AIAA JOURNAL
Vol. 40, No. 11, November 2002

Balanced Model Reduction via the Proper Orthogonal Decomposition

K. Willcox* and J. Peraire†
Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

A new method for performing a balanced reduction of a high-order linear system is presented. The technique combines the proper orthogonal decomposition and concepts from balanced realization theory. The method of snapshots is used to obtain low-rank, reduced-range approximations to the system controllability and observability grammians in either the time or frequency domain. The approximations are then used to obtain a balanced reduced-order model. The method is particularly effective when a small number of outputs is of interest. It is demonstrated for a linearized high-order system that models unsteady motion of a two-dimensional airfoil. Computation of the exact grammians would be impractical for such a large system. For this problem, very accurate reduced-order models are obtained that capture the required dynamics with just three states. The new models exhibit far superior performance than those derived using a conventional proper orthogonal decomposition. Although further development is necessary, the concept also extends to nonlinear systems.

Nomenclature

h	=	airfoil plunge displacement
K	=	proper orthogonal decomposition (POD) kernel
m	=	number of POD snapshots
n	=	number of states in computational fluid dynamics (CFD) model
n_r	=	number of states in reduced-order model
R	=	correlation matrix
T	=	matrix whose columns contain the balancing transformation vectors
$\boldsymbol{u}, \boldsymbol{U}$	=	vector containing inputs for models, time and frequency domain
W_c	=	controllability grammian
W_{co}	=	grammian product
W_o	=	observability grammian
$\boldsymbol{x}, \boldsymbol{X}$	=	aerodynamic state vector for CFD model, time and frequency domain
$\boldsymbol{x}_r$	=	aerodynamic state vector for reduced-order model
$\boldsymbol{y}, \boldsymbol{Y}$	=	vector containing outputs of CFD model, time and frequency domain
$\boldsymbol{y}_r$	=	vector containing outputs of reduced-order model
z	=	dual state vector for CFD model
σ_i	=	ith Hankel singular value
Ψ	=	basis vector
ω	=	forcing frequency

Introduction

MODEL reduction is a powerful tool that has been applied throughout many different disciplines, including controls, fluid dynamics, and structural dynamics. In many situations, high-order, complicated numerical models accurately represent the problem at hand, but are unsuitable for the desired application, for instance, for optimization or for control design. In particular, aeroservoelastic design can not be effected without low-order, high-fidelity models. Ideally, we would like to develop a model with a low number of states but that captures the system dynamics accurately over a range of frequencies and forcing inputs. This can be achieved via reduced-order modeling in which a high-order, high-fidelity model is projected onto a reduced-space basis. If the basis is chosen appropriately, the relevant high-fidelity system dynamics can be captured with a greatly reduced number of states. The range of validity of the reduced-order model is determined by the specifics of the reduction procedure.

Many methods have been suggested for determining an appropriate basis. Much of the work has been derived in a controls context. In particular, the idea of a balanced truncation has been shown to provide accurate low-order representations of state-space systems.[1] To determine the balanced realization, it is necessary to compute the grammians of the system, which use information pertaining to both system inputs and outputs. Although it is relatively straightforward to compute these matrices in a controls setting where the system order is moderate, the technique does not extend easily to high-order systems, where state orders exceed 10^4. For this reason, many of the control-based reduction concepts have not been transferred to other disciplines where model order is typically much higher, such as computational fluid dynamics (CFD). Several methods have been developed for computing approximations to the grammians for large systems, including the approximate subspace iteration,[2] least squares approximation,[3] and Krylov subspace methods[4,5]; however, these algorithms are complicated, computationally intensive, and restricted to linear systems. As an alternative means of performing the reduction, Padé approximations have also been used to approximate the system transfer function[6]; however, the resulting reduced-order models often suffer from instability.

The challenge has, therefore, been to develop effective reduction procedures suitable for very high-order systems. Much progress in this area has been made in the field of fluid dynamics.[7,8] One possibility for a basis is to compute the eigenmodes of the system.[9–11] Along with the use of static corrections,[12] this approach can lead to efficient models, and the eigenmodes themselves often lend physical insight to the problem. However, for these high-order systems, solution of such a large nonsymmetric eigenproblem is in itself a very difficult task and in many cases not a viable option. The proper orthogonal decomposition (POD) technique, also known as Karhunen–Loéve expansions (see Ref. 13), has been developed as an alternate method of deriving basis vectors for high-order systems and, in particular, has been widely applied to fluid dynamic problems.[14–16] Frequency-domain POD methods have also been developed and applied to a variety of flow problems.[17–19] However, in existing applications, only information pertaining to system inputs has been considered.

Presented as Paper 2001-2611 at the AIAA 15th Computational Fluid Dynamics Conference, Anaheim, CA, 11–14 June 2001; received 21 June 2001; revision received 28 March 2002; accepted for publication 15 June 2002.

*Assistant Professor, Aeronautics and Astronautics, Room 37-447, Aerospace Computational Design Laboratory; kwillcox@mit.edu. Member AIAA.

†Professor, Aeronautics and Astronautics, Room 37-451, Aerospace Computational Design Laboratory; peraire@mit.edu. Member AIAA.

The failure of current flow model reduction techniques to consider system outputs raises several issues. First, the typical goal of model reduction is to represent a particular system output accurately. When attempts are made to capture the appropriate dynamics without consideration of the relationship between this output and the system states, inefficient models may be obtained. If this additional information could be incorporated to the reduction procedure, it is expected that higher degrees of reduction could be achieved. Second, and more fundamental, a reduction procedure based only on system inputs may be highly dependent on the arbitrary scaling of the state variables. This could lead to reduced-space representations that are highly inaccurate.

Dowell and Hall[20] discuss the possibility of extracting eigenmodes or balanced modes directly from a reduced-order model derived using the POD. Although this approach might further reduce the order of the model, it does not take account of system inputs at the same fundamental level as the outputs. In fact, the balanced modes would be derived from a precomputed unbalanced subspace. In this work, a method will be presented that allows both inputs and outputs to be taken into account to obtain a balanced reduced-order model directly from a high-order system without the need for an intermediate reduced-order model. Lall et al.[21] noted the connection between the system grammians and the POD and used a Kahunen–Loéve decomposition to obtain an approximate balanced truncation. Here, we make use of a similar concept that does not require the construction of the approximate grammians, which in our case would be computationally prohibitive. Instead, the POD method of snapshots[15] will be used to approximate the grammians of the system in a very efficient way that does not require large computations or complicated algorithms. The method can also be implemented in the frequency domain, making it even more computationally efficient.

In this paper, the existing concepts of POD and balanced realization will be outlined. The new method that combines the two approaches will then be presented. Results will be shown for reduction of two high-order systems. The first case analyzed is a randomly generated state-space system whose exact balanced realization can be computed, allowing some insight to the performance of the method. The second example is the reduction of a CFD model that describes the unsteady linearized motion of a two-dimensional airfoil. In this case, reduction results will be compared to a full simulation of the CFD model and also to a conventional POD reduction approach. We then briefly discuss extension of the methodology to nonlinear problems, and, finally, we present some conclusions.

Model Order Reduction

Consider an nth-order linear system

$$\dot{\boldsymbol{x}} = A\boldsymbol{x} + B\boldsymbol{u} \tag{1}$$

$$\boldsymbol{y} = C\boldsymbol{x} \tag{2}$$

where $\boldsymbol{x}$ is the state vector, the vectors $\boldsymbol{u}$ and $\boldsymbol{y}$ contain the system inputs and outputs, respectively, and the order of the system n is high. The objective of the reduction procedure is to determine an n_rth-order reduced-space basis onto which the state vector can be projected, that is, $\boldsymbol{x} = V\boldsymbol{x}_r$, and an orthonormal set $\tilde{V}$, so that $\tilde{V}V = I$. This basis is chosen appropriately so that the reduced-order system

$$\dot{\boldsymbol{x}}_r = \tilde{V}AV\boldsymbol{x}_r + \tilde{V}B\boldsymbol{u} \tag{3}$$

$$\boldsymbol{y}_r = CV\boldsymbol{x}_r \tag{4}$$

accurately reproduces the desired dynamics of the original system (1) and (2) with many fewer states ($n_r \ll n$).

Balanced Truncation

The concept of a balanced truncation of a system was first introduced by Moore.[1] The underlying idea is to take account of both the inputs and outputs of the system when determining which states to retain in the reduced-state representation, but to do so with appropriate internal scaling. This scaling is important because a particular representation of the system is not unique: Any nonsingular linear transformation can be applied to the system (1) and (2) that effectively allows each state to be scaled by an arbitrary amount. For example, if we choose the transformation $\boldsymbol{x} = T\boldsymbol{x}_t$, we obtain the scaled system

$$\dot{\boldsymbol{x}}_t = T^{-1}AT\boldsymbol{x}_t + T^{-1}B\boldsymbol{u} \tag{5}$$

$$\boldsymbol{y} = CT\boldsymbol{x}_t \tag{6}$$

that is fully equivalent to Eqs. (1) and (2). A reduction procedure based only on system inputs or outputs may be heavily dependent on this arbitrary scaling of the states. In a balanced truncation, we, therefore, seek a reduction method that is independent of the particular system scaling.

Reduction of the system will be achieved by retaining only certain states in the representation. This is equivalent to defining a certain subspace within the state space. Two important subspaces are the controllable and observable subspaces. The controllable subspace is that set of states that can be obtained with zero initial state and a given input $\boldsymbol{u}(t)$ (also called the set of reachable states), whereas the observable subspace comprises those states that as initial conditions could produce a nonzero output $\boldsymbol{y}(t)$ with no external input. The controllability and observability grammians are each an $n \times n$ matrix whose eigenvectors span the controllable and observable subspaces, respectively. These matrices are defined for the linear system (1) and (2) as

$$W_c = \int_0^\infty e^{At}BB^*e^{A^*t}\,\mathrm{d}t \tag{7}$$

$$W_o = \int_0^\infty e^{A^*t}C^*Ce^{At}\,\mathrm{d}t \tag{8}$$

where the asterisk denotes the complex conjugate transpose.

When it is noted that for a single-input/single-output (SISO) system the quantity $\boldsymbol{x}_\delta(t) = e^{At}B$ is simply the impulse response of the system [set $u(t) = \delta(t)$ in Eq. (1)], the controllability grammian can also be written

$$W_c = \int_0^\infty \boldsymbol{x}_\delta(t)\boldsymbol{x}_\delta^*(t)\,\mathrm{d}t \tag{9}$$

For the observability grammian, we need to consider the dual system (1) and (2):

$$\dot{\boldsymbol{z}} = A^*\boldsymbol{z} + C^*\boldsymbol{u}_d \tag{10}$$

$$\boldsymbol{y}_d = B^*\boldsymbol{z} \tag{11}$$

Here, $\boldsymbol{z}$ is the dual state vector. Analogously to Eq. (9), we can write

$$W_o = \int_0^\infty z_\delta(t)z_\delta^*(t)\,\mathrm{d}t \tag{12}$$

where $z_\delta(t) = e^{A^*t}C^*$ is the impulse response of the dual SISO system.

To obtain a balanced realization of the system (1) and (2), a state transformation T is chosen so that the controllability and observability grammians are diagonal and equal. This so-called balancing transformation can be computed by first calculating the matrix $W_{co} = W_cW_o$ and then determining its eigenmodes:

$$W_{co} = T\Lambda T^{-1} \tag{13}$$

It can be seen that the eigenvectors of W_{co}, that is, $\boldsymbol{T}_i$, are the basis vectors that describe the balancing transformation, as follows. Consider the grammians of the transformed system (5) and (6). By replacing A with $\tilde{A} = T^{-1}AT$ and B with $\tilde{B} = T^{-1}B$ in Eq. (7), we obtain the following expression for the controllability grammian of the transformed system:

$$\tilde{W}_c = T^{-1}W_cT^{-1*} \tag{14}$$

Similarly, in Eq. (8), we replace A with $\tilde{A}$ and C with $\tilde{C} = CT$ to obtain

$$\tilde{W}_o = T^* W_o T \tag{15}$$

For a balanced system, we require $\tilde{W}_c = \tilde{W}_o = \Sigma$, where Σ is a diagonal matrix. From Eqs. (14) and (15), we can write

$$T^{-1} W_c = \Sigma T^* \tag{16}$$

$$W_o T = T^{-1*} \Sigma \tag{17}$$

or

$$T^{-1} W_c W_o T = \Sigma^2 \tag{18}$$

From Eq. (18), it can be seen that the transformation T that balances the system is that containing the eigenvectors of the grammian product W_{co} as its columns.

When Eqs. (18) and (13) are compared, it can also be seen that the eigenvalues λ_i contained in the diagonal matrix Λ are positive, real numbers, and $\sigma_i = \sqrt{\lambda_i}$ are known as the Hankel singular values of the system. The eigenvectors of W_{co} correspond to states through which the input is transmitted to the output. The magnitudes of the Hankel singular values describe the relative importance of these states and are independent of the particular realization of the system. In a balanced truncation, only those states are retained that correspond to large Hankel singular values.

When it is assumed that the nth-order system (1) and (2) has been transformed to a balanced realization, an error criterion for model reduction based on Hankel singular values can be derived.[22] A truncation of the balanced system is performed in which the first n_r states are retained, resulting in a reduced-order model of the form (3) and (4). The Hankel singular values of the neglected states give an error bound on the output:

$$\| \boldsymbol{y}(t) - \boldsymbol{y}_r(t) \| \leq 2 \sum_{i = n_r + 1}^{n} \sigma_i \| \boldsymbol{u}(t) \| \tag{19}$$

where $\| \ \|$ denotes the L_2 norm.

For large systems, it is not practical to compute the grammians explicitly using Eqs. (7) and (8). It is more convenient to note that W_c and W_o satisfy the Lyapunov equations

$$A W_c + W_c A^* + B B^* = 0 \tag{20}$$

$$A^* W_o + W_o A + C^* C = 0 \tag{21}$$

See Ref. 23 for more details. Methods have been suggested for solving Eqs. (20) and (21) when the systems are large, by the use of approximate subspace iteration,[2] least-squares approximation,[3] and Krylov subspace methods (see Refs. 4 and 5), to obtain low-rank approximations of the grammians. However, all of these techniques are expensive for very high-order systems and have only been demonstrated for much smaller problems than those encountered in complicated fluid dynamic applications. In this paper, we will introduce an efficient method for calculating very low-rank approximations to the grammians using the concepts of the POD that are outlined in the following section.

POD

The POD has been widely used to determine efficient bases for dynamic systems. It was introduced for the analysis of turbulence by Lumley[14] and is also known as the Karhunen-Loéve decomposition (see Ref. 13) and principal component analysis.[24] The basis vectors $\boldsymbol{\Psi}$ are chosen to maximize the following cost[16]:

$$\max_{\Phi} \langle |(\boldsymbol{x}, \boldsymbol{\Phi})|^2 \rangle / (\boldsymbol{\Phi}, \boldsymbol{\Phi}) = \langle |(\boldsymbol{x}, \boldsymbol{\Psi})|^2 \rangle / (\boldsymbol{\Psi}, \boldsymbol{\Psi}) \tag{22}$$

where $(\boldsymbol{x}, \boldsymbol{\Psi})$ denotes the scalar product of the basis vector $\boldsymbol{\Psi}$ with the state vector $\boldsymbol{x}(\theta, t)$, which depends on the spatial coordinates θ and time t, and where $\langle \ \rangle$ represents a time-averaging operation. It can be shown that a necessary condition for Eq. (22) to hold is that $\boldsymbol{\Psi}$ is an eigenfunction of the kernel K defined by

$$K(\theta, \theta') = \langle \boldsymbol{x}(\theta, t) \boldsymbol{x}^*(\theta', t) \rangle \tag{23}$$

Sirovich introduced the method of snapshots as a way of determining the eigenfunctions $\boldsymbol{\Psi}$ without explicitly calculating the kernel K (Ref. 15). The kernel can be approximated as

$$K(\theta, \theta') = \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{x}_i(\theta) \boldsymbol{x}_i^*(\theta') \tag{24}$$

where $\boldsymbol{x}_i(\theta)$ is the instantaneous system state or snapshot at a time t_i and the number of snapshots m is sufficiently large. The eigenvectors of K are of the form

$$\boldsymbol{\Psi} = \sum_{i=1}^{m} \beta_i \boldsymbol{x}_i \tag{25}$$

where the constants β_i can be seen to satisfy the eigenvector equation

$$R\beta = \Lambda\beta \tag{26}$$

and R is now the correlation matrix

$$R_{ik} = (1/m)(\boldsymbol{x}_i, \boldsymbol{x}_k) \tag{27}$$

Rather than performing a set of simulations to obtain the snapshots $\boldsymbol{x}_i$, the POD basis vectors can be obtained much more efficiently by taking advantage of linearity and the frequency domain. For a linear system, any general forcing function can be considered as a superposition of sinusoidally time-varying components each at a frequency ω:

$$\boldsymbol{u}(t) = Re\left\{ \int_{-\infty}^{\infty} \boldsymbol{U}(\omega) e^{j\omega t} \, \mathrm{d}\omega \right\} \tag{28}$$

Because the system is linear, the component of forcing at frequency ω induces a response that is also harmonic with frequency ω, that is, $\boldsymbol{x}(t) = \boldsymbol{X} e^{j\omega t}$ and $\boldsymbol{y} = \boldsymbol{Y} e^{j\omega t}$. The response due to each harmonic component could be computed separately and then recombined appropriately to obtain the overall response to the general forcing function. When a single temporal harmonic ω_k is considered, the state-space system (1) and (2) can be written in the frequency domain as

$$\boldsymbol{X}_k = (j\omega_k I - A)^{-1} B \boldsymbol{U}_k \tag{29}$$

$$\boldsymbol{Y}_k = C \boldsymbol{X}_k \tag{30}$$

As shown in Ref. 17, Eqs. (22) and (23) can be converted to the frequency domain using Parseval's theorem. The result is an approximation to the kernel similar to Eq. (24), but the summation over temporal states is replaced with a summation over frequencies. This can also be seen by considering an arbitrary snapshot $\boldsymbol{x}_i$ in Eq. (24), which is a linear combination of harmonic components. Because of temporal orthogonality, the cross products of components at different frequencies will be zero, thus they result in the same conclusion: The kernel can be approximated by a summation over frequencies.

The POD snapshots can, therefore, be obtained by choosing a set of sample frequencies $\{\omega_k\}$ based on the frequency content of problems of interest and solving the frequency-domain system (29) to obtain the responses $\{\boldsymbol{X}_k\}$. The resulting complex response can be used in a frequency-domain POD analysis as in Ref. 17, or the real and imaginary part of each complex response can be used as snapshots in a time domain POD analysis as in Ref. 19.

Balanced Truncation via the Method of Snapshots

Lall et al.[21] describe the connection between the POD and balanced truncation. Note the similarity between the POD kernel function K defined by Eq. (23) and the controllability grammian W_c defined by Eq. (9). In fact, as noted by Lall et al., if the fields $\boldsymbol{x}(\theta, t)$ in Eq. (23) are obtained by exciting the system with impulsive inputs, then the POD results in the construction of the controllability grammian. Further insight can be gained by considering the problem in the frequency domain. Consider an impulsive input: If $u(t) = \delta(t)$, then $U(\omega) = 1$ for all values of ω. From Eq. (29), the frequency-domain impulse response is, therefore, given by $X(\omega) = (j\omega I - A)^{-1}B$. Using Parseval's theorem to convert Eq. (7) to the frequency domain, the controllability grammian can be written as

$$W_c = \frac{1}{2\pi}\int_{-\infty}^{\infty}(j\omega I - A)^{-1}BB^*(-j\omega I - A^*)^{-1}\,\mathrm{d}\omega \tag{31}$$

This alternate representation of the grammian is also considered in Ref. 25.

Although the kernel is never explicitly computed in the POD frequency domain analysis, by choosing a finite set of discrete frequencies for the snapshots, Eq. (24) can be written

$$K = \frac{1}{m}\sum_{i=1}^{m}(j\omega_i I - A)^{-1}BB^*\left(-j\omega_i I - A^*\right)^{-1} \tag{32}$$

By comparing Eqs. (31) and (32), we can see that in the case of general inputs, the POD kernel is, therefore, an approximation to the controllability grammian over a chosen, restricted frequency range. The subspace spanned by the POD basis vectors approximates the controllability subspace.

It is a natural extension to consider a POD analysis that approximates the observability subspace. Furthermore, to obtain a balanced representation of the system, we can then use concepts from a traditional control balanced truncation. Lall et al.[21] used the direct POD method to obtain approximations to the system grammians. For a system of order n, this results in the construction of two $n \times n$ matrices. Clearly, for very large systems, this approach will be computationally infeasible, especially given that the matrices will not be sparse. Here we present an alternative approach that uses the POD method of snapshots to approximate the grammians, so that the large matrices need never be explicitly computed.

By obtaining snapshots of the dual system (10) and (11), and performing the POD method of snapshots analysis described earlier, we can calculate p eigenmodes of the observability kernel function. Let these eigenvectors be contained in the columns of the matrix X, with corresponding eigenvalues on the diagonal entries of the matrix Λ_o. Similarly, let the eigenvectors of the conventional (controllability) kernel K be contained in the columns of the matrix Y, with corresponding eigenvalues on the diagonal entries of the matrix Λ_c. Low-rank approximations to the controllability and observability grammians can then be made as follows:

$$W_c^p = Y\Lambda_c Y^* \tag{33}$$

$$W_o^p = X\Lambda_o X^* \tag{34}$$

where the superscript p denotes a pth-order approximation.

Through use of an efficient eigenvalue solver, the eigenmodes of the product $W_c^p W_o^p$ can then be calculated. In this work, ARPACK[26] was used to determine the eigenvalues. This package requires the user to supply only matrix–vector multiplications; hence, the large matrices W_c^p and W_o^p need never be explicitly formed.

The balancing algorithm can therefore be summarized as follows:

1) Use the method of snapshots to obtain p POD eigenmodes (Y, Λ_c) for the primal system.

2) Use the method of snapshots to obtain p POD eigenmodes (X, Λ_o) for the dual system.

3) Formulate the low-rank approximations $W_c^p = Y\Lambda_c Y^*$ and $W_o^p = X\Lambda_o X^*$. (The $n \times n$ matrices are never explicitly calculated.)

4) Obtain the eigenvectors of the product $W_c^p W_o^p$ to determine the balancing transformation T.

5) Retain only those eigenvectors in the reduced-space basis that correspond to large Hankel singular values.

Multiple Input/Output Case

The concept extends readily to the multi-input/multi-output case; however, it is important to treat the system in the correct manner if the correlation with the grammians is to be maintained.

Consider a system with q inputs

$$\boldsymbol{u} = [u_1 \quad u_2, \ldots, u_q]^T \tag{35}$$

The matrix B in Eq. (1) can be written

$$B = [\boldsymbol{b}_1 \quad \boldsymbol{b}_2, \ldots, \boldsymbol{b}_q] \tag{36}$$

We now inspect the form of the controllability grammian W_c defined by Eq. (7). Because of the nature of the outer product, the grammian of the multiple-input system can be written as a sum of grammian components that correspond to each of the inputs as follows:

$$W_c = \int_0^\infty e^{At}\boldsymbol{b}_1\boldsymbol{b}_1^* e^{A^*t}\,\mathrm{d}t + \int_0^\infty e^{At}\boldsymbol{b}_2\boldsymbol{b}_2^* e^{A^*t}\,\mathrm{d}t + \cdots + \int_0^\infty e^{At}\boldsymbol{b}_q\boldsymbol{b}_q^* e^{A^*t}\,\mathrm{d}t \tag{37}$$

The kernel function (23) should, therefore, be written as

$$K(\theta, \theta') = \langle \boldsymbol{x}^1(\theta, t)\boldsymbol{x}^{1*}(\theta', t) + \boldsymbol{x}^2(\theta, t)\boldsymbol{x}^{2*}(\theta', t) + \cdots + \boldsymbol{x}^q(\theta, t)\boldsymbol{x}^{q*}(\theta', t)\rangle \tag{38}$$

where $\boldsymbol{x}^j$ is the response of the system to forcing in u_j only. As described earlier, the eigenfunctions $\boldsymbol{\Psi}$ of the kernel can be written as linear combinations of snapshots

$$\boldsymbol{\Psi} = \sum_{j=1}^{q}\sum_{i=1}^{m_j}\beta_i^j \boldsymbol{x}_i^j \tag{39}$$

where the number of snapshots m_j can vary for different inputs j. Following the derivation by Sirovich,[15] we obtain an eigenvalue problem for the coefficients β_i^j. We find that the resulting system has an identical form to Eq. (26), with the total number of snapshots now being given by

$$m = \sum_{j=1}^{q} m_j$$

The POD can, therefore, be applied to a multiple-input problem, and the approximation of the controllability grammian will be maintained provided snapshots are obtained for each input in turn. The resulting collection of snapshots, $\boldsymbol{x}_i^j$, $i = 1, \ldots, m_j$, $j = 1, \ldots, q$, is then treated in the same way as for the SISO case. Analogous arguments can be applied to the dual problem so that, for multiple outputs, snapshots must be obtained for each output in turn. It is evident that, if one is concerned with a large number of outputs, then the balanced approach will be prohibitively expensive (and could potentially result in much lower reduction in the number of states). It is, therefore, important to characterize the problem at hand carefully before choosing a reduction methodology.

Results and Discussion

In this section, the performance of the method will be illustrated with two examples. The first is a randomly generated, moderately sized problem for which the exact balanced realization can be computed. The second is a realistic high-order fluid dynamic problem.

Randomly Generated State-Space System

In this example, we analyze a randomly generated SISO system of size $n = 100$. The matrix was chosen to be diagonal with eigenvalues distributed uniformly over the interval $[-1 \quad 0]$. The vectors $\boldsymbol{B}$ and $\boldsymbol{C}$ were also randomly generated. The exact balanced realization of the system was determined by solving the Lyapunov equations. The Hankel singular values were computed, and the first 10 are plotted in Fig. 1. As Fig. 1 shows, the magnitudes of the Hankel singular values decrease very rapidly. This indicates that a balanced truncation of

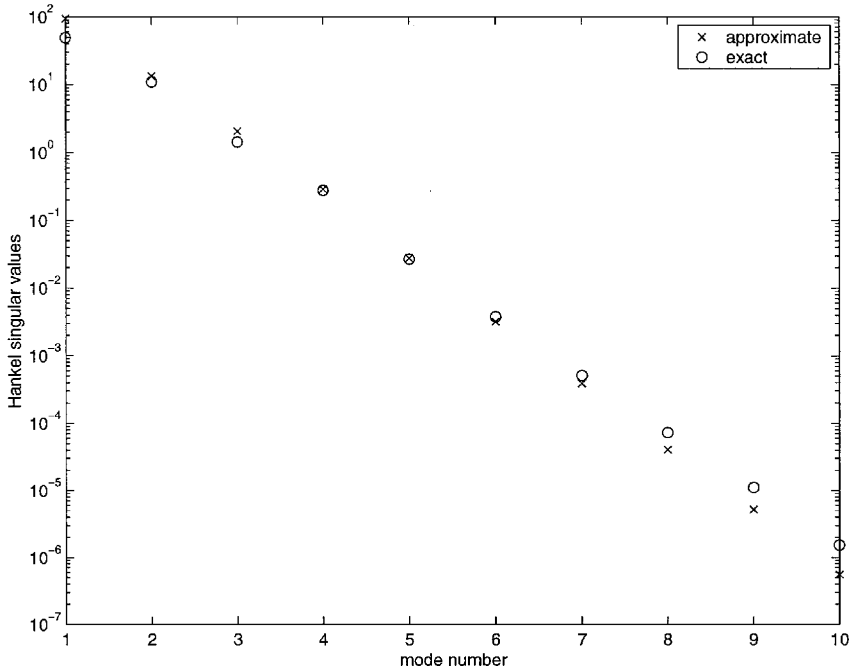

Fig. 1 Hankel singular values (o) and square roots of eigenvalues of the approximate grammian product (x), n = 100, p = 20, and m = 40.

the system could provide a very accurate representation with just a few states.

The approximate balancing method was then applied to the system. POD snapshots were taken from the primal and dual systems at frequency intervals of $\Delta\omega = 0.05$ from $\omega = 0$ to 1. There were 20 POD basis vectors calculated for each system and used to form the approximations to the appropriate grammian. The resulting square roots of the eigenvalues of the grammian product are also plotted in Fig. 1, and we see that the method approximates the dominant Hankel singular values very well.

CFD Model

Results will now be presented for a two-dimensional NACA 0012 airfoil operating in unsteady motion with a steady-state Mach number of 0.755. The airfoil is assumed to move in a single rigid plunging mode (vertical motion with no angular displacement). There is, therefore, a single system input: the instantaneous velocity of the airfoil. (A rigid vertical translation has no effect in a linearized analysis.) The output of interest for this case is the lift force generated on the airfoil. Although this is a simplified case, it is chosen to demonstrate the effectiveness of the methodology with a small number of inputs and outputs. Additional structural degrees of freedom, such as rigid angular displacement and velocity (pitching motion) could be addressed in an analogous way. The flow is assumed to be inviscid, and all motions are assumed to be small. Thus, the governing equations are the linearized Euler equations. The steady-state pressure contours for this problem are shown in Fig. 2. The CFD mesh has 3482 grid points, which corresponds to a total of $n = 13{,}928$ unknowns in the linear state-space system.

POD snapshots were obtained by causing the airfoil to plunge in sinusoidal motion at selected frequencies.[19] Frequencies were selected at 0.1 increments from $\omega = 0.1$ to 2.0. Snapshots were obtained at each frequency by solving the frequency-domain equations (29) and the corresponding frequency-domain equations for the dual system with a preconditioned complex GMRES algorithm. Thus, 40 snapshots were obtained for each problem (2 per frequency). The correlation matrices were calculated and the POD process used to determine the kernel eigenfunctions.

The low-rank approximations to the grammians were formed by taking 15 eigenmodes for each ($p = 15$). ARPACK was then used to calculate the first 10 eigenmodes of the grammian product. As mentioned earlier, it is not necessary to form the grammians or the product explicitly. Instead, for each problem, 15 eigenvectors, each

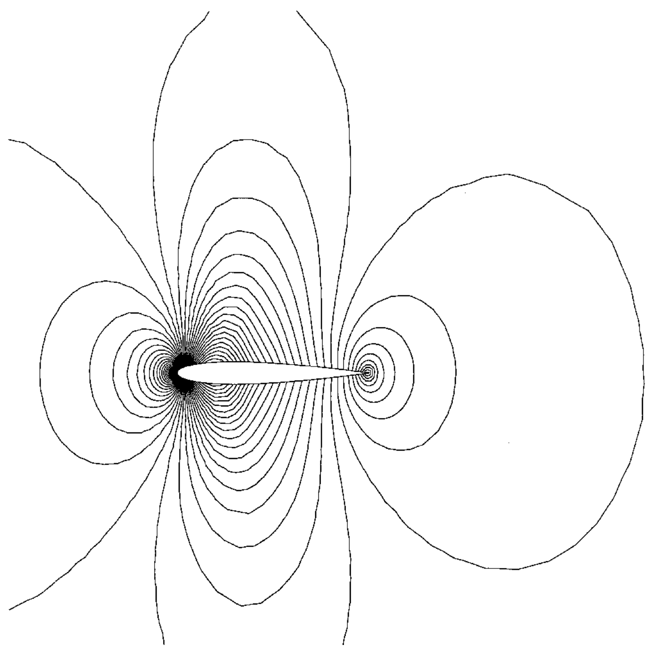

Fig. 2 Steady-state pressure contours for NACA 0012 airfoil: 3482 nodes, M = 0.755, and α = 0.016 deg.

of size n, and 15 eigenvalues were stored, and the matrix multiplications were computed as necessary. Because of the extremely low-rank approximation of the matrix, the eigenvalue solver converged very quickly. The resulting first 10 eigenvalues of the grammian product are plotted with $\times$ in Fig. 3. These eigenvalues approximate the squares of the Hankel singular values of the system, which are independent of the realization. In a balanced truncation, only those states are retained that correspond to large Hankel singular values. From Fig. 3, we see that the magnitudes of eigenvalues drop off very sharply, indicating that the reduced-order model will require only a few states. In fact, the first state contains most of the system "energy."

The accuracy of the reduced-order model obtained from the balanced truncation can be assessed via simulation results. Forced response of the airfoil to a pulse input in plunge is considered, and the

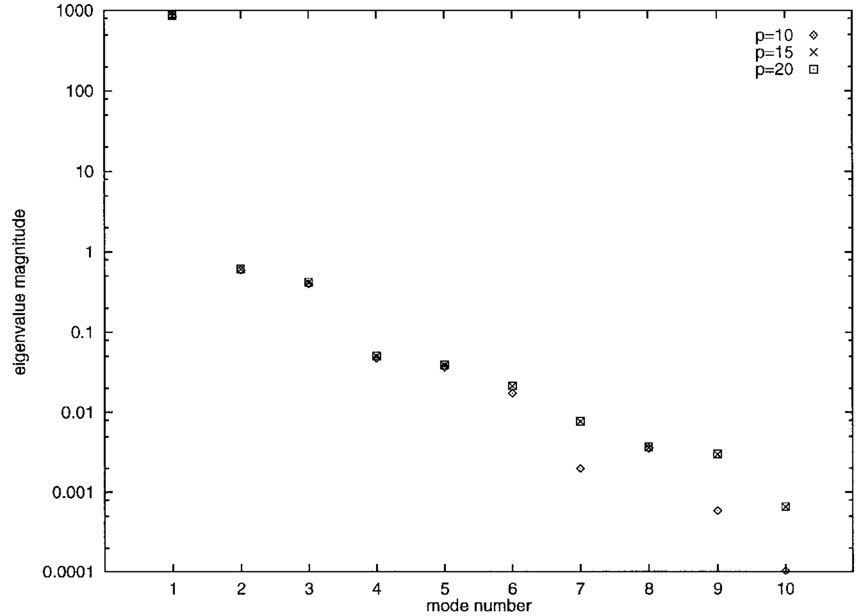

Fig. 3 Eigenvalues of the approximate grammian product (approximate the squares of the Hankel singular values of the system): n = 13,928, m = 40, and p = 10, 15, and 20.

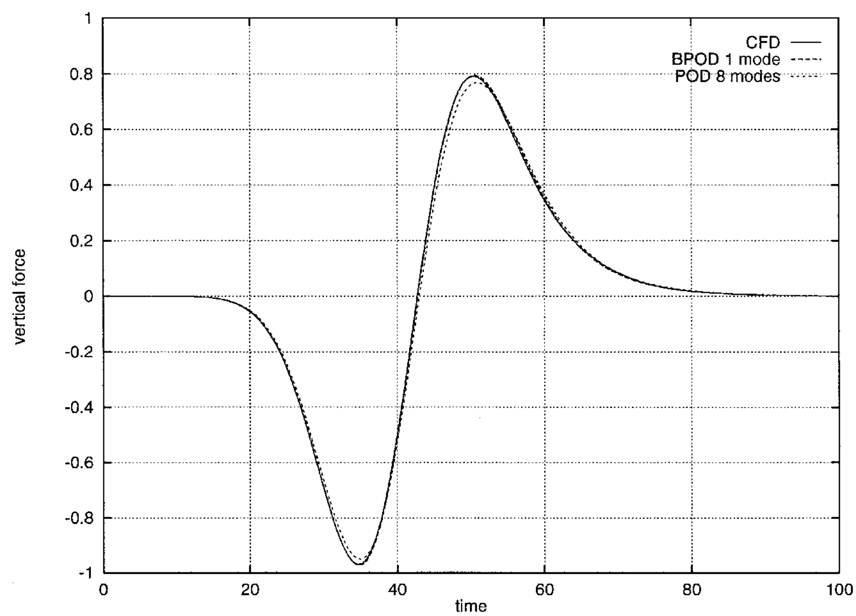

Fig. 4 Response of NACA 0012 airfoil to a pulse input in plunge: M = 0.755, α = 0.016 deg, and g = 0.01; results from CFD model (13,928 states, ——), conventional POD reduced-order model with eight states (- - - -), and balanced reduced-order model with one state (– – –).

results are compared to those obtained both with the high-order CFD code and with a reduced-order model derived via conventional POD. Note that although the basis for the conventional POD model is the same as that used to approximate the controllability grammian, the derivation of the balanced reduced-order model in no way depends on the conventional reduced-order model. The conventional model is presented simply for comparison with the new technique. Static corrections were also included in the reduced-order models to aid in capturing high-frequency dynamics.[19]

The plunge displacement h of the airfoil was prescribed to be

$$h(t) = \exp\left[-g(t-t_0)^2\right] \tag{40}$$

where g is a parameter that determines how sharp the pulse is and, thus, the value of the maximum significant frequency present. Figure 4 shows the results for $g = 0.01$, which corresponds to $\omega_{\max} = 0.48$ based on a 1% level. The solid line represents the force generated on the airfoil as a function of time as calculated with the high-order CFD model. The two dashed lines are the results obtained using a conventional POD model with eight states and the new balanced model with one state. As Fig. 4 shows, with just a single degree of freedom, the balanced reduced-order model captures the response of the CFD model almost exactly.

The same test was performed for a higher frequency pulse with $g = 0.1$. In this case, the highest significant frequency present at a 1% level is $\omega_{\max} = 1.34$. The frequency content in this pulse input, therefore, spans most of the range sampled by the snapshots. Figure 5 shows the results for the CFD model, along with the reduced-order models with eight (conventional) and three (balanced) degrees of freedom. Although the balanced model with three states captures the response extremely accurately, even with eight states the conventional model shows a significant error. This was the highest-order model that could be obtained using conventional POD because including additional basis vectors caused the reduced-order model to become unstable. To capture the response more accurately, it would be necessary to include more snapshots in the conventional

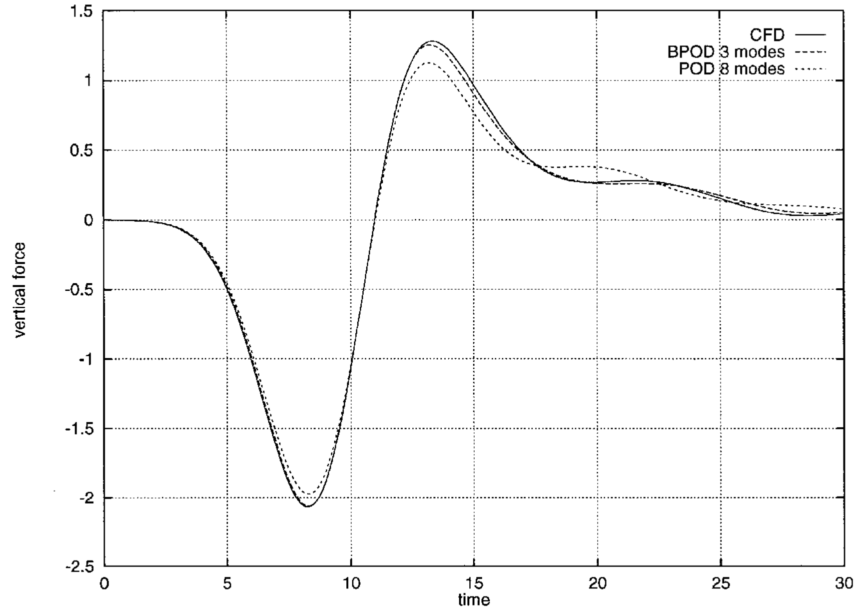

Fig. 5 Response of NACA 0012 airfoil to a pulse input in plunge: M = 0.755, α = 0.016 deg, and g = 0.1. Results from CFD model (13,928 states, ——), conventional POD reduced-order model (eight states, - - - -), and balanced reduced-order model (three states, – – –).

POD analysis. This relatively common occurrence with conventional POD methods seems to be addressed with the balancing approach.

Discussion

The computation of reduced-order models via the balancing method is slightly more than twice as expensive as conventional POD for the SISO system. The greatest computational cost occurs in the system solves that are required to obtain the flow snapshots. Because we must solve the dual system, twice as many snapshots are required for the balanced POD approach when there are equal numbers of inputs and outputs. Additionally, there is the small cost associated with calculating the eigenmodes of the grammian product. In many applications, however, one is less concerned with the cost of obtaining the reduced-order model and more so with the resulting size and quality of the model. As the preceding results show, by incorporating both input and output information, very accurate models can be obtained that have a minimal number of states.

For the example presented here, one might argue that the additional reduction to three states in the balanced reduced-order model is insignificant compared to the reduction from 13,928 to 8 states using conventional POD. In terms of simulation, this is true; however, in some applications, such as design of active control strategies, the smallest possible model is desired, as noted in Ref. 20. The additional reduction would also be significant for larger-scale problems such as turbomachinery applications with multiple blade passages. More important, reduction based on conventional POD is heavily dependent on the arbitrary scaling of this system. Using the balanced approach removes this dependency and reduces the potential for an inaccurate characterization of the system. Moreover, the models resulting from the balanced methodology do not appear to suffer from the same level of instability as those obtained with conventional POD. The method outlined here is far more efficient than attempts to calculate system grammians and accurately provides the required level of information. Also, because the snapshots can be obtained efficiently in the frequency domain, this method is appropriate for problems with spatial symmetry, such as turbomachinery flows.

When performing model reduction for any system, it is important to keep in mind the limitations of the reduced-order model. Strictly speaking, the reduced-order model is valid only over the range of conditions used to generate the reduced-order basis. If problem parameters are changed or active control is used to alter the dynamics, then it is possible (if not probable) that the dominant modes of the system will change. If a reduced-order model is to be used in such a context, it is vital that the controlled dynamics and parameter variations are included in the basis derivation process. This may be accomplished somewhat more easily using the frequency-domain approach because one must simply ensure that the relevant frequency range is sampled adequately.

Throughout the algorithm, there are several arbitrary decisions to be made. First, the snapshots must be selected. The range over which the sampling is performed is determined by assessing the important frequency range in the problems at hand. To determine the specific snapshot locations within this range, one uses a combination of experience and intuition. Often, the required density of snapshots will be determined a posteriori from the performance of the reduced-order model. If the desired dynamics cannot be accurately captured, more snapshots must be included in the POD process and the basis vectors recalculated. It is, therefore, important to validate the models against known results (in this case against the CFD model). Second, p, the number of POD eigenmodes to be used in the low-rank approximation of the grammians, must be chosen. Again, this will depend on the frequency range of interest, as well as on the number of modes to be retained in the reduced-order model. A fairly low number (15) was chosen for the results presented here; however, the performance of the models was found to be fairly robust with respect to this parameter. In Fig. 3, the eigenvalues of the grammian product are also plotted for $p = 10$ and 20. When the number of POD vectors was increased to 20, very little variation was seen in the eigenvalues. If only 10 POD vectors were used, the first five eigenvalues were virtually unchanged, whereas the next five showed some movement. This result is to be expected: to resolve q eigenvalues of the grammian product accurately, we should choose $p > q$ in the approximation of the matrices.

The method is flexible in that it can be applied as described to any linearized system. The approach also extends to nonlinear systems using concepts similar to those discussed by Lall et al.[21] It is relatively straightforward to obtain an approximation to the controllability subspace. For example, consider the nonlinear system

$$\dot{\boldsymbol{x}} = f[\boldsymbol{x}(\theta, t), \boldsymbol{u}(t)] \tag{41}$$

$$\boldsymbol{y} = g[\boldsymbol{x}(\theta, t)] \tag{42}$$

The POD eigenfunctions would be calculated using snapshots from simulation of the nonlinear system (41) and (42). The difficulty arises with the approximation of the observability subspace. The concept of a dual system does not exist in a nonlinear setting. Two possibilities present themselves. The first is to linearize the nonlinear system (41) and (42) and formulate the dual linearized system (the adjoint). The snapshots would then be obtained from a combination of nonlinear (primal) and linearized (dual) systems. The second approach is to follow the method outlined in Ref. 21 that defines an empirical observability grammian based on system outputs for various initial conditions. For the high-order systems encountered in CFD applications, this second approach, although more accurate, would be computationally very expensive. Work is underway to develop a better approach for handling large nonlinear systems.

Once the grammians have been suitably approximated, the balancing method can then be used to calculate a linear transformation of the nonlinear state to obtain the nonlinear reduced-order model

$$\dot{\boldsymbol{x}}_r = T^{-1} f[T\boldsymbol{x}_r(\theta, t), \boldsymbol{u}(t)] \tag{43}$$

$$\boldsymbol{y}_r = g[T\boldsymbol{x}_r(\theta, t)] \tag{44}$$

Conclusions

A new method for computing an approximate balanced truncation of a linear state-space system has been presented. By the use of the method of snapshots to perform a POD analysis of the primal and dual systems, low-rank, reduced-range approximations to the controllability and observability grammians are obtained very efficiently. This POD analysis can be performed in either the time or frequency domain. By the incorporation of information pertaining to both inputs and outputs, the resulting reduced-order models capture the desired system dynamics with a very low number of states. The required size of the models is significantly lower than for those developed using a conventional POD approach, and they demonstrate improved robustness over a larger range of inputs. Results have been presented for a very high-order system, and the method has been shown to work extremely effectively. The concept is applicable to general linearized systems and, with some modifications, can be extended to nonlinear systems.

Acknowledgment

Support for this work has been provided by the Singapore MIT Alliance.

References

[1]Moore, B., "Principal Component Analysis in Linear Systems: Controllability, Observability, and Model Reduction," *IEEE Transactions on Automatic Control*, Vol. AC-26, No. 1, 1981, pp. 17–31.

[2]Baker, M., Mingori, D., and Goggins, P., "Approximate Subspace Iteration for Constructing Internally Balanced Reduced-Order Models of Unsteady Aerodynamic Systems," AIAA Paper 96-1441, 1996.

[3]Scottedward Hodel, A., "Least Squares Approximate Solution of the Lyapunov Equation," *Proceedings of the 30th IEEE Conference on Decision and Control*, IEEE Publications, Piscataway, NJ, 1991.

[4]Jaimoukha, I., and Kasenally, E., "Krylov Subspace Methods for Solving Large Lyapunov Equations," *SIAM Journal of Numerical Analysis*, Vol. 31, No. 1, 1994, pp. 227–251.

[5]Gudmundsson, T., and Laub, A., "Approximate Solution of Large Sparse Lyapunov Equations," *IEEE Transactions on Automatic Control*, Vol. 39, No. 5, 1994, pp. 1110–1114.

[6]Gallivan, K., Grimme, E., and Van Dooren, P., "Padé Approximation of Large-Scale Dynamic Systems with Lanczos Methods," *Proceedings of the 33rd IEEE Conference on Decision and Control*, IEEE Publications, Piscataway, NJ, 1994, pp. 443–448.

[7]Hall, K., Florea, R., and Lanzkron, P., "A Reduced Order Model of Unsteady Flows in Turbomachinery," American Society of Mechanical Engineers, ASME Paper 94-GT-291, 1994.

[8]Dowell, E., Hall, K., Thomas, J., Florea, R., Epureanu, B., and Heeg, J., "Reduced Order Models in Unsteady Aerodynamics," *Engineering Mechanics*, Vol. 6, No. 4, 1999, pp. 229–252.

[9]Romanowski, M., and Dowell, E., "Using Eigenmodes to Form an Efficient Euler Based Unsteady Aerodynamics Analysis," *Proceedings of the Special Symposium on Aeroelasticity and Fluid/Structure Interaction Problems, ASME International Mechanical Engineering Congress and Exposition*, AD-Vol. 44, American Society of Mechanical Engineers, Fairfield, NJ, 1994, pp. 147–160.

[10]Mahajan, A., Dowell, E., and Bliss, D., "Eigenvalue Calculation Procedure for an Euler/Navier–Stokes Solver with Application to Flows over Airfoils," *Journal of Computational Physics*, Vol. 97, No. 2, 1991, pp. 398–413.

[11]Hall, K., "Eigenanalysis of Unsteady Flows About Airfoils, Cascades, and Wings," *AIAA Journal*, Vol. 32, No. 12, 1994, pp. 2426–2432.

[12]Bisplinghoff, R., and Ashley, H., *Principles of Aeroelasticity*, Dover, New York, 1975.

[13]Loéve, M., *Probability Theory*, Van Nostrand, New York, 1955.

[14]Lumley, J., "The Structures of Inhomogeneous Turbulent Flow," *Atmospheric Turbulence and Radio Wave Propagation*, edited by A. M. Yaglom and V. I. Tatarski, Nauka, Moscow, 1967, pp. 166–178.

[15]Sirovich, L., "Turbulence and the Dynamics of Coherent Structures. Part 1: Coherent Structures," *Quarterly of Applied Mathematics*, Vol. 45, No. 3, 1987, pp. 561–571.

[16]Berkooz, G., Holmes, P., and Lumley, J., "The Proper Orthogonal Decomposition in the Analysis of Turbulent Flows," *Annual Review of Fluid Mechanics*, Vol. 25, 1993, pp. 539–575.

[17]Kim, T., "Frequency-Domain Karhunen–Loeve Method and Its Application to Linear Dynamic Systems," *AIAA Journal*, Vol. 36, No. 11, 1998, pp. 2117–2123.

[18]Hall, K. C., Thomas, J. P., and Dowell, E. H., "Proper Orthogonal Decomposition Technique for Transonic Unsteady Aerodynamic Flows," *AIAA Journal*, Vol. 38, No. 10, 2000, pp. 1853–1862.

[19]Willcox, K., *Reduced-Order Aerodynamic Models for Aeroelastic Control of Turbomachines*, Ph.D. Dissertation, Dept. of Aeronautics and Astronautics, Massachusetts Inst. of Technology, Cambridge, MA, Feb. 2000.

[20]Dowell, E., and Hall, K., "Modeling of Fluid–Structure Interaction," *Annual Review of Fluid Mechanics*, Vol. 33, 2001, pp. 445–90.

[21]Lall, S., Marsden, J., and Glavaski, S., "Empirical Model Reduction of Controlled Nonlinear Systems," *Proceedings of the IFAC World Congress*, Vol. F, International Federation of Automatic Control, Beijing, 1999, pp. 473–478.

[22]Enns, D., "Model Reduction with Balanced Realizations: An Error Bound and a Frequency Weighted Generalization," *Proceedings of the IEEE Conference on Decision and Control*, IEEE Publications, Piscataway, NJ, 1984, pp. 127–132.

[23]Kwakernaak, H., and Sivan, R., *Linear Optimal Control Systems*, Wiley-Interscience, New York, 1972.

[24]Hotelling, H., "Analysis of a Complex of Statistical Variables with Principal Components," *Journal of Educational Psychology*, Vol. 24, No. 7, 1933, pp. 417–441 and 498–520.

[25]Zhou, K., Salomon, G., and Wu, E., "Balanced Realization and Model Reduction for Unstable Systems," *International Journal of Robust and Nonlinear Control*, Vol. 9, No. 3, 1999, pp. 183–198.

[26]Lehoucq, R., Sorensen, D., and Yang, C., *ARPACK Users Guide: Solution of Large Scale Eigenvalue Problems by Implicitly Restarted Arnoldi Methods*, Rice Univ. Press, Houston, TX, 1997.

C. Pierre
Associate Editor