}_{1/2}(\boldsymbol{P}) &\geq \min_{\boldsymbol{P}} J(\boldsymbol{P}) + \min_{\boldsymbol{P}} L(\boldsymbol{P}) \\ &= J_{\min} + L_{\min}.\end{aligned} \tag{51}$$

Corollary 1: The relation (42) holds provided

$$\begin{aligned}|W_B(z_1, z_2)| &= \sqrt{\alpha}|W_1(z_1, z_2)| \\ |W_C(z_1, z_2)| &= \sqrt{\beta}|W_2(z_1, z_2)|.\end{aligned} \tag{52}$$

Corollary 2: If (42) holds, then (37) is changed to

$$\boldsymbol{P} = \sqrt{\frac{\beta}{\alpha}} \boldsymbol{P}_b \tag{53a}$$

or equivalently

$$\boldsymbol{T} = \left(\frac{\beta}{\alpha}\right)^{1/4} \boldsymbol{T}_b \boldsymbol{U} \tag{53b}$$

and the equality sign in (41) holds. Moreover, (38) becomes

$$\overline{M}_{1/2}(\boldsymbol{P}) = \left(\sum_{i=1}^{m+n} \sigma_i\right)\left(\sum_{i=1}^{m+n} \sigma_i + 2\sqrt{\alpha\beta}\right). \tag{54}$$

The optimal filter structures that minimize $\overline{M}_{1/2}(\boldsymbol{P})$ (17) subject to the minimization of $J(\boldsymbol{P})$ can readily be synthesized by substituting (37b) into (4).

Remark 3: Notice that (53) can be considered to be an extension of the 1-D closed-form solution reported in [5] to the 2-D case. In Corollary 2 it is mentioned that (53) can be derived from (37) as a special case for the system such that (42) is satisfied. In other words, unlike the solution given by (53), (37) can be applied to the general systems where (42) is not always satisfied. It should be pointed out that neither the 1-D version of the closed-form solution (37) stated in Theorem 2 nor the 1-D counterpart of arguments stated in Theorem 3 has been reported in [5].

IV. Filter Synthesis with Minimum Weighted L_2-Sensitivity

In this section, we synthesize the 2-D filter structures that minimize a weighted L_2-sensitivity measure defined by

$$m_2 = \left\|\frac{\delta H(z_1,z_2)}{\delta \boldsymbol{A}}\right\|_2^2 + \left\|\frac{\delta H(z_1,z_2)}{\delta \boldsymbol{b}}\right\|_2^2 + \left\|\frac{\delta H(z_1,z_2)}{\delta \boldsymbol{c}^t}\right\|_2^2 \tag{55}$$

instead of (10). Referring to (11) and (14), we can write (55) as

$$\begin{aligned} m_2 &= \mathrm{tr}[\boldsymbol{K}_A] + \mathrm{tr}[\boldsymbol{K}_{oB}] + \mathrm{tr}[\boldsymbol{K}_{cC}] \\ &= \mathrm{tr}\left[\sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \boldsymbol{M}_A(i,j)\boldsymbol{M}_A^t(i,j)\right] + \mathrm{tr}[\boldsymbol{K}_{oB}] \\ &\quad + \mathrm{tr}[\boldsymbol{K}_{cC}] \end{aligned} \tag{56}$$

where $\boldsymbol{K}_A$ is obtained by the general expression of (15) with

$$\boldsymbol{Y}(z_1,z_2) = W_A(z_1,z_2)\boldsymbol{G}^t(z_1,z_2)\boldsymbol{F}^t(z_1,z_2)$$

and $\boldsymbol{M}_A(i,j)$ is derived from

$$\begin{aligned} \boldsymbol{M}_A(i,j) &= \sum\sum_{(0,0)\le(k,r)<(i,j)} w_A(k,r)\boldsymbol{M}(i-k,j-r) \\ \boldsymbol{M}(i,j) &= \sum\sum_{(0,0)\le(k,r)<(i,j)} \boldsymbol{g}^t(k,r)\boldsymbol{f}^t(i-k,j-r). \\ \boldsymbol{f}(i,j) &= \boldsymbol{A}^{(i-1,j)}\begin{bmatrix}\boldsymbol{I}_m & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\boldsymbol{b} + \boldsymbol{A}^{(i,j-1)}\begin{bmatrix}\boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}_n\end{bmatrix}\boldsymbol{b} \\ \boldsymbol{g}(i,j) &= \boldsymbol{c}\boldsymbol{A}^{(i-1,j)}\begin{bmatrix}\boldsymbol{I}_m & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix} + \boldsymbol{c}\boldsymbol{A}^{(i,j-1)}\begin{bmatrix}\boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}_n\end{bmatrix} \\ \boldsymbol{A}^{(1,0)} &= \begin{bmatrix}\boldsymbol{I}_m & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\boldsymbol{A}, \\ \boldsymbol{A}^{(0,1)} &= \begin{bmatrix}\boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}_n\end{bmatrix}\boldsymbol{A} \\ \boldsymbol{A}^{(0,0)} &= \boldsymbol{I}_{m+n}, \quad \boldsymbol{A}^{(-i,j)} = \boldsymbol{A}^{(i,-j)} = \boldsymbol{0}, \qquad (i,j\ge 1) \\ \boldsymbol{A}^{(i,j)} &= \boldsymbol{A}^{(1,0)}\boldsymbol{A}^{(i-1,j)} + \boldsymbol{A}^{(0,1)}\boldsymbol{A}^{(i,j-1)} \\ &= \boldsymbol{A}^{(i-1,j)}\boldsymbol{A}^{(1,0)} + \boldsymbol{A}^{(i,j-1)}\boldsymbol{A}^{(0,1)}, \qquad (i,j) > (0,0) \end{aligned}$$

with $w_A(k,r)$ being the unit-sample response of $W_A(z_1,z_2)$. Applying the coordinate transformation defined by (3) to the original filter, (56) becomes

$$\begin{aligned} \overline{m}_2(\boldsymbol{P}) &= \mathrm{tr}\left[\sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \boldsymbol{M}_A(i,j)\boldsymbol{P}^{-1}\boldsymbol{M}_A^t(i,j)\boldsymbol{P}\right] \\ &\quad + \mathrm{tr}[\boldsymbol{K}_{oB}\boldsymbol{P}] + \mathrm{tr}[\boldsymbol{K}_{cC}\boldsymbol{P}^{-1}] \end{aligned} \tag{57}$$

where $\boldsymbol{P}$ is as defined in (17). According to the partition of (18), $\boldsymbol{M}_A(i,j)$ can be represented as

$$\boldsymbol{M}_A(i,j) = \begin{bmatrix}\boldsymbol{M}_A^{(1)}(i,j) & \boldsymbol{M}_A^{(2)}(i,j)\\ \boldsymbol{M}_A^{(3)}(i,j) & \boldsymbol{M}_A^{(4)}(i,j)\end{bmatrix}. \tag{58}$$

Taking (21) into account, it follows from (57) that

$$\begin{aligned} \frac{\partial \overline{m}_2(\boldsymbol{P})}{\partial \boldsymbol{P}_1} &= \boldsymbol{F}_1(\boldsymbol{P}) - \boldsymbol{P}_1^{-1}\boldsymbol{F}_2(\boldsymbol{P})\boldsymbol{P}_1^{-1} \\ \frac{\partial \overline{m}_2(\boldsymbol{P})}{\partial \boldsymbol{P}_4} &= \boldsymbol{F}_3(\boldsymbol{P}) - \boldsymbol{P}_4^{-1}\boldsymbol{F}_4(\boldsymbol{P})\boldsymbol{P}_4^{-1} \end{aligned} \tag{59}$$

where

$$\begin{aligned} \boldsymbol{F}_1(\boldsymbol{P}) &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}[\boldsymbol{M}_A^{(1)}(i,j)\boldsymbol{P}_1^{-1}\boldsymbol{M}_A^{(1)}(i,j)^t \\ &\quad + \boldsymbol{M}_A^{(2)}(i,j)\boldsymbol{P}_4^{-1}\boldsymbol{M}_A^{(2)}(i,j)^t] + \boldsymbol{K}_{oB}^{(1)} \\ \boldsymbol{F}_2(\boldsymbol{P}) &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}[\boldsymbol{M}_A^{(1)}(i,j)^t\boldsymbol{P}_1\boldsymbol{M}_A^{(1)}(i,j) \\ &\quad + \boldsymbol{M}_A^{(3)}(i,j)^t\boldsymbol{P}_4\boldsymbol{M}_A^{(3)}(i,j)] + \boldsymbol{K}_{cC}^{(1)} \\ \boldsymbol{F}_3(\boldsymbol{P}) &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}[\boldsymbol{M}_A^{(3)}(i,j)\boldsymbol{P}_1^{-1}\boldsymbol{M}_A^{(3)}(i,j)^t \\ &\quad + \boldsymbol{M}_A^{(4)}(i,j)\boldsymbol{P}_4^{-1}\boldsymbol{M}_A^{(4)}(i,j)^t] + \boldsymbol{K}_{oB}^{(4)} \\ \boldsymbol{F}_4(\boldsymbol{P}) &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}[\boldsymbol{M}_A^{(2)}(i,j)^t\boldsymbol{P}_1\boldsymbol{M}_A^{(2)}(i,j) \\ &\quad + \boldsymbol{M}_A^{(4)}(i,j)^t\boldsymbol{P}_4\boldsymbol{M}_A^{(4)}(i,j)] + \boldsymbol{K}_{cC}^{(4)}. \end{aligned}$$

Letting the two equations in (59) be null yields

$$\begin{aligned} \boldsymbol{P}_1\boldsymbol{F}_1(\boldsymbol{P})\boldsymbol{P}_1 &= \boldsymbol{F}_2(\boldsymbol{P}) \\ \boldsymbol{P}_4\boldsymbol{F}_3(\boldsymbol{P})\boldsymbol{P}_4 &= \boldsymbol{F}_4(\boldsymbol{P}) \end{aligned} \tag{60}$$

respectively. From (25) and (60) it follows that the values $\boldsymbol{P}_1^{(i+1)}$ and $\boldsymbol{P}_4^{(i+1)}$ satisfying

$$\begin{aligned} \boldsymbol{P}_1^{(i+1)}\boldsymbol{F}_1(\boldsymbol{P}^{(i)})\boldsymbol{P}_1^{(i+1)} &= \boldsymbol{F}_2(\boldsymbol{P}^{(i)}) \\ \boldsymbol{P}_4^{(i+1)}\boldsymbol{F}_3(\boldsymbol{P}^{(i)})\boldsymbol{P}_4^{(i+1)} &= \boldsymbol{F}_4(\boldsymbol{P}^{(i)}) \end{aligned} \tag{61}$$

respectively, are given by

$$\begin{aligned} \boldsymbol{P}_1^{(i+1)} &= \boldsymbol{F}_1^{-(1/2)}(\boldsymbol{P}^{(i)})[\boldsymbol{F}_1^{1/2}(\boldsymbol{P}^{(i)})\boldsymbol{F}_2(\boldsymbol{P}^{(i)})\boldsymbol{F}_1^{1/2}(\boldsymbol{P}^{(i)})]^{1/2} \\ &\quad \cdot \boldsymbol{F}_1^{-(1/2)}(\boldsymbol{P}^{(i)}) \\ \boldsymbol{P}_4^{(i+1)} &= \boldsymbol{F}_3^{-(1/2)}(\boldsymbol{P}^{(i)})[\boldsymbol{F}_3^{1/2}(\boldsymbol{P}^{(i)})\boldsymbol{F}_4(\boldsymbol{P}^{(i)})\boldsymbol{F}_3^{1/2}(\boldsymbol{P}^{(i)})]^{1/2} \\ &\quad \cdot \boldsymbol{F}_3^{-(1/2)}(\boldsymbol{P}^{(i)}) \end{aligned} \tag{62}$$

where $\boldsymbol{P}^{(i)}$ is the solution of the previous iteration. The initial estimate $\boldsymbol{P}^{(0)}$ in the above iteration is given by (37a). This iteration process continues until

$$|\overline{m}_2(\boldsymbol{P}^{(i+1)}) - \overline{m}_2(\boldsymbol{P}^{(i)})| < \varepsilon \tag{63}$$

where $\varepsilon > 0$ is a prescribed tolerance.

While the convergence of the iterative algorithm described in (62) remains to be proved, the algorithm was applied to quite a number of simulation examples and fast convergence was observed in all the cases. A sample of these examples will be illustrated in the next section. As a remark on this convergence issue, we note that an interesting iterative algorithm, based on the concept of gradient flow for frequency weighted sensitivity minimization of 1-D discrete-time systems, was proposed in [7] and extended to 2-D Fornasini–Marchesini model in [15]. Although the nonlinear setting in (62) differs from that of [7] and [15], the technique employed there to show the convergence of the algorithms appears worthwhile to analyze in order to show the convergence of (62) or similar algorithms.

Given the L_2-optimal matrix $\boldsymbol{P} = \boldsymbol{P}_1 \oplus \boldsymbol{P}_4$ which is positive-definite and symmetric, the corresponding L_2-optimal transformation matrix can be constructed as

$$\boldsymbol{T} = [\boldsymbol{P}_1^{1/2} \oplus \boldsymbol{P}_4^{1/2}][\boldsymbol{U}_1 \oplus \boldsymbol{U}_4] \tag{64}$$

where $\boldsymbol{U}_1$ and $\boldsymbol{U}_4$ are arbitrary $m \times m$ and $n \times n$ orthogonal matrices, respectively. It is possible to synthesize the L_2-optimal filter structures such that (57) is minimum by substituting (64) into (4).

Remark 4: As was shown in [22], the orthogonal matrices $\boldsymbol{U}_1$ and $\boldsymbol{U}_4$ in (64) can be used to obtain a state-space realization with more zero or one entries, which further reduces the L_2 sensitivity. An alternative approach to accomplish this is to use singular value decomposition (SVD) [23], [24] as follows.

Let us denote

$$\tilde{\boldsymbol{A}} = \tilde{\boldsymbol{T}}^{-1}\boldsymbol{A}\tilde{\boldsymbol{T}} = \begin{bmatrix} \tilde{\boldsymbol{A}}_1 & \tilde{\boldsymbol{A}}_2 \\ \tilde{\boldsymbol{A}}_3 & \tilde{\boldsymbol{A}}_4 \end{bmatrix}, \quad \tilde{\boldsymbol{T}} = \boldsymbol{P}_1^{1/2} \oplus \boldsymbol{P}_4^{1/2} \tag{65}$$

and apply SVD to $\tilde{\boldsymbol{A}}_2$

$$\tilde{\boldsymbol{A}}_2 = \boldsymbol{R}\boldsymbol{S}\boldsymbol{Q}^T \tag{66}$$

where $\boldsymbol{R}$ and $\boldsymbol{Q}$ are $m \times m$ and $n \times n$ orthogonal matrices, respectively, and

$$\boldsymbol{S}_2 = \left[\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & \boldsymbol{0} \\ & & \sigma^{r_2} & \\ \hline & \boldsymbol{0} & & \boldsymbol{0} \end{array}\right]$$

with r_2 being the rank of $\boldsymbol{A}_2$. Evidently, if we let $\boldsymbol{U}_1 = \boldsymbol{R}$, $\boldsymbol{U}_2 = \boldsymbol{Q}$, then $\overline{\boldsymbol{A}} = \boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T}$ has the form

$$\overline{\boldsymbol{A}} = \begin{bmatrix} \overline{\boldsymbol{A}}_1 & \boldsymbol{S}_2 \\ \overline{\boldsymbol{A}}_3 & \overline{\boldsymbol{A}}_4 \end{bmatrix} \tag{67}$$

where $\boldsymbol{S}_2$ has $(mn - r_2)$ zero entries. Alternatively, SVD may be applied to the matrix $\tilde{\boldsymbol{A}}_3$ to yield $(mn - r_3)$ zero entries where r_3 is the rank of $\boldsymbol{A}_3$.

V. Illustrative Examples

The frequency weighting functions $W_A(z_1, z_2)$, $W_B(z_1, z_2)$, and $W_C(z_1, z_2)$ can be either of 2-D finite impulse response (FIR) or infinite impulse response (IIR) digital filters. For simplicity, let these be given by the following 2-D lowpass filters:

$$W_A(z_1, z_2) = \sum_{i=0}^{20}\sum_{j=0}^{20} w_a(i,j) z_1^{-i} z_2^{-j}$$
$$W_B(z_1, z_2) = \sum_{i=0}^{20}\sum_{j=0}^{20} w_b(i,j) z_1^{-i} z_2^{-j}$$
$$W_C(z_1, z_2) = \frac{N(z_1, z_2)}{D_1(z_1)D_2(z_2)}$$

where

$$w_a(i,j) = 0.256322 \exp[-0.103203\{(i-4)^2 + (j-4)^2\}]$$
$$w_b(i,j) = 0.256322 \exp[-0.103203\{(i-5)^2 + (j-i)^2\}]$$
$$N(z_1, z_2) = \sum_{i=0}^{4}\sum_{j=0}^{4} b_{ij} z_1^{-i} z_2^{-j}$$
$$D(z_k) = 1.0 - 1.11425 z_k^{-1} + 0.75745 z_k^{-2} - 0.34255 z_k^{-3} + 0.10171 z_k^{-4}, \qquad k = 1, 2$$

$$\begin{bmatrix} b_{00} & b_{01} & \cdots & b_{04} \\ b_{10} & b_{11} & \cdots & b_{14} \\ \vdots & \vdots & \ddots & \vdots \\ b_{40} & b_{41} & \cdots & b_{44} \end{bmatrix} = \begin{bmatrix} 0.12814 & 0.64232 & 0.74979 & 0.64232 & 0.12814 \\ 0.64232 & 0.33077 & 0.68889 & 0.33077 & 0.64232 \\ 0.74979 & 0.68889 & 1.34339 & 0.68889 & 0.74979 \\ 0.64232 & 0.33077 & 0.68889 & 0.33077 & 0.64232 \\ 0.12814 & 0.64232 & 0.74979 & 0.64232 & 0.12814 \end{bmatrix} \times 10^{-2}.$$

A factorization of (8) is now assumed to be

$$W_1(z_1, z_2) = 1$$
$$W_2(z_1, z_2) = W_A(z_1, z_2)$$

Example 1: (2, 2)th-Order Filter

Consider the LSS model (1) specified by

$$\boldsymbol{A} = \begin{bmatrix} 1.88899 & -0.91219 & -1.0 & 0.0 \\ 1.0 & 0.0 & 0.0 & 0.0 \\ 0.02771 & -0.02580 & 1.88899 & 1.0 \\ -0.02580 & 0.02431 & -0.91219 & 0.0 \end{bmatrix}$$
$$\boldsymbol{b} = [0.219089 \quad 0.0 \quad -0.028889 \quad 0.091219]^t$$
$$\boldsymbol{c} = [0.28889 \quad -0.091219 \quad -0.219089 \quad 0.0]$$

where $m = n = 2$.

Applying Parseval's relation to (14) and (15), it follows that

$$\boldsymbol{K}_{o1} = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \boldsymbol{g}_{o1}^t(i,j)\boldsymbol{g}_{o1}(i,j)$$
$$\boldsymbol{K}_{c2} = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \boldsymbol{f}_{c2}(i,j)\boldsymbol{f}_{c2}^t(i,j) \tag{68}$$

where

$$\boldsymbol{g}_{o1}(i,j) = \sum\sum_{(0,0)\le(k,r)<(i,j)} w_1(k,r)\boldsymbol{g}(i-k,j-r)$$
$$\boldsymbol{f}_{c2}(i,j) = \sum\sum_{(0,0)\le(k,r)<(i,j)} w_2(k,r)\boldsymbol{f}(i-k,j-r)$$
$$W_k(z_1,z_2) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} w_k(i,j)z_1^{-i}z_2^{-j}, \qquad k=1,2.$$

Here, $\boldsymbol{f}(i,j)$ and $\boldsymbol{g}(i,j)$ are as defined in (56). From (56) and (68), the 2-D weighted Gramians $\boldsymbol{K}_A^{(i)}$, $\boldsymbol{K}_{o1}^{(i)}$, $\boldsymbol{K}_{c2}^{(i)}$, $\boldsymbol{K}_{oB}^{(i)}$, and $\boldsymbol{K}_{cC}^{(i)}$, $i = 1,4$ can be calculated by truncation $0 \le i \le 100$ and $0 \le j \le 100$[2] as

$$\boldsymbol{K}_A^{(1)} = \begin{bmatrix} 460405.02074980 & -406380.17548504 \\ -406380.17548504 & 364565.10231906 \end{bmatrix}$$
$$\boldsymbol{K}_A^{(4)} = \begin{bmatrix} 390597141.53364521 & 383598953.44964486 \\ 383598953.44964486 & 390423379.40672630 \end{bmatrix}$$
$$\boldsymbol{K}_{o1}^{(1)} = \begin{bmatrix} 11.33639752 & -10.35083304 \\ -10.35083304 & 9.66068618 \end{bmatrix}$$
$$\boldsymbol{K}_{o1}^{(4)} = \begin{bmatrix} 638.95921778 & 622.90731381 \\ 622.90731381 & 638.94448493 \end{bmatrix}$$
$$\boldsymbol{K}_{c2}^{(1)} = \begin{bmatrix} 2869.13457721 & 2822.98523270 \\ 2822.98523270 & 2869.02675750 \end{bmatrix}$$
$$\boldsymbol{K}_{c2}^{(4)} = \begin{bmatrix} 4.92296703 & -4.33804824 \\ -4.33804824 & 3.94010274 \end{bmatrix}$$
$$\boldsymbol{K}_{oB}^{(1)} = \begin{bmatrix} 23.29761522 & -20.35185194 \\ -20.35185194 & 18.37114758 \end{bmatrix}$$
$$\boldsymbol{K}_{oB}^{(4)} = \begin{bmatrix} 21436.38888158 & 21164.41125664 \\ 21164.41125664 & 21436.07017164 \end{bmatrix}$$
$$\boldsymbol{K}_{cC}^{(1)} = \begin{bmatrix} 50.99062638 & 50.04288786 \\ 50.04288786 & 50.98841828 \end{bmatrix}$$
$$\boldsymbol{K}_{cC}^{(4)} = \begin{bmatrix} 0.12967208 & -0.11585595 \\ -0.11585595 & 0.10657265 \end{bmatrix}.$$

Then, the original weighted sensitivities (14) and (56) become

$$M_{1/2} = J + L = 7507830.8616$$
$$m_2 = 781888507.4065$$

respectively, where $J = 7464814.5185$ and $L = 43016.3431$.

[2]This region was chosen according to the memory capacity of computers in the laboratory as well as the approximation accuracy in the truncation.

The suboptimal $\boldsymbol{P}_c = \boldsymbol{P}_1 \oplus \boldsymbol{P}_4$ matrix that minimizes $L(\boldsymbol{P})$ in (17) is calculated from (44) as

$$\boldsymbol{P}_1 = \begin{bmatrix} 7.84351089 & 8.18028539 \\ 8.18028539 & 8.86708460 \end{bmatrix}$$
$$\boldsymbol{P}_4 = \begin{bmatrix} 0.01483796 & -0.01394208 \\ -0.01394208 & 0.01349485 \end{bmatrix}$$

or equivalently

$$\boldsymbol{T}_1 = \begin{bmatrix} 2.06433731 & 1.89262314 \\ 1.76150446 & 2.40087206 \end{bmatrix}$$
$$\boldsymbol{T}_4 = \begin{bmatrix} 0.08854411 & -0.08365345 \\ -0.06955674 & 0.09304145 \end{bmatrix}.$$

In this case, from (17) and (57) we have

$$\overline{M}_{1/2} = J + L_{\min} = 8057.9652$$
$$\overline{m}_2 = 980226506$$

respectively, where $J = 7998.2396$ and $L_{\min} = 59.7256$.

The L_1/L_2-optimal $\boldsymbol{P}$ matrix which minimizes (17) subject to the minimization of $J(\boldsymbol{P})$ can be computed from (37a) as

$$\boldsymbol{P}_1 = \begin{bmatrix} 5.14975986 & 5.30489866 \\ 5.30489866 & 5.62128787 \end{bmatrix}$$
$$\boldsymbol{P}_4 = \begin{bmatrix} 0.01884643 & -0.01733058 \\ -0.01733058 & 0.01662501 \end{bmatrix}$$

or equivalently

$$\boldsymbol{T}_1 = \begin{bmatrix} 1.41877744 & 1.77110995 \\ 1.15269142 & 2.07185674 \end{bmatrix}$$
$$\boldsymbol{T}_4 = \begin{bmatrix} 0.12303452 & -0.06090101 \\ -0.10149961 & 0.07951626 \end{bmatrix}.$$

As a result, the L_1/L_2-optimal filter structure is synthesized from (4) as shown at the bottom of this page and it follows from (17) and (57) that

$$\overline{M}_{1/2} = J_{\min} + L = 6459.9130$$
$$\overline{m}_2 = 163012.2215$$

respectively, where $J_{\min} = 6391.6953$ and $L = 68.2177$.

Applying the iterative procedure (62) produces

$$\boldsymbol{P}_1 = \begin{bmatrix} 12.78717263 & 13.55220273 \\ 13.55220273 & 14.75846399 \end{bmatrix}$$
$$\boldsymbol{P}_4 = \begin{bmatrix} 0.00869668 & -0.00794560 \\ -0.00794560 & 0.00751555 \end{bmatrix}$$

$$\overline{\boldsymbol{A}} = \begin{bmatrix} 0.95926221 & -0.13460051 & -0.28387657 & 0.14051640 \\ 0.15109352 & 0.92972779 & 0.15793663 & -0.07817724 \\ 0.06626498 & -0.01761805 & 0.99244647 & 0.15503995 \\ -0.02334944 & 0.03626901 & -0.14459846 & 0.89654353 \end{bmatrix}$$
$$\overline{\boldsymbol{b}} = [0.50550229 \quad -0.28123959 \quad 0.90459477 \quad 2.30185664]^t$$
$$\overline{\boldsymbol{c}} = [0.30472326 \quad 0.32266325 \quad -0.02695551 \quad 0.01334274]$$

after 20 iterations or equivalently

$$\boldsymbol{T}_1 = \begin{bmatrix} 2.21652433 & 2.80609916 \\ 1.85565665 & 3.36377799 \end{bmatrix}$$
$$\boldsymbol{T}_4 = \begin{bmatrix} 0.07803397 & -0.05106247 \\ -0.06253108 & 0.06004513 \end{bmatrix}.$$

where truncation $0 \le i \le 100$ and $0 \le j \le 100$ was used to compute $F_k(\boldsymbol{P}),\ k = 1, 2, 3, 4$. Substituting the above coordinate transformation into (4) provides the L_2-optimal filter structure as shown at the bottom of this page and (17) and (57) were used to calculate

$$\overline{M}_{1/2} = J + L = 11546.8608$$
$$\overline{m}_2 = 84193.9719$$

respectively, where $J = 11481.6550$ and $L = 65.2058$.

Example 2: (3, 3)th Order Filter

Let the LSS model (1) be given by

$$\boldsymbol{A}_1 = \begin{bmatrix} 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 1.0 \\ 0.38315 & -1.38605 & 1.90670 \end{bmatrix}$$
$$\boldsymbol{A}_2 = \begin{bmatrix} -0.06280 & 0.06190 & 0.00654 \\ -0.02810 & 0.03956 & -0.02248 \\ 1.24452 & -0.57092 & 2.05865 \end{bmatrix}$$
$$\boldsymbol{A}_3 = \begin{bmatrix} -0.00003 & 0.00038 & -0.00053 \\ -0.00001 & 0.00018 & -0.00026 \\ -0.00008 & 0.00023 & -0.00017 \end{bmatrix}$$
$$\boldsymbol{A}_4 = \begin{bmatrix} 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 1.0 \\ 0.38238 & -1.38178 & 1.90253 \end{bmatrix}$$
$$\boldsymbol{b}_1 = \boldsymbol{b}_2 = [0.0 \quad 0.0 \quad 1.0]^t$$
$$\boldsymbol{c}_1 = [0.01141 \quad -0.00540 \quad 0.01956]$$
$$\boldsymbol{c}_2 = [0.01164 \quad -0.00545 \quad 0.01960]$$

where $m = n = 3$.

Using (56) and (68), the submatrices of the 2-D weighted Gramians $\boldsymbol{K}_A^{(i)}$, $\boldsymbol{K}_{o1}^{(i)}$, $\boldsymbol{K}_{c2}^{(i)}$, $\boldsymbol{K}_{oB}^{(i)}$, and $\boldsymbol{K}_{cC}^{(i)}$, $i = 1$, 4 can be calculated by truncation $0 \le i \le 100$ and $0 \le j \le 100$ as shown at the bottom of the next page. Then (14) and (56) becomes

$$M_{1/2} = J + L = 299673.7313$$
$$m_2 = 10441330.7603$$

respectively, where $J = 296773.3396$ and $L = 2900.3916$.

Applying (44), the suboptimal filter structures are realized from

$$\boldsymbol{P}_1 = \begin{bmatrix} 1994.65983573 & 1090.47959350 & 349.84385826 \\ 1090.47959350 & 810.29935629 & 464.59809968 \\ 349.84385826 & 464.59809968 & 447.86957497 \end{bmatrix}$$
$$\boldsymbol{P}_4 = \begin{bmatrix} 6.11907288 & 3.09433408 & 0.66823909 \\ 3.09433408 & 2.08175279 & 0.97490842 \\ 0.66823909 & 0.97490842 & 0.93506167 \end{bmatrix}$$

or equivalently,

$$\boldsymbol{T}_1 = \begin{bmatrix} 13.82537475 & 30.61643239 & 29.43047598 \\ 16.78176757 & 22.16623497 & 6.10980039 \\ 18.56399712 & 8.45750822 & -5.63188606 \end{bmatrix}$$
$$\boldsymbol{T}_4 = \begin{bmatrix} 0.59412325 & 1.56120892 & 1.82447723 \\ 0.72216789 & 1.15695186 & 0.47083831 \\ 0.79947389 & 0.46108297 & -0.28862721 \end{bmatrix}.$$

From (17) and (57), this gives

$$\overline{M}_{1/2} = J + L_{\min} = 364.6126$$
$$\overline{m}_2 = 2186.2190$$

respectively, where $J = 351.6035$ and $L_{\min} = 13.0091$.

Making use of (37a), the L_1/L_2-optimal filter structures that minimize (17) subject to the minimization of $J(\boldsymbol{P})$ are constructed from

$$\boldsymbol{P}_1 = \begin{bmatrix} 695.06511011 & 443.59300157 & 226.86622553 \\ 443.59300157 & 382.35826949 & 282.87830911 \\ 226.86622553 & 282.87830911 & 287.77432092 \end{bmatrix}$$
$$\boldsymbol{P}_4 = \begin{bmatrix} 2.71258093 & 1.73686609 & 0.89019654 \\ 1.73686609 & 1.50929838 & 1.12229674 \\ 0.89019654 & 1.12229674 & 1.14720037 \end{bmatrix}$$

or equivalently

$$\boldsymbol{T}_1 = \begin{bmatrix} 13.73007262 & 19.38689860 & 11.43233915 \\ 15.36067722 & 12.09880780 & -0.16344619 \\ 16.16489972 & 2.79936292 & -4.31670073 \end{bmatrix}$$
$$\boldsymbol{T}_4 = \begin{bmatrix} 0.86378782 & 1.21374840 & 0.70232925 \\ 0.96989606 & 0.75372228 & -0.02242183 \\ 1.02128137 & 0.16658800 & -0.27646548 \end{bmatrix}.$$

$$\overline{\boldsymbol{A}} = \begin{bmatrix} 0.96517195 & -0.16243289 & -0.11672716 & 0.07638182 \\ 0.12649366 & 0.92381805 & 0.06439353 & -0.04213668 \\ 0.13175457 & -0.04244315 & 0.97923151 & 0.12870234 \\ -0.06389460 & 0.11195028 & -0.16569878 & 0.90975849 \end{bmatrix}$$
$$\overline{\boldsymbol{b}} = [0.32772442 \quad -0.18079196 \quad 1.95851839 \quad 3.55877769]^t$$
$$\overline{\boldsymbol{c}} = [0.47106057 \quad 0.50381352 \quad -0.01709638 \quad 0.01118722]$$

Then, making use of (4), we get

$$\overline{\boldsymbol{A}}_1 = \begin{bmatrix} 0.83377989 & -0.21625751 & -0.01783948 \\ 0.27580984 & 0.51201201 & -0.32656076 \\ -0.12545803 & 0.44975172 & 0.56090810 \end{bmatrix}$$
$$\overline{\boldsymbol{A}}_2 = \begin{bmatrix} 0.30544369 & 0.16449039 & 0.03618117 \\ -0.38465345 & -0.20745137 & -0.04669960 \\ 0.28655056 & 0.15175330 & 0.03160229 \end{bmatrix}$$
$$\overline{\boldsymbol{A}}_3 = \begin{bmatrix} -0.00034161 & 0.00120968 & -0.00058481 \\ -0.00168438 & 0.00016443 & 0.00204681 \\ -0.00114298 & 0.00183361 & -0.00013724 \end{bmatrix}$$
$$\overline{\boldsymbol{A}}_4 = \begin{bmatrix} 0.83476698 & -0.21581510 & -0.01797649 \\ 0.27709299 & 0.51222077 & -0.32714155 \\ -0.12456645 & 0.45339646 & 0.55554225 \end{bmatrix}$$
$$\overline{\boldsymbol{b}}_1 = [0.11589425 \quad -0.14568253 \quad 0.10786033]^t$$
$$\overline{\boldsymbol{b}}_2 = [1.81737136 \quad -2.28750015 \quad 1.71803526]^t$$
$$\overline{\boldsymbol{c}}_1 = [0.38989791 \quad 0.21062649 \quad 0.04689093]$$
$$\overline{\boldsymbol{c}}_2 = [0.02478567 \quad 0.01328537 \quad 0.00287859]$$

and from (17) and (57) it follows that

$$\overline{M}_{1/2} = J_{\min} + L = 290.7477$$
$$\overline{m}_2 = 2806.3110$$

respectively, where $J_{\min} = 276.0360$ and $L = 14.7116$.

Applying the iterative procedure (62) provides

$$\boldsymbol{P}_1 = \begin{bmatrix} 2402.69261014 & 1326.43635801 & 446.26772332 \\ 1326.43635801 & 927.17620872 & 512.42934385 \\ 446.26772332 & 512.42934385 & 480.37920997 \end{bmatrix}$$
$$\boldsymbol{P}_4 = \begin{bmatrix} 5.13089004 & 2.66738136 & 0.67461188 \\ 2.66738136 & 1.78259209 & 0.86291979 \\ 0.67461188 & 0.86291979 & 0.81547874 \end{bmatrix}$$

after 20 iterations or equivalently

$$\boldsymbol{T}_1 = \begin{bmatrix} 17.53074796 & 34.34813573 & 30.25840475 \\ 19.02504678 & 22.70962425 & 7.03539409 \\ 19.82782611 & 7.62476990 & -5.39438650 \end{bmatrix}$$
$$\boldsymbol{T}_4 = \begin{bmatrix} 0.65589244 & 1.51045141 & 1.55538795 \\ 0.73604761 & 1.04267141 & 0.39199789 \\ 0.78027301 & 0.37396246 & -0.25846633 \end{bmatrix}.$$

where truncation $0 \le i \le 100$ and $0 \le j \le 100$ was used to compute $F_k(\boldsymbol{P})$, $k = 1, 2, 3, 4$. Substituting the above

$$\boldsymbol{K}_A^{(1)} = \begin{bmatrix} 1251.75449772 & -2979.86329982 & 2999.81755684 \\ -2979.86329982 & 7122.30587652 & -7133.50326166 \\ 2999.81755684 & -7133.50326166 & 7299.95917720 \end{bmatrix}$$
$$\boldsymbol{K}_A^{(4)} = \begin{bmatrix} 845585.48203204 & -1993349.60116238 & 2015318.08545784 \\ -1993349.60116238 & 4714069.08436384 & -4744374.88497449 \\ 2015318.08545784 & -4744374.88497449 & 4863101.78275287 \end{bmatrix}$$
$$\boldsymbol{K}_{o1}^{(1)} = \begin{bmatrix} 0.00149858 & -0.00364930 & 0.00356937 \\ -0.00364930 & 0.00956165 & -0.00905795 \\ 0.00356937 & -0.00905795 & 0.00933207 \end{bmatrix}$$
$$\boldsymbol{K}_{o1}^{(4)} = \begin{bmatrix} 0.15953189 & -0.38560521 & 0.37922176 \\ -0.38560521 & 1.00770155 & -0.95804046 \\ 0.37922176 & -0.95804046 & 0.99147667 \end{bmatrix}$$
$$\boldsymbol{K}_{c2}^{(1)} = \begin{bmatrix} 45322.52827960 & 43822.10189449 & 39634.12075737 \\ 43822.10189449 & 45307.52406750 & 43818.67539592 \\ 39634.12075737 & 43818.67539592 & 45327.66655788 \end{bmatrix}$$
$$\boldsymbol{K}_{c2}^{(4)} = \begin{bmatrix} 77.13949821 & 74.86433871 & 67.80484994 \\ 74.86433871 & 77.73814927 & 75.31371509 \\ 67.80484994 & 75.31371509 & 78.03763354 \end{bmatrix}$$
$$\boldsymbol{K}_{oB}^{(1)} = \begin{bmatrix} 0.00657555 & -0.01610628 & 0.01597188 \\ -0.01610628 & 0.04000610 & -0.03922995 \\ 0.01597188 & -0.03922995 & 0.04019986 \end{bmatrix}$$
$$\boldsymbol{K}_{oB}^{(4)} = \begin{bmatrix} 4.24697220 & -10.09219961 & 10.16484848 \\ -10.09219961 & 24.09804943 & -24.12544059 \\ 10.16484848 & -24.12544059 & 24.73648669 \end{bmatrix}$$
$$\boldsymbol{K}_{cC}^{(1)} = \begin{bmatrix} 947.12826344 & 900.34470992 & 772.59639177 \\ 900.34470992 & 946.91813917 & 900.22072602 \\ 772.59639177 & 900.22072602 & 947.21153262 \end{bmatrix}$$
$$\boldsymbol{K}_{cC}^{(4)} = \begin{bmatrix} 1.97861983 & 1.88693064 & 1.62312130 \\ 1.88693064 & 1.99044115 & 1.89578584 \\ 1.62312130 & 1.89578584 & 1.99633146 \end{bmatrix}.$$

TABLE I
WEIGHTED SENSITIVITY ANALYSIS

Realization	Example 1	Example 2
Original	$M_{1/2} = 7.5078 \times 10^6$	$M_{1/2} = 2.9967 \times 10^5$
	$m_2 = 7.8189 \times 10^8$	$m_2 = 1.0441 \times 10^7$
Suboptimal	$\overline{M}_{1/2} = 8.0580 \times 10^3$	$\overline{M}_{1/2} = 3.6461 \times 10^2$
	$\overline{m}_2 = 9.8023 \times 10^4$	$\overline{m}_2 = 2.1862 \times 10^3$
L_1/L_2-Optimal	$\overline{M}_{1/2} = 6.4599 \times 10^3$	$\overline{M}_{1/2} = 2.9075 \times 10^2$
	$\overline{m}_2 = 1.6301 \times 10^5$	$\overline{m}_2 = 2.8063 \times 10^3$
L_2-Optimal	$\overline{M}_{1/2} = 1.1547 \times 10^4$	$\overline{M}_{1/2} = 3.8044 \times 10^2$
	$\overline{m}_2 = 8.4194 \times 10^4$	$\overline{m}_2 = 2.1519 \times 10^3$
$(J_{min} + L_{min})$	6.4514×10^3	2.8905×10^2

coordinate transformation into (4) results in

$$\overline{\boldsymbol{A}}_1 = \begin{bmatrix} 0.79612923 & -0.21332680 & -0.06943544 \\ 0.23792156 & 0.37583635 & -0.40698693 \\ -0.10257831 & 0.44748310 & 0.73473442 \end{bmatrix}$$

$$\overline{\boldsymbol{A}}_2 = \begin{bmatrix} 0.22375543 & 0.22818147 & 0.13019843 \\ -0.22779964 & -0.23189051 & -0.13252835 \\ 0.12926548 & 0.13011081 & 0.07252603 \end{bmatrix}$$

$$\overline{\boldsymbol{A}}_3 = \begin{bmatrix} -0.00038721 & 0.00272144 & -0.00071499 \\ -0.00109674 & -0.00103953 & 0.00243030 \\ -0.00121810 & 0.00214948 & 0.00091477 \end{bmatrix}$$

$$\overline{\boldsymbol{A}}_4 = \begin{bmatrix} 0.79627366 & -0.23797139 & -0.10339016 \\ 0.21193424 & 0.37311787 & -0.45053023 \\ -0.06836792 & 0.40837312 & 0.73313847 \end{bmatrix}$$

$$\overline{\boldsymbol{b}}_1 = [0.11118275 \quad -0.11288667 \quad 0.06372873]^t$$

$$\overline{\boldsymbol{b}}_2 = [2.85345887 \quad -2.46012520 \quad 1.18577329]^t$$

$$\overline{\boldsymbol{c}}_1 = [0.48512286 \quad 0.41842076 \quad 0.20174307]$$

$$\overline{\boldsymbol{c}}_2 = [0.01891648 \quad 0.01922876 \quad 0.01090239]$$

which is L_2 optimal and from (17) and (57) we have

$$\overline{M}_{1/2} = J + L = 380.4402$$
$$\overline{m}_2 = 2151.9307$$

respectively, where $J = 367.3009$ and $L = 13.1394$.

The simulation results of the above examples are summarized in terms of the weighted sensitivities $\overline{M}_{1/2}$ and $\overline{m}_2$ in Table I. It is observed that the weighted sensitivity $\overline{M}_{1/2}$ of the L_1/L_2-optimal filter structures is very close to the value of $J_{\min} + L_{\min}$. Also, $\overline{m}_2$ of the L_1/L_2-optimal filter structures is not far away from $\overline{m}_2$ of the L_2-optimal filter structures.

VI. CONCLUSION

Two frequency-weighted sensitivity measures have been defined as a generalization of those reported in [10] and [13]. To construct the 2-D coordinate transformation matrix such that the weighted L_1/L_2 mixed sensitivity is optimal in a certain sense, an analytical method has been developed to obtain the closed-form solution. The 1-D version of the analytical method can be viewed as an alternative to the weighted sensitivity minimization algorithm reported in [4] and is much simpler than the algorithm which relies on the Lagrange multiplier method. In addition, the 1-D version of this closed-form solution has not been reported in [5]. An iterative procedure has been proposed to find the optimal coordinate transformation that minimizes the weighted L_2-sensitivity measure. The merit of this procedure is that the estimate at each iteration can be derived analytically. Our first contribution in this paper has been the introduction of general unconstrained frequency weights for 2-D state-space digital filters. The second is to present a novel closed-form solution for obtaining the 2-D filter structures that minimize $\overline{M}_{1/2}(\boldsymbol{P})$, subject to the minimization of $J(\boldsymbol{P})$. The third is to develop a procedure for iteratively finding the optimal coordinate transformation that yields the filter structures with minimum weighted L_2-sensitivity. We have illustrated the utility of the proposed technique with two numerical examples.

It should be noted that the approach presented here can be extended to the M-dimensional case where $M > 2$ in a straightforward manner, provided the multidimensional LSS model reported in [25] is employed. In addition, similar arguments can be applied to the Fornasini–Marchesini second LSS model [17], [15].

REFERENCES

[1] L. Thiele, "Design of sensitivity and round-off noise optimal state-space discrete systems," *Int. J. Circuit Theory Appl.*, vol. 12, pp. 39–46, Jan. 1984.

[2] ——, "On the sensitivity of linear state-space systems," *IEEE Trans. Circuits Syst.*, vol. CAS-33, pp. 502–510, May 1986.

[3] M. Iwatsuki, M. Kawamata, and T. Higuchi, "Statistical sensitivity and minimum sensitivity structures with fewer coefficients in discrete time linear systems," *IEEE Trans. Circuits Syst.*, vol. 37, pp. 72–80, Jan. 1989.

[4] G. Li and M. Gevers, "Optimal finite precision implementation of a state-estimate feedback controller," *IEEE Trans. Circuits Syst.*, vol. 37, pp. 1487–1498, Dec. 1990.

[5] G. Li, B. D. O. Anderson, M. Gevers, and J. E. Perkins, "Optimal FWL design of state-space digital systems with weighted sensitivity minimization and sparseness consideration," *IEEE Trans. Circuits Syst. I*, vol. 39, pp. 365–377, May 1992.

[6] W.-Y. Yan and J. B. Moore, "On L^2-sensitivity minimization of linear state-space systems," *IEEE Trans. Circuits Syst. I*, vol. 39, pp. 641–648, Aug. 1992.

[7] G. Li and M. Gevers, "Optimal synthetic FWL design of state-space digital filters," in *Proc. 1992 IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 4, pp. 429–432.

[8] W. J. Lutz and S. L. Hakimi, "Design of multi-input multi-output systems with minimum sensitivity," *IEEE Trans. Circuits Syst.*, vol. 35, pp. 1114–1122, Sept. 1988.

[9] A. Zilouchian and R. L. Carroll, "A coefficient sensitivity bound in 2-D state-space digital filtering," *IEEE Trans. Circuits Syst.*, vol. CAS-33, pp. 665–667, June 1986.

[10] M. Kawamata, T. Lin, and T. Higuchi, "Minimization of sensitivity of 2-D state-space digital filters and its relation to 2-D balanced realizations," in *Proc. 1987 IEEE Int. Symp. Circuits Systems*, pp. 710–713.

[11] T. Hinamoto, T. Hamanaka, and S. Maekawa, "Synthesis of 2-D state-space digital filters with low sensitivity based on the Fornasini-Marchesini model," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-38, pp. 1587–1594, Sept. 1990.

[12] T. Hinamoto, T. Takao, and M. Muneyasu, "Synthesis of 2-D separable-denominator digital filters with low sensitivity," J. Franklin Inst., vol. 329, pp. 1063–1080, 1992.

[13] T. Hinamoto and T. Takao, "Synthesis of 2-D state-space filter structures with low frequency-weighted sensitivity," *IEEE Trans. Circuits Syst. II*, vol. 39, pp. 646–651, Sept. 1992.

[14] ——, "Minimization of frequency-weighting sensitivity in 2-D systems based on the Fornasini-Marchesini second model," in *Proc. 1992*

IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 401–404; also "On the frequency-weighting sensitivity of 2-D state-space digital filters based on the Fornasini-Marchesini second model," *IEICE Trans. Fundamentals*, vol. E75-A, pp. 813–820, July 1992.
[15] G. Li, "On frequency weighted minimal L_2 sensitivity of 2-D systems using Fornasini-Marchesini LSS model," *IEEE Trans. Circuits Syst. I*, vol. 44, pp. 642–646, July 1997.
[16] R. P. Roesser, "A discrete state-space model for linear image processing," *IEEE Trans. Automat. Contr.*, vol. AC-20, pp. 1–10, Feb. 1975.
[17] E. Fornasini and G. Marchesini, "Doubly-indexed dynamical systems: State-space models and structural properties," *Math. Syst. Theory*, vol. 12, pp. 59–72, 1978.
[18] S. Kung, B. C. Lévy, M. Morf, and T. Kailath, "New results in 2-D systems theory, Part II: 2-D state-space models—Realization and the notions of controllability, observability, and minimality," *Proc. IEEE*, vol. 65, pp. 945–961, June 1977.
[19] L. L. Scharf, *Statistical Signal Processing*. Reading, MA: Addison-Wesley, 1991.
[20] C. T. Mullis and R. A. Roberts, "Synthesis of minimum roundoff noise fixed point digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-23, pp. 551–562, 1976.
[21] S. Shokoohi, "Block-balanced realizations," in *Proc. 1984 IEEE Int. Symp. Circuits Systems*, pp. 825–829.
[22] C. Xiao, P. Agathoklis, and D. J. Hill, "Coefficient sensitivity and structure optimization of multidimensional state-space digital filters," in *Proc. 1997 IEEE Int. Symp. Circuits Systems*, pp. 2465–2468.
[23] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd ed. Baltimore, MD: Johns Hopkins Univ. Press, 1989.
[24] B. W. Bowmar and J. C. Hung, "Minimum roundoff noise digital filters with some power-of-two coefficients," *IEEE Trans. Circuits Syst.*, vol. CAS-31, pp. 833–840, Oct. 1984.
[25] D. S. K. Chan, "The structure of recursible multidimensional discrete systems," *IEEE Trans. Automat. Contr.*, vol. AC-25, pp. 663–673, Aug. 1980.

Takao Hinamoto (M'77–SM'84) received the B.E. degree in electrical engineering from Okayama University, Okayama, Japan, in 1969, the M.E. degree in electrical engineering from Kobe University, Kobe, Japan, in 1971, and the Dr. Eng. degree in electrical engineering from Osaka University, Osaka, Japan, in 1977.

From 1972 to 1988, he was with the Faculty of Engineering, Kobe University, Kobe, Japan. From September 1979 to March 1981, he was on leave from Kobe University as a Visiting Member of Staff in the Department of Electrical Engineering, Queen's University, Kingston, ONT, Canada. During 1988–1991, he was Professor of Electronic Circuits in the Faculty of Engineering, Tottori University, Tottori, Japan. Since January 1992, he has been Professor of Electronic Control in the Department of Electrical Engineering, Hiroshima University, Higashi-Hiroshima, Japan. His teaching and research interests include digital signal processing, system theory, and control systems engineering. He has coedited and coauthored the book *Two-Dimensional Signal and Image Processing* (Tokyo, Japan: SICE, 1996).

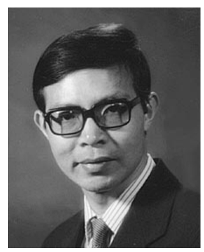
Dr. Hinamoto was an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS II during 1993–1995 and served as the Chair of the 12th Digital Signal Processing (DSP) Symposium held in Hiroshima in November 1997, which is the 12th of a series of annual symposia sponsored by the DSP Technical Committee of the IEICE. He also served as the Guest Editor of the special section on DSP in the August 1998 issue of the *IEICE Transactions on Fundamentals*.

Yoshitaka Zempo received the B.E. and M.E. degrees in electrical engineering from Hiroshima University in 1993 and 1995, respectively.

In April 1995 he joined Yamaguchi Television and Radio Broadcasting Station, Inc., Tokuyama, Japan. His research interests are in digital signal processing and system theory.

Yoshio Nishino received the B.E., M.E., and Dr. Eng. degrees in applied mathematics and physics from Kyoto University in 1971, 1973, and 1979, respectively.

He was a Lecturer in the Faculty of Engineering, Kyoto University, Kyoto, Japan, from 1979 to 1981. Since 1982 he has been an Associate Professor of the Faculty of Engineering, Hiroshima University, Hiroshima, Japan. His study is concerned with applications of differential geometry and group theory to problems arising in physics and engineering.



Wu-Sheng Lu (S'81–M'85–SM'90–F'99) received the B.S. degree in mathematics from Fudan University, Fudan, China, and the M.S. degree in electrical engineering and the Ph.D. degree in control science from the University of Minnesota, Minneapolis, in 1964, 1983, and 1984, respectively.

He was a post-doctoral fellow at the University of Victoria, Victoria, BC, Canada in 1985 and a visiting assistant professor at the University of Minnesota in 1986. Since 1987 he has been with the University of Victoria, where he is currently Professor. His teaching and research interests are in the areas of digital signal processing and robotics. He is the co-author with A. Antoniou of *Two-Dimensional Digital Filters* (New York: Marcel Dekker, 1992).

Dr. Lu served as an Associate Editor of the *Canadian Journal of Electrical and Computer Engineering* in 1989 and was the Editor of the same journal from 1990 to 1992. He was an Associate Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS II from 1993 to 1995 and is presently an Associate Editor of *Multidimensional Systems and Signal Processing.*