L_2-Sensitivity Minimization of One- and Two-Dimensional State–Space Digital Filters Subject to L_2-Scaling Constraints

Takao Hinamoto, *Fellow, IEEE*, Ken-ichi Iwata, and Wu-Sheng Lu, *Fellow, IEEE*

***Abstract*—This paper investigates the problem of minimizing an L_2-sensitivity measure subject to L_2-norm dynamic-range scaling constraints for one-dimensional (1-D) as well as a class of two-dimensional (2-D) state–space digital filters, where the 2-D digital filters are described by a transposed structure of the Fornasini–Marchesini second local state–space model. In each case, a novel iterative technique is developed to solve the constraint optimization problem directly. The proposed solution methods are largely based on the use of a Lagrange function and some matrix-theoretic techniques. Numerical examples are presented to demonstrate the effectiveness of the proposed techniques.**

***Index Terms*—A class of 2-D state–space digital filters, Lagrange's function, L_2-norm dynamic-range scaling constraints, L_2-sensitivity minimization, optimal realization, state–space digital filters.**

I. Introduction

FOR over three decades, the issue on finite word length (FWL) effects has been a significant research topic in the implementation of fixed-point state–space digital filters. A great deal of effort has been made to synthesize the optimal FWL state–space filter structures that minimize the FWL effects on the efficiency and performance of the filter actually implemented. This paper is concerned with the problem of minimizing coefficient sensitivity in state–space digital filters. Given a transfer function with infinite accuracy coefficients that meets specification requirements, we often need to implement its state–space model using a finite binary representation. This requires the truncation or rounding of coefficients in the state–space model to fit the FWL constraints. As a result, the characteristics of a stable filter might be so altered that the filter may become unstable. This motivates the study of the coefficient sensitivity minimization problem. In the literature, techniques for synthesizing the state–space descriptions that minimize the coefficient sensitivity can be divided into two main classes: L_1/L_2-mixed sensitivity minimization [1]–[5] and L_2-sensitivity minimization [6]–[10]. In [6]–[10], it has been argued that the sensitivity measure based on the L_2-norm only is more natural and reasonable relative to the L_1/L_2-mixed sensitivity minimization. For 2-D state–space digital filters, the L_1/L_2-mixed sensitivity minimization problem [11]–[15] and L_2-sensitivity minimization problem [10], [16]–[19] have also been investigated. However, to our best knowledge, little has been done for the minimization of L_2-sensitivity subject to the L_2-norm dynamic-range scaling constraints for state–space digital filters [20], although it has been known that the use of scaling constraints can be beneficial for suppressing overflow oscillations [21], [22].

This paper investigates the problem of minimizing an L_2-sensitivity measure subject to L_2-norm dynamic-range scaling constraints for state–space digital filters as well as a class of 2-D state–space digital filters. To this end, an expression for evaluating the L_2-sensitivity is introduced, and the L_2-sensitivity minimization problem subject to the scaling constraints is formulated in each case. An iterative algorithm is then developed in each case to solve the constraint optimization problem directly. Our solution methods are largely based on the use of a Lagrange function and some matrix-theoretic techniques. Next, the coordinate transformation matrix is adjusted without altering the L_2-sensitivity to satisfy the scaling constraints, and is used to construct the optimal state–space filter structure that minimizes the L_2-sensitivity measure subject to the scaling constraints. Two numerical examples are presented to demonstrate the utility of the proposed algorithms.

Unlike the work reported in [20], the proposed iterative technique relies on neither converting the problem into an unconstrained optimization formulation nor using a quasi-Newton algorithm. From computer simulation results, it has turned out that the proposed iterative technique requires less than half amount of computations to attain practically the same convergence accuracy as compared with the technique reported in [20].

Throughout this paper, $\boldsymbol{I}_n$ denotes the identity matrix of dimension $n \times n$. The transpose (conjugate transpose) of a matrix $\boldsymbol{A}$ and trace of a square matrix $\boldsymbol{A}$ are denoted by $\boldsymbol{A}^T(\boldsymbol{A}^*)$ and $\mathrm{tr}[\boldsymbol{A}]$, respectively. The ith diagonal element of a square matrix $\boldsymbol{A}$ is denoted by $(\boldsymbol{A})_{ii}$.

II. L_2-Sensitivity Minimization for 1-D Filters

A. L_2-Sensitivity Analysis

Consider a stable, controllable, and observable, state–space digital filter $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)_n$ described by

$$\begin{aligned} \boldsymbol{x}(k+1) &= \boldsymbol{A}\boldsymbol{x}(k) + \boldsymbol{b}u(k) \\ y(k) &= \boldsymbol{c}\boldsymbol{x}(k) + du(k) \end{aligned} \tag{1}$$

Manuscript received February 9, 2005; revised June 6, 2005. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Sheila S. Hemami.

T. Hinamoto is with the Graduate School of Engineering, Hiroshima University, Hiroshima 739-8527, Japan (e-mail: hinamoto@hiroshima-u.ac.jp).

K. Iwata was with the Graduate School of Engineering, Hiroshima University, Hiroshima 739-8527, Japan. He is now with Mitsubishi Electric Corporation, Himeji, Japan (e-mail: iwata@hiroshima-u.ac.jp).

W.-S. Lu is with the Department of Electrical and Computer Engineering, University of Victoria, Victoria, BC V8W 3P6, Canada (e-mail: wslu@ece.uvic.ca).

Digital Object Identifier 10.1109/TSP.2006.870584

where $\boldsymbol{x}(k)$ is an $n \times 1$ state-variable vector, $u(k)$ is a scalar input, $y(k)$ is a scalar output, and $\boldsymbol{A}$, $\boldsymbol{b}$, $\boldsymbol{c}$ and d are real constant matrices of appropriate dimensions. The transfer function of the filter in (1) is given by

$$H(z) = \boldsymbol{c}(z\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b} + d. \tag{2}$$

The L_2-sensitivity of the filter in (1) is defined as follows.

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

$$\boldsymbol{S}_{\boldsymbol{X}} = \frac{\partial f}{\partial \boldsymbol{X}}, \quad (\boldsymbol{S}_{\boldsymbol{X}})_{ij} = \frac{\partial f}{\partial x_{ij}} \tag{3}$$

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

Definition 2: Let $\boldsymbol{X}(z)$ be an $m \times n$ complex matrix-valued function of a complex variable z and let $x_{pq}(z)$ be the (p,q)th entry of $\boldsymbol{X}(z)$. The L_2-norm of $\boldsymbol{X}(z)$ is then defined as

$$\begin{aligned} \|\boldsymbol{X}(z)\|_2 &= \left[\frac{1}{2\pi}\int_0^{2\pi}\sum_{p=1}^{m}\sum_{q=1}^{n}\left|x_{pq}(e^{j\omega})\right|^2 d\omega\right]^{\frac{1}{2}} \\ &= \left(\operatorname{tr}\left[\frac{1}{2\pi j}\oint_{|z|=1}\boldsymbol{X}(z)\boldsymbol{X}^*(z)\frac{dz}{z}\right]\right)^{\frac{1}{2}}. \end{aligned} \tag{4}$$

From (2) and Definitions 1 and 2, the overall L_2-sensitivity measure for the filter in (1) is defined as

$$\begin{aligned} S_1 &= \left\|\frac{\partial H(z)}{\partial \boldsymbol{A}}\right\|_2^2 + \left\|\frac{\partial H(z)}{\partial \boldsymbol{b}}\right\|_2^2 + \left\|\frac{\partial H(z)}{\partial \boldsymbol{c}^T}\right\|_2^2 \\ &= \left\|[\boldsymbol{F}(z)\boldsymbol{G}(z)]^T\right\|_2^2 + \left\|\boldsymbol{G}^T(z)\right\|_2^2 + \|\boldsymbol{F}(z)\|_2^2 \end{aligned} \tag{5}$$

where

$$\boldsymbol{F}(z) = (z\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b}, \quad \boldsymbol{G}(z) = \boldsymbol{c}(z\boldsymbol{I}_n - \boldsymbol{A})^{-1}.$$

The term d in (2) and the sensitivity with respect to it are coordinate independent, and therefore they are neglected here.

It is easy to show that the L_2-sensitivity measure in (5) can be expressed as

$$S_1 = \operatorname{tr}[\boldsymbol{M}(\boldsymbol{I}_n)] + \operatorname{tr}[\boldsymbol{W}_o] + \operatorname{tr}[\boldsymbol{K}_c] \tag{6}$$

where

$$\begin{aligned} \boldsymbol{K}_c &= \frac{1}{2\pi j}\oint_{|z|=1}\boldsymbol{F}(z)\boldsymbol{F}^T(z^{-1})\frac{dz}{z} \\ \boldsymbol{W}_o &= \frac{1}{2\pi j}\oint_{|z|=1}\boldsymbol{G}^T(z)\boldsymbol{G}(z^{-1})\frac{dz}{z} \\ \boldsymbol{M}(\boldsymbol{P}) &= \frac{1}{2\pi j}\oint_{|z|=1}[\boldsymbol{F}(z)\boldsymbol{G}(z)]^T\boldsymbol{P}^{-1}\boldsymbol{F}(z^{-1})\boldsymbol{G}(z^{-1})\frac{dz}{z}. \end{aligned}$$

The matrices $\boldsymbol{K}_c$ and $\boldsymbol{W}_o$ in (6) are called the controllability and observability Gramians, respectively, and can be obtained by solving the following Lyapunov equations [23]:

$$\begin{aligned} \boldsymbol{K}_c &= \boldsymbol{A}\boldsymbol{K}_c\boldsymbol{A}^T + \boldsymbol{b}\boldsymbol{b}^T \\ \boldsymbol{W}_o &= \boldsymbol{A}^T\boldsymbol{W}_o\boldsymbol{A} + \boldsymbol{c}^T\boldsymbol{c}. \end{aligned} \tag{7}$$

In order to evaluate the term $\operatorname{tr}[\boldsymbol{M}(\boldsymbol{I}_n)]$ in (6), one only needs to compute $\boldsymbol{M}(\boldsymbol{P})$ with $\boldsymbol{P} = \boldsymbol{I}_n$. However, the general solution method for computing $\boldsymbol{M}(\boldsymbol{P})$ with an arbitrary positive-definite matrix $\boldsymbol{P}$ will be required in our subsequent development. We shall address this issue shortly.

If a coordinate transformation defined by

$$\overline{\boldsymbol{x}}(k) = \boldsymbol{T}^{-1}\boldsymbol{x}(k) \tag{8}$$

is applied to the filter in (1), then the new realization $(\overline{\boldsymbol{A}}, \overline{\boldsymbol{b}}, \overline{\boldsymbol{c}}, d)_n$ can be characterized by

$$\begin{aligned} \overline{\boldsymbol{A}} &= \boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T}, \quad \overline{\boldsymbol{b}} = \boldsymbol{T}^{-1}\boldsymbol{b}, \quad \overline{\boldsymbol{c}} = \boldsymbol{c}\boldsymbol{T} \\ \overline{\boldsymbol{K}}_c &= \boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T}, \quad \overline{\boldsymbol{W}}_o = \boldsymbol{T}^T\boldsymbol{W}_o\boldsymbol{T}. \end{aligned} \tag{9}$$

From (2) and (9), it is clear that the transfer function $H(z)$ is invariant under the coordinate transformation in (8). In addition, under the coordinate transformation in (8), the Gramian $\boldsymbol{M}(\boldsymbol{I}_n)$ becomes $\boldsymbol{T}^T\boldsymbol{M}(\boldsymbol{P})\boldsymbol{T}$ and the L_2-sensitivity measure in (6) is changed to

$$S_1(\boldsymbol{P}) = \operatorname{tr}[\boldsymbol{M}(\boldsymbol{P})\boldsymbol{P}] + \operatorname{tr}[\boldsymbol{W}_o\boldsymbol{P}] + \operatorname{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}] \tag{10}$$

where $\boldsymbol{P} = \boldsymbol{T}\boldsymbol{T}^T$. Noting that

$$\begin{aligned} \overline{\boldsymbol{F}}(z)\overline{\boldsymbol{G}}(z) &= \boldsymbol{T}^{-1}\boldsymbol{F}(z)\boldsymbol{G}(z)\boldsymbol{T} \\ &= [\boldsymbol{T}^{-1}\ \boldsymbol{0}]\begin{bmatrix} z\boldsymbol{I}_n - \boldsymbol{A} & -\boldsymbol{b}\boldsymbol{c} \\ \boldsymbol{0} & z\boldsymbol{I}_n - \boldsymbol{A} \end{bmatrix}^{-1}\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{T} \end{bmatrix} \end{aligned} \tag{11}$$

where

$$\overline{\boldsymbol{F}}(z) = (z\boldsymbol{I}_n - \overline{\boldsymbol{A}})^{-1}\overline{\boldsymbol{b}}, \quad \overline{\boldsymbol{G}}(z) = \overline{\boldsymbol{c}}(z\boldsymbol{I}_n - \overline{\boldsymbol{A}})^{-1}$$

and denoting the observability Gramian of the system in (11) by $\boldsymbol{Y}$, it can be shown for an arbitrary $\boldsymbol{P} = \boldsymbol{T}\boldsymbol{T}^T$, the matrix $\boldsymbol{M}(\boldsymbol{P})$ can be obtained by solving the Lyapunov equation

$$\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{b}\boldsymbol{c} \\ \boldsymbol{0} & \boldsymbol{A} \end{bmatrix}^T\boldsymbol{Y}\begin{bmatrix} \boldsymbol{A} & \boldsymbol{b}\boldsymbol{c} \\ \boldsymbol{0} & \boldsymbol{A} \end{bmatrix} + \begin{bmatrix} \boldsymbol{P}^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \tag{12}$$

and then taking the lower-right $n \times n$ block of $\boldsymbol{Y}$ as $\boldsymbol{M}(\boldsymbol{P})$, i.e.,

$$\boldsymbol{M}(\boldsymbol{P}) = [\boldsymbol{0} \quad \boldsymbol{I}_n]\boldsymbol{Y}\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{I}_n \end{bmatrix}. \tag{13}$$

Moreover, if the L_2-norm dynamic-range scaling constraints are imposed on the new state-variable vector $\overline{\boldsymbol{x}}(k)$, it is required that

$$(\overline{\boldsymbol{K}}_c)_{ii} = (\boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T})_{ii} = 1, \qquad \text{for } i = 1, 2, \cdots, n. \tag{14}$$

The problem of L_2-sensitivity minimization subject to L_2-norm dynamic-range scaling constraints can now be formulated as follows: *For given* $\boldsymbol{A}$, $\boldsymbol{b}$, *and* $\boldsymbol{c}$, *obtain an* $n \times n$

nonsingular matrix $\boldsymbol{T}$ that minimizes the sensitivity measure $S_1(\bar{\boldsymbol{P}})$ in (10) *subject to the scaling constraints in* (14).

B. L_2-Sensitivity Minimization

The problem of minimizing $S_1(\boldsymbol{P})$ in (10) subject to the constraints in (14) is a constrained nonlinear optimization problem where the variable matrix is $\boldsymbol{P}$. If we sum the n constraints in (14) up, then we have

$$\mathrm{tr}[\boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T}] = \mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}] = n. \tag{15}$$

Consequently, the problem of minimizing (10) subject to the constraints in (14) can be "relaxed" into the following problem:

$$\begin{aligned} &\text{minimize } S_1(\boldsymbol{P}) \text{ in } (10) \\ &\text{subject to } \mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}] = n. \end{aligned} \tag{16}$$

Although clearly a solution of problem (16) is not necessarily a solution of the problem of minimizing (10) subject to the constraints in (14), it is important to emphasize that the ultimate solution we seek is not matrix $\boldsymbol{P}$ but a nonsingular matrix $\boldsymbol{T}$ that is related to the solution of the problem of minimizing (10) subject to the constraints in (14) as $\boldsymbol{P} = \boldsymbol{T}\boldsymbol{T}^T$. If matrix $\boldsymbol{P}$ is a solution of problem (16) and $\boldsymbol{P}^{1/2}$ denotes a matrix square root of $\boldsymbol{P}$, i.e., $\boldsymbol{P} = \boldsymbol{P}^{1/2}\boldsymbol{P}^{1/2}$, then it is easy to see that any matrix $\boldsymbol{T}$ of the form $\boldsymbol{T} = \boldsymbol{P}^{1/2}\boldsymbol{U}$, where $\boldsymbol{U}$ is an arbitrary orthogonal matrix, still holds the relation $\boldsymbol{P} = \boldsymbol{T}\boldsymbol{T}^T$. As will be shown shortly, under the constraint $\mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}] = n$ in (16), there exists an orthogonal matrix $\boldsymbol{U}$ such that matrix $\boldsymbol{T} = \boldsymbol{P}^{1/2}\boldsymbol{U}$ satisfies the constraints in (14), where $\boldsymbol{P}^{1/2}$ is a square root of the solution matrix $\boldsymbol{P}$ for problem (16).

It is for these reasons, we now address problem (16) as the first step of our solution strategy. To solve (16), we define the Lagrange function of the problem as

$$\begin{aligned} J_1(\boldsymbol{P},\lambda) = \mathrm{tr}\,[\boldsymbol{M}(\boldsymbol{P})\boldsymbol{P}]\,\mathrm{tr}[\boldsymbol{W}_o\boldsymbol{P}] + \mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}] \\ +\lambda\left(\mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}] - n\right) \end{aligned} \tag{17}$$

where λ is a Lagrange multiplier. It is well known that the solution of problem (16) must satisfy the Karush–Kuhn–Tucker (KKT) conditions $\partial J_1(\boldsymbol{P},\lambda)/\partial\boldsymbol{P} = \boldsymbol{0}$ and $\partial J_1(\boldsymbol{P},\lambda)/\partial\lambda = 0$, where the gradients are found to be [24]

$$\begin{aligned} \frac{\partial J_1(\boldsymbol{P},\lambda)}{\partial \boldsymbol{P}} &= \boldsymbol{M}(\boldsymbol{P}) - \boldsymbol{P}^{-1}\boldsymbol{N}(\boldsymbol{P})\boldsymbol{P}^{-1} \\ &\quad + \boldsymbol{W}_o - (\lambda+1)\boldsymbol{P}^{-1}\boldsymbol{K}_c\boldsymbol{P}^{-1} \\ \frac{\partial J_1(\boldsymbol{P},\lambda)}{\partial \lambda} &= \mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}] - n \end{aligned} \tag{18}$$

where $\boldsymbol{N}(\bar{\boldsymbol{P}})$ can be obtained by solving the following Lyapunov equation:

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{b}\boldsymbol{c} \\ \boldsymbol{0} & \boldsymbol{A} \end{bmatrix} \boldsymbol{Z} \begin{bmatrix} \boldsymbol{A} & \boldsymbol{b}\boldsymbol{c} \\ \boldsymbol{0} & \boldsymbol{A} \end{bmatrix}^T + \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{P} \end{bmatrix}$$

and then taking the upper-left $n \times n$ block of $\boldsymbol{Z}$, i.e.,

$$\boldsymbol{N}(\boldsymbol{P}) = [\boldsymbol{I}_n \quad \boldsymbol{0}]\boldsymbol{Z}\begin{bmatrix} \boldsymbol{I}_n \\ \boldsymbol{0} \end{bmatrix}.$$

Note that matrix $\boldsymbol{Z}$ corresponds to the controllability Gramian of the system in (11).

Hence, the KKT conditions become

$$\begin{aligned} \boldsymbol{P}\boldsymbol{F}(\boldsymbol{P})\boldsymbol{P} &= \boldsymbol{G}(\boldsymbol{P},\lambda) \\ \mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}] &= n \end{aligned} \tag{19}$$

where

$$\begin{aligned} \boldsymbol{F}(\boldsymbol{P}) &= \boldsymbol{M}(\boldsymbol{P}) + \boldsymbol{W}_o \\ \boldsymbol{G}(\boldsymbol{P},\lambda) &= \boldsymbol{N}(\boldsymbol{P}) + (\lambda+1)\boldsymbol{K}_c. \end{aligned}$$

The first equation in (19) is highly nonlinear with respect to $\boldsymbol{P}$. An effective approach to solving the first equation in (19) is to "relax" it into the following recursive second-order matrix equation:

$$\boldsymbol{P}_{i+1}\boldsymbol{F}(\boldsymbol{P}_i)\boldsymbol{P}_{i+1} = \boldsymbol{G}(\boldsymbol{P}_i,\lambda_i) \tag{20}$$

where $\boldsymbol{P}_i$ is assumed to be known from the previous recursion. Noting that $\boldsymbol{P}\boldsymbol{W}\boldsymbol{P} = \boldsymbol{M}$ has the unique solution [5]

$$\boldsymbol{P} = \boldsymbol{W}^{-\frac{1}{2}}\left[\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{M}\boldsymbol{W}^{\frac{1}{2}}\right]^{\frac{1}{2}}\boldsymbol{W}^{-\frac{1}{2}} \tag{21}$$

where $\boldsymbol{W} > 0$ and $\boldsymbol{M} \geq 0$ are symmetric, the solution $\boldsymbol{P}_{i+1}$ of (20) is given by

$$\boldsymbol{P}_{i+1} = \boldsymbol{F}(\boldsymbol{P}_i)^{-\frac{1}{2}}\left[\boldsymbol{F}(\boldsymbol{P}_i)^{\frac{1}{2}}\boldsymbol{G}(\boldsymbol{P}_i,\lambda_i)\boldsymbol{F}(\boldsymbol{P}_i)^{\frac{1}{2}}\right]^{\frac{1}{2}}\boldsymbol{F}(\boldsymbol{P}_i)^{-\frac{1}{2}}. \tag{22}$$

To derive a recursive formula for the Lagrange multiplier λ, we employ (19) to write

$$\mathrm{tr}\,[\boldsymbol{P}\boldsymbol{F}(\boldsymbol{P})] = \mathrm{tr}\left[\boldsymbol{N}(\boldsymbol{P})\boldsymbol{P}^{-1}\right] + n(\lambda+1) \tag{23}$$

which naturally suggests the following recursion for λ:

$$\lambda_{i+1} = \frac{\mathrm{tr}\,[\boldsymbol{P}_i\boldsymbol{F}(\boldsymbol{P}_i)] - \mathrm{tr}\left[\boldsymbol{N}(\boldsymbol{P}_i)\boldsymbol{P}_i^{-1}\right]}{n} - 1. \tag{24}$$

In the above algorithm, λ_i is obtained from the previous iteration. The iteration process starts with $\boldsymbol{P}_1 = \boldsymbol{I}_n$ and any value of $\lambda_1 > 0$ and continues until (19) is satisfied within a prescribed numerical tolerance.

As the second step of the solution strategy, we now turn our attention to the construction of the optimal coordinate transformation matrix $\boldsymbol{T}$ that solves the problem of minimizing (10) subject to the constraints in (14). As analyzed earlier, the optimal $\boldsymbol{T}$ assumes the form

$$\boldsymbol{T} = \boldsymbol{P}^{\frac{1}{2}}\boldsymbol{U} \tag{25}$$

where $\boldsymbol{P}^{1/2}$ is the square root of the matrix $\boldsymbol{P}$ obtained above, and $\boldsymbol{U}$ is an $n \times n$ orthogonal matrix to be determined as follows. From (9) and (25), it follows that

$$\begin{aligned} \overline{\boldsymbol{K}}_c &= \boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T} \\ &= \boldsymbol{U}^T\boldsymbol{P}^{-\frac{1}{2}}\boldsymbol{K}_c\boldsymbol{P}^{-\frac{1}{2}}\boldsymbol{U}. \end{aligned} \tag{26}$$

In order to find an orthogonal matrix $\boldsymbol{U}$ such that the matrix $\overline{\boldsymbol{K}}_c$ in (26) satisfies the constraints in (14), we perform the eigenvalue-eigenvector decomposition for the positive definite matrix $\boldsymbol{P}^{-1/2}\boldsymbol{K}_c\boldsymbol{P}^{-1/2}$ as

$$\boldsymbol{P}^{-\frac{1}{2}}\boldsymbol{K}_c\boldsymbol{P}^{-\frac{1}{2}} = \boldsymbol{R}\boldsymbol{\Theta}\boldsymbol{R}^T \tag{27}$$

where $\boldsymbol{\Theta} = \mathrm{diag}\{\theta_1, \theta_2, \cdots, \theta_n\}$ with $\theta_i > 0$ and $\boldsymbol{R}$ is an orthogonal matrix. Next, an orthogonal matrix $\boldsymbol{S}$ such that

$$\boldsymbol{S}\boldsymbol{\Theta}\boldsymbol{S}^T = \begin{bmatrix} 1 & * & \cdots & * \\ * & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ * & \cdots & * & 1 \end{bmatrix} \tag{28}$$

can be obtained by numerical manipulations [22, p. 278]. Using (26), (27), and (28), it can be readily verified that the orthogonal matrix $\boldsymbol{U} = \boldsymbol{R}\boldsymbol{S}^T$ leads to a $\overline{\boldsymbol{K}}_c$ in (26) whose diagonal elements are equal to unity; hence, the constraints in (14) are now satisfied. This matrix $\boldsymbol{T}$ together with (25) gives the solution of the problem of minimizing (10) subject to the constraints in (14) as

$$\boldsymbol{T} = \boldsymbol{P}^{\frac{1}{2}}\boldsymbol{R}\boldsymbol{S}^T. \tag{29}$$

III. L_2-Sensitivity Minimization for 2-D Digital Filters

A. L_2-Sensitivity Analysis

Consider a local state–space model $(\boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{b}, \boldsymbol{c}_1, \boldsymbol{c}_2, d)_n$ [25] for a class of 2-D state–space digital filters that is described by

$$\begin{bmatrix} \boldsymbol{x}(i+1, j+1) \\ y(i,j) \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{A}_2 \\ \boldsymbol{c}_1 & \boldsymbol{c}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{x}(i, j+1) \\ \boldsymbol{x}(i+1, j) \end{bmatrix} + \begin{bmatrix} \boldsymbol{b} \\ d \end{bmatrix} u(i,j) \tag{30}$$

where $\boldsymbol{x}(i,j)$ is an $n \times 1$ local state vector, $u(i,j)$ is a scalar input, $y(i,j)$ is a scalar output, and $\boldsymbol{A}_1$, $\boldsymbol{A}_2$, $\boldsymbol{b}$, $\boldsymbol{c}_1$, $\boldsymbol{c}_2$ and d are real constant matrices of appropriate dimensions. The 2-D filter in (30) is assumed to be stable, locally controllable, and locally observable. The transfer function of the 2-D filter in (30) is given by

$$H(z_1, z_2) = \left(z_1^{-1}\boldsymbol{c}_1 + z_2^{-1}\boldsymbol{c}_2\right) \cdot \left(\boldsymbol{I}_n - z_1^{-1}\boldsymbol{A}_1 - z_2^{-1}\boldsymbol{A}_2\right)^{-1} \boldsymbol{b} + d. \tag{31}$$

A block diagram of the local state–space (LSS) model in (30) is shown in Fig. 1. It is interesting to note that

$$H^T(z_1, z_2) = \boldsymbol{b}^T \left(\boldsymbol{I}_n - z_1^{-1}\boldsymbol{A}_1^T - z_2^{-1}\boldsymbol{A}_2^T\right)^{-1} \cdot \left(z_1^{-1}\boldsymbol{c}_1^T + z_2^{-1}\boldsymbol{c}_2^T\right) + d \tag{32}$$

can be viewed as a transfer function of the Fornasini–Marchesini second LSS model [26]. Since $H(z_1, z_2) = H^T(z_1, z_2)$, the LSS model in (30) corresponds to a transposed structure of the Fornasini–Marchesini second LSS model.

From (31) and Definition 1, it can easily be verified that

$$\begin{aligned} \frac{\partial H(z_1, z_2)}{\partial \boldsymbol{A}_k} &= z_k^{-1}\left[\boldsymbol{F}(z_1, z_2)\boldsymbol{G}(z_1, z_2)\right]^T \\ \frac{\partial H(z_1, z_2)}{\partial \boldsymbol{b}} &= \boldsymbol{G}^T(z_1, z_2) \\ \frac{\partial H(z_1, z_2)}{\partial \boldsymbol{c}_k^T} &= z_k^{-1}\boldsymbol{F}(z_1, z_2), \quad k = 1, 2 \end{aligned} \tag{33}$$

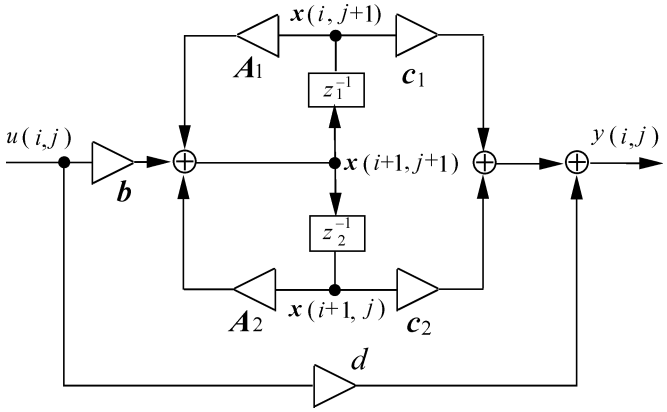

Fig. 1. LSS model for 2-D filters.

where

$$\begin{aligned} \boldsymbol{F}(z_1, z_2) &= \left(\boldsymbol{I}_n - z_1^{-1}\boldsymbol{A}_1 - z_2^{-1}\boldsymbol{A}_2\right)^{-1}\boldsymbol{b} \\ \boldsymbol{G}(z_1, z_2) &= \left(z_1^{-1}\boldsymbol{c}_1 + z_2^{-1}\boldsymbol{c}_2\right) \cdot \left(\boldsymbol{I}_n - z_1^{-1}\boldsymbol{A}_1 - z_2^{-1}\boldsymbol{A}_2\right)^{-1}. \end{aligned}$$

The term d in (31) and its sensitivity are independent on the LSS coordinate, and therefore they are neglected here.

Definition 3: 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_2 norm of $\boldsymbol{X}(z_1, z_2)$ is then defined as

$$\begin{aligned} &\|\boldsymbol{X}(z_1, z_2)\|_2 \\ &= \left[\frac{1}{(2\pi)^2}\int_0^{2\pi}\int_0^{2\pi}\sum_{p=1}^{m}\sum_{q=1}^{n}\left|x_{pq}(e^{j\omega_1}, e^{j\omega_2})\right|^2 d\omega_1 d\omega_2\right]^{\frac{1}{2}} \\ &= \left(\mathrm{tr}\left[\frac{1}{(2\pi j)^2}\oint_{\Gamma_1}\oint_{\Gamma_2}\boldsymbol{X}(z_1, z_2)\boldsymbol{X}^*(z_1, z_2)\frac{dz_1 dz_2}{z_1 z_2}\right]\right)^{\frac{1}{2}} \end{aligned} \tag{34}$$

where $\Gamma_i = \{z_i : |z_i| = 1\}$ for $i = 1, 2$.

From (33) and Definition 3, the overall L_2-sensitivity measure for the 2-D filter in (30) is evaluated by

$$\begin{aligned} S_2 &= \sum_{k=1}^{2}\left\|\frac{\partial H(z_1, z_2)}{\partial \boldsymbol{A}_k}\right\|_2^2 + \left\|\frac{\partial H(z_1, z_2)}{\partial \boldsymbol{b}}\right\|_2^2 \\ &\quad + \sum_{k=1}^{2}\left\|\frac{\partial H(z_1, z_2)}{\partial \boldsymbol{c}_k^T}\right\|_2^2 \\ &= 2\left\|\left[\boldsymbol{F}(z_1, z_2)\boldsymbol{G}(z_1, z_2)\right]^T\right\|_2^2 + \left\|\boldsymbol{G}^T(z_1, z_2)\right\|_2^2 \\ &\quad + 2\left\|\boldsymbol{F}(z_1, z_2)\right\|_2^2. \end{aligned} \tag{35}$$

The L_2-sensitivity measure in (35) can be written as

$$S_2 = 2\mathrm{tr}\left[\boldsymbol{M}(\boldsymbol{I}_n)\right] + \mathrm{tr}[\boldsymbol{W}_o] + 2\mathrm{tr}[\boldsymbol{K}_c] \tag{36}$$

where

$$\begin{aligned} \boldsymbol{K}_c &= \frac{1}{(2\pi j)^2}\oint_{\Gamma_1}\oint_{\Gamma_2}\boldsymbol{F}(z_1, z_2)\boldsymbol{F}^T\left(z_1^{-1}, z_2^{-1}\right)\frac{dz_1 dz_2}{z_1 z_2} \\ \boldsymbol{W}_o &= \frac{1}{(2\pi j)^2}\oint_{\Gamma_1}\oint_{\Gamma_2}\boldsymbol{G}^T(z_1, z_2)\boldsymbol{G}\left(z_1^{-1}, z_2^{-1}\right)\frac{dz_1 dz_2}{z_1 z_2} \end{aligned}$$

$$\boldsymbol{M}(\boldsymbol{P}) = \frac{1}{(2\pi j)^2} \oint_{\Gamma_1} \oint_{\Gamma_2} [\boldsymbol{F}(z_1, z_2)\boldsymbol{G}(z_1, z_2)]^T \boldsymbol{P}^{-1} \cdot \boldsymbol{F}\left(z_1^{-1}, z_2^{-1}\right) \boldsymbol{G}\left(z_1^{-1}, z_2^{-1}\right) \frac{dz_1 dz_2}{z_1 z_2}.$$

Matrices $\boldsymbol{K}_c$, $\boldsymbol{W}_o$, and $\boldsymbol{M}(\boldsymbol{P})$ are called 2-D Gramians and can be derived from

$$\begin{aligned} \boldsymbol{K}_c &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \boldsymbol{f}(i,j)\boldsymbol{f}^T(i,j) \\ \boldsymbol{W}_o &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \boldsymbol{g}^T(i,j)\boldsymbol{g}(i,j) \\ \boldsymbol{M}(\boldsymbol{P}) &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \boldsymbol{H}^T(i,j)\boldsymbol{P}^{-1}\boldsymbol{H}(i,j) \end{aligned} \tag{37}$$

where

$$\begin{aligned} \boldsymbol{f}(i,j) &= \boldsymbol{A}^{(i,j)}\boldsymbol{b} \\ \boldsymbol{g}(i,j) &= \boldsymbol{c}_1\boldsymbol{A}^{(i-1,j)} + \boldsymbol{c}_2\boldsymbol{A}^{(i,j-1)} \\ \boldsymbol{A}^{(0,0)} &= \boldsymbol{I}_n, \quad \boldsymbol{A}^{(i,j)} = \boldsymbol{0}, \quad i<0 \text{ or } j<0 \\ \boldsymbol{A}^{(i,j)} &= \boldsymbol{A}_1\boldsymbol{A}^{(i-1,j)} + \boldsymbol{A}_2\boldsymbol{A}^{(i,j-1)} \\ &= \boldsymbol{A}^{(i-1,j)}\boldsymbol{A}_1 + \boldsymbol{A}^{(i,j-1)}\boldsymbol{A}_2, \quad (i,j) > (0,0) \\ \boldsymbol{H}(i,j) &= \sum_{(0,0)\le(k,r)\le(i,j)}\sum \boldsymbol{f}(k,r)\boldsymbol{g}(i-k, j-r) \end{aligned}$$

with the partial ordering for integer pairs (i,j) used in [27, p. 2].

If a coordinate transformation defined by

$$\overline{\boldsymbol{x}}(i,j) = \boldsymbol{T}^{-1}\boldsymbol{x}(i,j) \tag{38}$$

is applied to the 2-D filter in (30), we obtain a new realization $(\overline{\boldsymbol{A}}_1, \overline{\boldsymbol{A}}_2, \overline{\boldsymbol{b}}, \overline{\boldsymbol{c}}_1, \overline{\boldsymbol{c}}_2, d)_n$ characterized by

$$\begin{aligned} \overline{\boldsymbol{A}}_1 &= \boldsymbol{T}^{-1}\boldsymbol{A}_1\boldsymbol{T}, \quad \overline{\boldsymbol{A}}_2 = \boldsymbol{T}^{-1}\boldsymbol{A}_2\boldsymbol{T} \\ \overline{\boldsymbol{b}} &= \boldsymbol{T}^{-1}\boldsymbol{b}, \quad \overline{\boldsymbol{c}}_1 = \boldsymbol{c}_1\boldsymbol{T}, \quad \overline{\boldsymbol{c}}_2 = \boldsymbol{c}_2\boldsymbol{T} \\ \overline{\boldsymbol{K}}_c &= \boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T}, \quad \overline{\boldsymbol{W}}_o = \boldsymbol{T}^T\boldsymbol{W}_o\boldsymbol{T}. \end{aligned} \tag{39}$$

Noting that the coordinate transformation in (38) transforms the Gramian $\boldsymbol{M}(\boldsymbol{I}_n)$ into $\boldsymbol{T}^T\boldsymbol{M}(\boldsymbol{P})\boldsymbol{T}$, (36) becomes

$$S_2(\boldsymbol{P}) = 2\mathrm{tr}\,[\boldsymbol{M}(\boldsymbol{P})\boldsymbol{P}] + \mathrm{tr}[\boldsymbol{W}_o\boldsymbol{P}] + 2\mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}] \tag{40}$$

where $\boldsymbol{P} = \boldsymbol{T}\boldsymbol{T}^T$. Moreover, if the L_2-norm dynamic-range scaling constraints are imposed on the LSS vector $\overline{\boldsymbol{x}}(i,j)$, then

$$(\overline{\boldsymbol{K}}_c)_{ii} = (\boldsymbol{T}^{-1}\boldsymbol{K}_c\boldsymbol{T}^{-T})_{ii} = 1, \qquad \text{for } i = 1, 2, \cdots, n \tag{41}$$

are required.

The problem considered here is as follows: *For given* $\boldsymbol{A}_1$, $\boldsymbol{A}_2$, $\boldsymbol{b}$, $\boldsymbol{c}_1$, *and* $\boldsymbol{c}_2$, *obtain an* $n \times n$ *nonsingular matrix* $\boldsymbol{T}$ *which minimizes* (40) *subject to the scaling constraints in* (41).

B. L_2-Sensitivity Minimization

In order to minimize (40) over an $n \times n$ symmetric positive-definite matrix $\boldsymbol{P}$ subject to the "relaxed" constraint $\mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}] = n$ derived from (41), we define the Lagrange function

$$J_2(\boldsymbol{P}, \lambda) = 2\mathrm{tr}\,[\boldsymbol{M}(\boldsymbol{P})\boldsymbol{P}] + \mathrm{tr}[\boldsymbol{W}_o\boldsymbol{P}] + 2\mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}] + \lambda\left(\mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}] - n\right) \tag{42}$$

where λ is a Lagrange multiplier. As is well known, the solution of the minimization problem of (42) must satisfy the KKT conditions $\partial J_2(\boldsymbol{P}, \lambda)/\partial \boldsymbol{P} = \boldsymbol{0}$ and $\partial_2(\boldsymbol{P}, \lambda)/\partial\lambda = 0$, where the gradients are found to be [24]

$$\begin{aligned} \frac{\partial J_2(\boldsymbol{P}, \lambda)}{\partial \boldsymbol{P}} &= 2\boldsymbol{M}(\boldsymbol{P}) - 2\boldsymbol{P}^{-1}\boldsymbol{N}(\boldsymbol{P})\boldsymbol{P}^{-1} \\ &\quad + \boldsymbol{W}_o - (\lambda+2)\boldsymbol{P}^{-1}\boldsymbol{K}_c\boldsymbol{P}^{-1} \\ \frac{\partial J_2(\boldsymbol{P}, \lambda)}{\partial \lambda} &= \mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}] - n \end{aligned} \tag{43}$$

where $\boldsymbol{N}(\boldsymbol{P})$ is derived from

$$\boldsymbol{N}(\boldsymbol{P}) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \boldsymbol{H}(i,j)\boldsymbol{P}\boldsymbol{H}^T(i,j).$$

Hence the KKT conditions become

$$\begin{aligned} \boldsymbol{P}\boldsymbol{F}(\boldsymbol{P})\boldsymbol{P} &= \boldsymbol{G}(\boldsymbol{P}, \lambda) \\ \mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}] &= n \end{aligned} \tag{44}$$

where

$$\begin{aligned} \boldsymbol{F}(\boldsymbol{P}) &= 2\boldsymbol{M}(\boldsymbol{P}) + \boldsymbol{W}_o \\ \boldsymbol{G}(\boldsymbol{P}, \lambda) &= 2\boldsymbol{N}(\boldsymbol{P}) + (\lambda+2)\boldsymbol{K}_c. \end{aligned}$$

The first equation in (44) is highly nonlinear with respect to $\boldsymbol{P}$. An effective approach to solving the first equation in (44) is to *relax* it into the following recursive second-order matrix equation:

$$\boldsymbol{P}_{i+1}\boldsymbol{F}(\boldsymbol{P}_i)\boldsymbol{P}_{i+1} = \boldsymbol{G}(\boldsymbol{P}_i, \lambda_i) \tag{45}$$

where $\boldsymbol{P}_i$ is assumed to be known from the previous recursion and then the solution $\boldsymbol{P}_{i+1}$ is given by

$$\boldsymbol{P}_{i+1} = \boldsymbol{F}(\boldsymbol{P}_i)^{-\frac{1}{2}}\left[\boldsymbol{F}(\boldsymbol{P}_i)^{\frac{1}{2}}\boldsymbol{G}(\boldsymbol{P}_i, \lambda_i)\boldsymbol{F}(\boldsymbol{P}_i)^{\frac{1}{2}}\right]^{\frac{1}{2}}\boldsymbol{F}(\boldsymbol{P}_i)^{-\frac{1}{2}}. \tag{46}$$

To derive a recursive formula for the Lagrange multiplier λ, we use (44) to write

$$\mathrm{tr}\,[\boldsymbol{P}\boldsymbol{F}(\boldsymbol{P})] = 2\mathrm{tr}\left[\boldsymbol{N}(\boldsymbol{P})\boldsymbol{P}^{-1}\right] + n(\lambda+2) \tag{47}$$

that naturally reveals the following recursion for λ:

$$\lambda_{i+1} = \frac{\mathrm{tr}\,[\boldsymbol{P}_i\boldsymbol{F}(\boldsymbol{P}_i)] - 2\mathrm{tr}\left[\boldsymbol{N}(\boldsymbol{P}_i)\boldsymbol{P}_i^{-1}\right]}{n} - 2. \tag{48}$$

The iteration process starts with $\boldsymbol{P}_1 = \boldsymbol{I}_n$ and any value of $\lambda_1 > 0$ and continues until (44) is satisfied within a prescribed numerical tolerance.

Finally, the process in (25)–(29) is applied to the resulting optimal matrix $\boldsymbol{P}$ in order to construct the optimal transformation matrix $\boldsymbol{T} = \boldsymbol{P}^{1/2}\boldsymbol{R}\boldsymbol{S}^T$ minimizing (40) subject to the constraints in (41).

IV. Numerical Examples

Example 1: Let a state–space digital filter in (1) be specified by

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0.453\,770 & -1.556\,160 & 1.974\,860 \end{bmatrix}$$
$$\boldsymbol{b} = [0 \quad 0 \quad 0.242\,096]^T$$
$$\boldsymbol{c} = [0.095\,706 \quad 0.095\,086 \quad 0.327\,556]$$
$$d = 0.015\,940.$$

Performing the computation of (7), (12) and (13), the Gramians $\boldsymbol{K}_c$, $\boldsymbol{W}_o$, and $\boldsymbol{M}(\boldsymbol{I}_3)$ are calculated as

$$\boldsymbol{K}_c = \begin{bmatrix} 1.000\,000 & 0.872\,501 & 0.562\,821 \\ 0.872\,501 & 1.000\,000 & 0.872\,501 \\ 0.562\,821 & 0.872\,501 & 1.000\,000 \end{bmatrix}$$
$$\boldsymbol{W}_o = \begin{bmatrix} 0.820\,741 & -2.035\,328 & 1.628\,161 \\ -2.035\,328 & 5.307\,273 & -4.264\,903 \\ 1.628\,161 & -4.264\,903 & 3.941\,491 \end{bmatrix}$$
$$\boldsymbol{M}(\boldsymbol{I}_3) = \begin{bmatrix} 8.921\,380 & -22.046\,457 & 17.916\,285 \\ -22.046\,457 & 55.671\,710 & -46.052\,011 \\ 17.916\,285 & -46.052\,011 & 42.522\,082 \end{bmatrix}.$$

The L_2-sensitivity measure in (6) is then computed as

$$S_1 = 120.184\,677.$$

Choosing $\boldsymbol{P}_1 = \boldsymbol{I}_3$ and $\lambda_1 = 100$ in (22) and (24) as the initial estimates, it took the proposed iterative algorithm 500 iterations to converge to

$$\bar{\boldsymbol{P}} = \begin{bmatrix} 2.307\,529 & 1.375\,667 & 0.514\,400 \\ 1.375\,667 & 1.103\,115 & 0.678\,193 \\ 0.514\,400 & 0.678\,193 & 0.666\,912 \end{bmatrix}$$

which yields

$$\boldsymbol{T} = \begin{bmatrix} 0.906\,372 & 0.756\,223 & 0.956\,110 \\ 0.196\,978 & 0.857\,123 & 0.574\,155 \\ -0.369\,823 & 0.597\,630 & 0.415\,910 \end{bmatrix}.$$

In this case, the Gramian $\boldsymbol{M}(\boldsymbol{P})$ is computed from (12) and (13) as

$$\boldsymbol{M}(\boldsymbol{P}) = \begin{bmatrix} 1.908\,677 & -0.301\,984 & -1.313\,686 \\ -0.301\,984 & 1.701\,052 & 0.430\,349 \\ -1.313\,686 & 0.430\,349 & 1.395\,025 \end{bmatrix}$$

and the L_2-sensitivity measure in (10) is minimized subject to the scaling constraints in (14) to

$$S_1(\boldsymbol{P}) = 8.672\,129.$$

Profiles of the L_2-sensitivity, parameter λ, as well as $\mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}]$ during the first 500 iterations of the algorithm are shown in Figs. 2 and 3, respectively. Together, these figures clearly reveal a two-stage convergence behavior of the algorithm in that the first stage (which consists of just one iteration) of the algorithm reduces the L_2-sensitivity drastically without maintaining

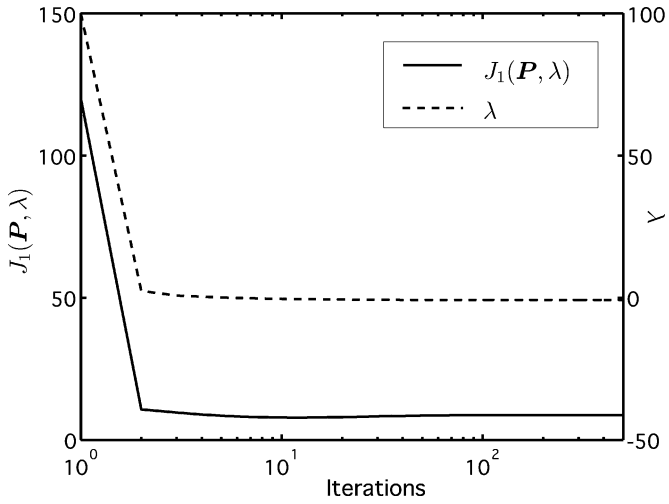

Fig. 2. L_2-sensitivity and λ performances.

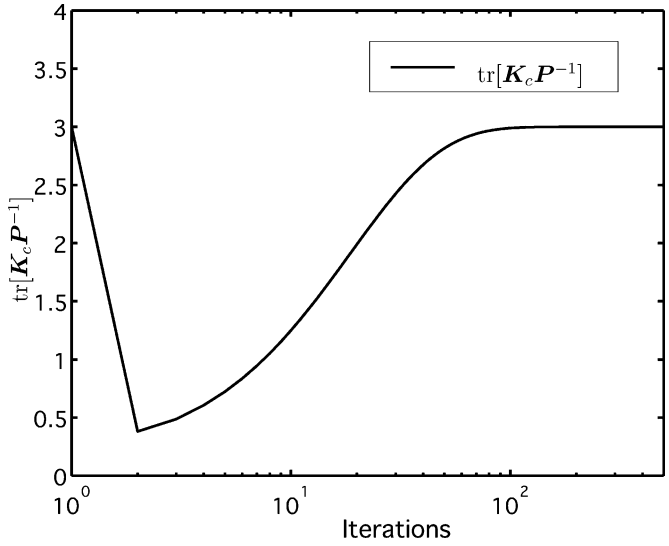

Fig. 3. $\mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}]$ performance.

the constraint $\mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}] = n$, and the second stage of the algorithm is able to restore the constraint $\mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}] = n$ while further reducing the L_2-sensitivity slightly.

For comparison purposes, only the iterative algorithm in (22) is applied by letting $\lambda_i = 0$ for any i and setting $\boldsymbol{P}_1 = \boldsymbol{I}_3$ in order to minimize the L_2-sensivivity measure in (10) (without considering the scaling constraints in (14)), and after 500 iterations it converges to

$$\boldsymbol{P} = \begin{bmatrix} 4.774\,934 & 2.835\,816 & 1.053\,819 \\ 2.835\,816 & 2.287\,705 & 1.415\,049 \\ 1.053\,819 & 1.415\,049 & 1.403\,809 \end{bmatrix}$$

which yields

$$\boldsymbol{T} = \begin{bmatrix} 2.185\,162 & 0.0 & 0.0 \\ 1.297\,760 & 0.776\,868 & 0.0 \\ 0.482\,261 & 1.015\,861 & 0.373\,174 \end{bmatrix}$$

and $S_1(\boldsymbol{P}) = 7.832\,680$. Note that this method is essentially the same as in [10]. The above coordinate transformation matrix $\boldsymbol{T}$ is then scaled by an appropriate nonsingular diagonal matrix, so that the scaling constraints in (14) are satisfied. Then, the result is

$$S_1(\boldsymbol{P}) = 9.822\,372$$

where $\boldsymbol{P} = \boldsymbol{T}\boldsymbol{T}^T$ and

$$\boldsymbol{T} = \begin{bmatrix} 1.0 & 0.0 & 0.0 \\ 0.593\,896 & 0.562\,461 & 0.0 \\ 0.220\,698 & 0.735\,495 & 0.307\,225 \end{bmatrix}.$$

Applying the technique reported in [20] yields

$$S_1(\boldsymbol{P}) = 8.683\,274$$

where

$$\boldsymbol{T} = \begin{bmatrix} 0.913\,655 & -0.857\,313 & 0.877\,296 \\ 0.905\,773 & -0.121\,938 & 0.493\,844 \\ 0.576\,905 & 0.415\,235 & 0.377\,361 \end{bmatrix}.$$

From these results, it is observed that the proposed technique offers a smaller value of the L_2-sensitivity measure subject to the scaling constraints relative to a method (which is essentially the same as in [10]) for performing the scaling so as to satisfy the constraints in (14) after minimizing the L_2-sensitivity measure in (10) as well as the existing method in [20]. It has also turned out that the proposed iterative technique requires less than half amount of computations to attain practically the same convergence accuracy as compared with the method reported in [20].

Example 2: Consider a class of 2-D digital filters in (30) specified by

$$\boldsymbol{A}_1 = \begin{bmatrix} 0 & 0.481\,228 & 0 & 0 \\ 0 & 0 & 0.510\,378 & 0 \\ 0 & 0 & 0 & 0.525\,287 \\ -0.031\,857 & 0.298\,663 & -0.808\,282 & 1.044\,600 \end{bmatrix}$$

$$\boldsymbol{A}_2 = \begin{bmatrix} -0.226\,080 & 0.776\,837 & 0.024\,693 & -0.000\,933 \\ -0.843\,550 & 1.610\,400 & -0.309\,366 & 0.065\,898 \\ -1.260\,339 & 2.005\,100 & -0.453\,220 & 0.203\,118 \\ -1.121\,498 & 1.636\,435 & -0.590\,516 & 0.562\,890 \end{bmatrix}$$

$$\boldsymbol{b} = [0 \quad 0 \quad 0 \quad 0.198\,473]^T$$

$$\boldsymbol{c}_1 = [-0.567\,054 \quad 0.231\,913 \quad 0.197\,016 \quad 0.239\,932]$$

$$\boldsymbol{c}_2 = [0.464\,344 \quad 0.441\,837 \quad -0.061\,100 \quad 0.105\,505]$$

$$d = 0.009\,43.$$

Using (37), the Gramians $\boldsymbol{K}_c, \boldsymbol{W}_o$, and $\boldsymbol{M}(\boldsymbol{I}_4)$ are calculated as

$$\boldsymbol{K}_c = \begin{bmatrix} 1.000\,000 & 0.987\,279 & 0.940\,868 & 0.844\,274 \\ 0.987\,279 & 1.000\,000 & 0.976\,755 & 0.888\,478 \\ 0.940\,868 & 0.976\,755 & 1.000\,000 & 0.952\,963 \\ 0.844\,274 & 0.888\,478 & 0.952\,963 & 1.000\,000 \end{bmatrix}$$

$$\boldsymbol{W}_o = 10\begin{bmatrix} 1.337\,108 & -1.304\,050 & 0.189\,462 & -0.556\,646 \\ -1.304\,050 & 1.637\,345 & -0.429\,399 & 0.576\,183 \\ 0.189\,462 & -0.429\,399 & 2.122\,604 & -2.191\,942 \\ -0.556\,646 & 0.576\,183 & -2.191\,942 & 2.672\,474 \end{bmatrix}$$

$$\boldsymbol{M}(\boldsymbol{I}_4) = 10^3\begin{bmatrix} 1.043\,052 & -1.097\,577 & 0.637\,299 & -0.982\,714 \\ -1.097\,577 & 1.238\,937 & -0.830\,495 & 1.153\,558 \\ 0.637\,299 & -0.830\,495 & 2.324\,302 & -2.574\,574 \\ -0.982\,714 & 1.153\,558 & -2.574\,574 & 3.019\,844 \end{bmatrix}$$

where the infinite sums in (37) were truncated with $(i, j) = (100, 100)$. The L_2-sensitivity measure S_2 in (36) is then computed as

$$S_2 = 15\,337.965\,477.$$

Choosing $\boldsymbol{P}_1 = \boldsymbol{I}_4$ and $\lambda_1 = 100$ in (46) and (48) as the initial estimates, it took the proposed iterative algorithm 2000 iterations to converge to

$$\boldsymbol{P} = \begin{bmatrix} 1.688\,238 & 1.500\,480 & 1.311\,958 & 1.149\,324 \\ 1.500\,480 & 1.373\,665 & 1.224\,597 & 1.077\,089 \\ 1.311\,958 & 1.224\,597 & 1.148\,785 & 1.035\,997 \\ 1.149\,324 & 1.077\,089 & 1.035\,997 & 0.980\,059 \end{bmatrix}$$

which yields

$$\boldsymbol{T} = \begin{bmatrix} -1.174\,654 & 0.198\,729 & 0.493\,644 & 0.158\,893 \\ -1.115\,961 & 0.031\,366 & 0.322\,000 & 0.153\,714 \\ -1.034\,448 & -0.100\,879 & 0.255\,344 & -0.057\,664 \\ -0.962\,590 & -0.009\,915 & 0.107\,523 & -0.204\,501 \end{bmatrix}.$$

In this case, the Gramian $\boldsymbol{M}(\boldsymbol{P})$ is computed from (37) as

$$\boldsymbol{M}(\boldsymbol{P}) = 10^3 \times \begin{bmatrix} 0.605\,548 & -0.599\,700 & 0.142\,673 & -0.347\,188 \\ -0.599\,700 & 0.672\,622 & -0.215\,513 & 0.360\,191 \\ 0.142\,673 & -0.215\,513 & 1.146\,333 & -1.234\,023 \\ -0.347\,188 & 0.360\,191 & -1.234\,023 & 1.478\,310 \end{bmatrix}$$

over $(0, 0) \leq (i, j) \leq (100, 100)$, and the L_2-sensitivity measure in (40) is minimized subject to the scaling constraints in (41) to

$$S_2(\boldsymbol{P}) = 372.776\,304.$$

Profiles of the L_2-sensitivity, parameter λ, as well as $\mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}]$ during the first 2000 iterations of the algorithm are shown in Figs. 4 and 5, respectively. From these figures, a two-stage convergence behavior similar to that of Example 1 can also be observed.

For comparison purposes, only the iterative algorithm in (46) is applied by letting $\lambda_i = 0$ for any i and setting $\boldsymbol{P}_1 = \boldsymbol{I}_4$ in order to minimize the L_2-sensivivity measure in (40) (without considering the scaling constraints in (41)), and after 2000 iterations it converges to

$$\boldsymbol{P} = \begin{bmatrix} 2.095\,327 & 1.862\,416 & 1.628\,491 & 1.426\,572 \\ 1.862\,416 & 1.705\,267 & 1.520\,239 & 1.337\,047 \\ 1.628\,491 & 1.520\,239 & 1.426\,747 & 1.286\,632 \\ 1.426\,572 & 1.337\,047 & 1.286\,632 & 1.217\,307 \end{bmatrix}$$

which yields

$$\boldsymbol{T} = \begin{bmatrix} 1.447\,524 & 0.0 & 0.0 & 0.0 \\ 1.286\,621 & 0.223\,322 & 0.0 & 0.0 \\ 1.125\,018 & 0.325\,839 & 0.234\,330 & 0.0 \\ 0.985\,525 & 0.309\,192 & 0.329\,243 & -0.205\,050 \end{bmatrix}$$

and $S_2(\boldsymbol{P}) = 372.464\,800$. The above coordinate transformation matrix $\boldsymbol{T}$ is then scaled by an appropriate nonsingular diagonal matrix, so that the scaling constraints in (41) are satisfied. Then, the result is

$$S_2(\boldsymbol{P}) = 458.063\,668$$

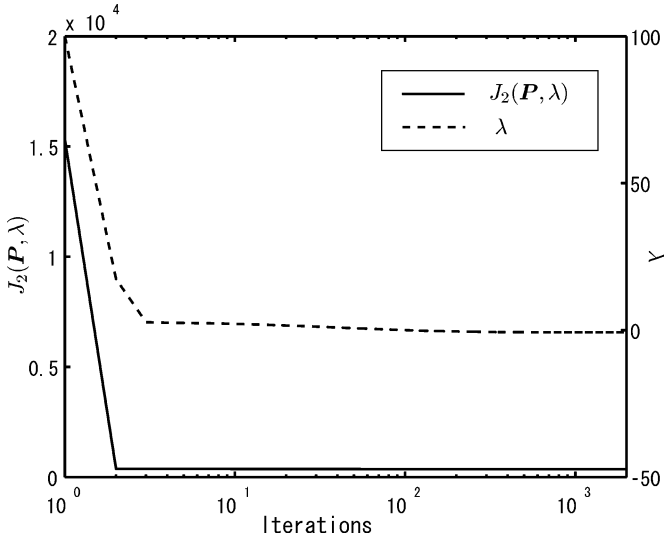

Fig. 4. L_2-sensitivity and λ performances.

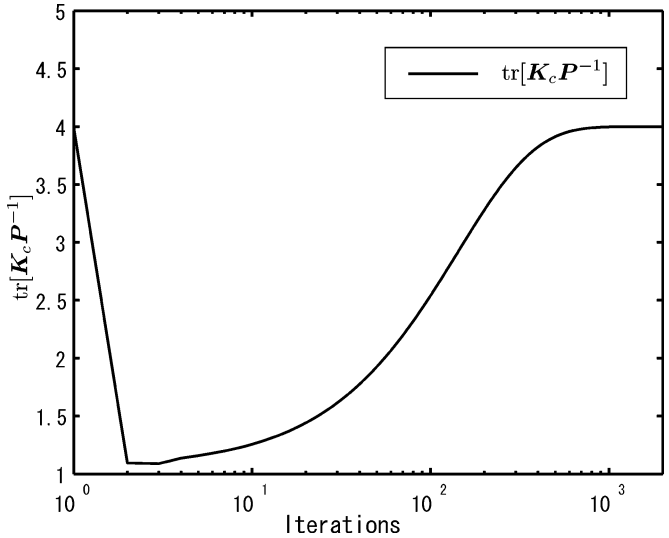

Fig. 5. $\mathrm{tr}[\boldsymbol{K}_c\boldsymbol{P}^{-1}]$ performance.

where $\boldsymbol{P} = \boldsymbol{T}\boldsymbol{T}^T$ and

$$\boldsymbol{T} = \begin{bmatrix} 1.0 & 0.0 & 0.0 & 0.0 \\ 0.888\,843 & 0.187\,005 & 0.0 & 0.0 \\ 0.777\,201 & 0.272\,850 & 0.171\,289 & 0.0 \\ 0.680\,835 & 0.258\,910 & 0.240\,668 & -0.250\,773 \end{bmatrix}.$$

From the above result, it is observed that the constrained optimization technique proposed here offers smaller L_2-sensitivity subject to the scaling constraints relative to a method for carrying out the scaling so as to satisfy the constraints in (41) after minimizing the L_2-sensitivity measure in (40).

V. Conclusion

The problem of minimizing an L_2-sensitivity measure subject to L_2-norm dynamic-range scaling constraints for state–space digital filters as well as a class of 2-D state–space digital filters have been investigated. In each case, a novel iterative algorithm has been developed to solve the constraint optimization problem directly. This has been performed by using a Lagrange function and some matrix-theoretic techniques. The optimal state–space filter structure has been constructed by applying the coordinate-transformation matrix ajusted to satisfy the scaling constraints without altering the L_2-sensitivity. In addition, it has been shown that the 2-D LSS model used to describe a class of 2-D state–space digital filters corresponds to a transposed structure of the Fornasini–Marchesini second LSS model. Our computer simulation results have demonstrated the effectiveness of the proposed techniques compared with the existing methods.

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] V. Tavsanoglu and L. Thiele, "Optimal design of state–space digital filters by simultaneous minimization of sensitivity and roundoff noise," *IEEE Trans. Circuits Syst.*, vol. CAS-31, no. 10, pp. 884–888, Oct. 1984.

[3] L. Thiele, "On the sensitivity of linear state–space systems," *IEEE Trans. Circuits Syst.*, vol. CAS-33, no. 5, pp. 502–510, May 1986.

[4] 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, no. 1, pp. 72–80, Jan. 1989.

[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, no. 5, 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, no. 8, 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 (ICASSP)*, vol. 4, pp. 429–432.

[8] M. Gevers and G. Li, *Parameterizations in Control, Estimation and Filtering Problems: Accuracy Aspects*. New York: Springer-Verlag, 1993.

[9] C. Xiao, "Improved L_2-sensitivity for state–space digital system," *IEEE Trans. Signal Process.*, vol. 45, no. 4, pp. 837–840, Apr. 1997.

[10] T. Hinamoto, S. Yokoyama, T. Inoue, W. Zeng, and W.-S. Lu, "Analysis and minimization of L_2-sensitivity for linear systems and two-dimensional state–space filters using general controllability and observability Gramians," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 49, no. 9, pp. 1279–1289, Sep. 2002.

[11] 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.

[12] 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 Process.*, vol. ASSP-38, no. 9, pp. 1587–1594, Sep. 1990.

[13] 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.

[14] T. Hinamoto and T. Takao, "Synthesis of 2-D state–space filter structures with low frequency-weighted sensitivity," *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.*, vol. 39, no. 9, pp. 646–651, Sep. 1992.

[15] ——, "Minimization of frequency-weighting sensitivity in 2-D systems based on the Fornasini–Marchesini second model," in *Proc. 1992 IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, pp. 401–404.

[16] G. Li, "On frequency weighted minimal L_2 sensitivity of 2-D systems using Fornasini–Marchesini LSS model," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 44, no. 7, pp. 642–646, Jul. 1997.

[17] ——, "Two-dimensional system optimal realizations with L_2-sensitivity minimization," *IEEE Trans. Signal Process.*, vol. 46, no. 3, pp. 809–813, Mar. 1998.

[18] T. Hinamoto, Y. Zempo, Y. Nishino, and W.-S. Lu, "An analytical approach for the synthesis of two-dimensional state–space filter structures with minimum weighted sensitivity," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 46, no. 10, pp. 1172–1183, Oct. 1999.

[19] T. Hinamoto and Y. Sugie, "L_2-sensitivity analysis and minimization of 2-D separable-denominator state–space digital filters," *IEEE Trans. Signal Process.*, vol. 50, no. 12, pp. 3107–3114, Dec. 2002.

[20] T. Hinamoto, H. Ohnishi, and W.-S. Lu, "Minimization of L_2-sensitivity For state–space digital filters subject to L_2-scaling constraints," in *Proc. 2004 IEEE Int. Symp. Circuits Systems*, vol. III, pp. 137–140.

[21] C. T. Mullis and R. A. Roberts, "Synthesis of minimum roundoff noise fixed-point digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-23, no. 9, pp. 551–562, Sep. 1976.
[22] S. Y. Hwang, "Minimum uncorrelated unit noise in state–space digital filtering," *IEEE Trans. Acoust, Speech, Signal Process.*, vol. 25, no. 4, pp. 273–281, Aug. 1977.
[23] T. Kailath, *Linear System*. Englewood Cliffs, NJ: Prentice-Hall, 1980.
[24] L. L. Scharf, *Statistical Signal Processing*. Reading, MA: Addison-Wesley, 1991.
[25] T. Hinamoto, "A novel local state–space model for 2-D digital filters and its properties," in *Proc. 2001 IEEE Int. Symp. Circuits Systems*, vol. 2, pp. 545–548.
[26] E. Fornasini and G. Marchesini, "Doubly-indexed dynamical systems: State–space models and structural properties," *Math Syst. Theory*, vol. 12, pp. 59–72, 1978.
[27] R. P. Roessor, "A discrete state–space model for linear image processing," *IEEE Trans. Autom. Control*, vol. AC-20, no. 1, pp. 1–10, Feb. 1975.

Takao Hinamoto (M'77–SM'84–F'01) received the B.E. degree from Okayama University, Okayama, Japan, in 1969, the M.E. degree from Kobe University, Kobe, Japan, in 1971, and the Dr.Eng. degree from Osaka University, Osaka, Japan, in 1977, all in electrical engineering.

From 1972 to 1988, he was with the Faculty of Engineering, Kobe University. From 1979 to 1981, he was a Visiting Member of Staff in the Department of Electrical Engineering, Queen's University, Kingston, ON, Canada, on leave from Kobe University. From 1988 to 1991, he was a Professor of electronic circuits in the Faculty of Engineering, Tottori University, Tottori, Japan. Since January 1992, he has been a Professor of electronic control in the Department of Electrical Engineering, Hiroshima University, Hiroshima, Japan. His research interests include digital signal processing, system theory, and control engineering. He has published about 350 papers in these areas and is the coeditor and coauthor of the book *Two-Dimensional Signal and Image Processing* (Tokyo, Japan: SICE, 1996).

Dr. Hinamoto served as an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—II: ANALOG AND DIGITAL SIGNAL PROCESSING from 1993 to 1995 and as an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I: FUNDAMENTAL THEORY AND APPLICATIONS from 2002 to 2003. He was the Guest Editor of the special sections on Digital Signal Processing (DSP) in the August 1998 issue as well as Adaptive Signal Processing and Its Applications in the March 2005 issue of the *IEICE Transactions on Fundamentals*. He served as the Chair of the 12th DSP Symposium held in Hiroshima, Japan, in November 1997, sponsored by the DSP Technical Committee of IEICE. He also served as the General Chair of the 47th IEEE International Midwest Symposium on Circuits and Systems held in Hiroshima, Japan, in July 2004. Since 1995, he has been a member of the Steering Committee of the IEEE International Midwest Symposium on Circuits and Systems, and since 1998, a member of the Digital Signal Processing Technical Committee in the IEEE Circuits and Systems Society. He was a member of the Technical Program Committee for the International Symposium of Circuits and Systems (ISCAS) 1999 and an International Coordinator of the Organizing Committee for ISCAS 2004. From 1993 to 2000, he served as Senator or Member of the Board of Directors in the Society of Instrument and Control Engineers (SICE), and from 1999 to 2001, he was the Chair of the Chugoku Chapter of SICE. He played a leading role in establishing the Hiroshima Section of IEEE and served as the Interim Chair of the Section. From June 2003 to May 2004, he served as the Chair of the DSP Technical Committee of IEICE and the Chair of the Chugoku Chapter of IEICE. He is a recipient of the IEEE Third Millennium Medal and a Fellow of IEICE.

Ken-ichi Iwata received the B.E. and M.E. degrees in electrical engineering from Hiroshima University, Hiroshima, Japan, in 2003 and 2005, respectively.

He was engaged in research on digital signal processing during his graduate studies. Since April 2005, he has been with Mitsubishi Electric Corporation, Himeji, Japan.

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

He was a Postdoctoral Fellow at the University of Victoria, Victoria, BC, Canada, in 1985 and a Visiting Assistant Professor with the University of Minnesota in 1986. Since 1987, he has been with the University of Victoria, where he is currently a Professor. His current teaching and research interests are in the general areas of digital signal processing and application of optimization methods. He is the coauthor with A. Antoniou of the book *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 Editor of the same journal from 1990 to 1992. He served as an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS, PART II from 1993 to 1995 and for Part I of the same journal from 1999 to 2001. Presently, he is serving as Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS, PART I, and the *International Journal of Multidimensional Systems and Signal Processing*.