Proceedings of IMECE2007
2007 ASME International Mechanical Engineering Congress and Exposition
November 11-15, 2007, Seattle, Washington, USA

IMECE2007-42031

A REVIEW OF PROPER MODELING TECHNIQUES

Tulga Ersal
The University of Michigan
Department of Mechanical Engineering
Ann Arbor, Michigan, USA
tersal@umich.edu

Hosam K. Fathy
The University of Michigan
Department of Mechanical Engineering
Ann Arbor, Michigan, USA
hfathy@umich.edu

Loucas S. Louca
University of Cyprus
Department of Mechanical and Manufacturing Engineering
Nicosia, Cyprus
lslouca@ucy.ac.cy

D. Geoff Rideout
Memorial University of Newfoundland
Department of Engineering and Applied Science
St. John's, NL, Canada
grideout@engr.mun.ca

Jeffrey L. Stein
The University of Michigan
Department of Mechanical Engineering
Ann Arbor, Michigan, USA
stein@umich.edu

ABSTRACT

A dynamic system model is *proper* for a particular application if it achieves the accuracy required by the application with minimal complexity. Because model complexity often – but not always – correlates inversely with simulation speed, a proper model is often alternatively defined as one balancing accuracy and speed. Such balancing is crucial for applications requiring both model accuracy and speed, such as system optimization and hardware-in-the-loop simulation. Furthermore, the simplicity of proper models conduces to control system analysis and design, particularly given the ease with which lower-order controllers can be implemented compared to higher-order ones. The literature presents many algorithms for *deducing* proper models from simpler ones or *reducing* complex models until they become proper. This paper presents a broad survey of the proper modeling literature. To simplify the presentation, the algorithms are classified into *frequency-, projection-, optimization-, and energy-based*, based on the metrics they use for obtaining proper models. The basic mechanics, properties, advantages and limitations of the methods are discussed, along with the relationships between different techniques, with the intention of helping the modeler to identify the most suitable proper modeling method for their application.

Keywords: proper modeling, model simplification, model reduction, model deduction, model partitioning

I. INTRODUCTION

Mathematical simulation models are indispensable to engineering system analysis, design, and control development, particularly during preliminary design stages. They enable virtual experiments when physical experimentation is either too expensive, time consuming, infeasible or even impossible to conduct.

The viability of a model for system development purposes rests on its *accuracy* and *simplicity*. Model accuracy is critical for understanding, optimizing, and controlling the dynamics of a given system effectively. Model simplicity, on the other hand, is essential for tractability in system identification and optimization. Simpler models are also easier to inspect for physical insights than more complex ones, and can lead to lower-order controllers that are easier to implement. Finally, simpler models are often – but not always – faster to simulate, which can be crucial for applications such as hardware-in-the-loop simulation or embedded model-reference control. In summary, model accuracy and simplicity are often both crucial for effective system identification, analysis, optimization, and control.

Seeking model accuracy and simplicity simultaneously, however, typically engenders a tradeoff: increasing the accuracy of a system model often necessitates increasing the complexity of the model to a level more commensurate with the complexity of the real system. In other words, the requirements of model accuracy and simplicity often compete, and must hence be traded off. This competition typically grows as engineering systems become larger, more complex, and more integrated: a trend in many engineering disciplines. There is a growing need for system models that mitigate this competition and balance accuracy and simplicity by only capturing the dynamics necessary for their respective applications.

The literature, in recognition of this need, deems a dynamic system model *proper* [1] if it provides the accuracy required for a given application with minimal complexity. By balancing accuracy and simplicity, proper models prove useful in optimization [2], real-time simulation [3], control design [4], and other applications requiring both model accuracy and simplicity, such as sensitivity analysis, Monte Carlo simulation, or system identification.

Obtaining a proper model, however, is not an easy task. It is not always obvious which phenomena are important for a specific application, i.e., what to include in a model and what to neglect. Hence, dynamic system models are seldom proper at the outset. To remedy this problem, the literature proposes many techniques for obtaining proper models.

This paper provides a broad review of the proper modeling literature. Some of these techniques begin with simple models and increment their complexity until they meet their respective accuracy requirements: a process known as *model deduction*. Most techniques, however, begin with excessively complex models and then *reduce* them until they become proper.

The ultimate goal of both model deduction and reduction techniques is the same, regardless of how it is achieved: given a dynamic system model, balance its accuracy and complexity by massaging it to include only the most salient dynamics of the given system. This implies that every proper modeling algorithm must have at its core a *metric* for quantifying the relative importance of modeling the different dynamics of a given system. Based on the metrics they use for proper modeling, this paper classifies the proper modeling techniques presented in the literature into *frequency*-, *projection*-, *optimization*-, and *energy-based*.

This classification is neither a universally adopted convention nor is it strict. In fact, the section will show that a given proper modeling technique can often conceptually belong to more than one of these categories. However, the authors have found this classification intuitively appealing and convenient for presentation and pedagogy, and hence adopt it herein.

This review focuses mostly on model reduction and deduction techniques applicable to finite-dimensional, lumped-parameter, continuous-time models of deterministic dynamic systems, with some brief references to infinite-dimensional and stochastic systems. The review also emphasizes that there does not exist a "universal" proper modeling algorithm applicable to all proper modeling problems in all domains. Rather, different proper modeling algorithms are ideally suited to different problem domains, and one must therefore choose between proper modeling algorithms judiciously based on the given problem space. The paper concludes with a brief examination of ongoing challenges in proper modeling, and how further research can address them. Similar reviews exist in the literature [5-15], but this work is unique in its use of proper modeling as a broad contextual framework within which different algorithms are compared and contrasted.

II. FREQUENCY-BASED TECHNIQUES

The fundamental metric used by *frequency-based* proper modeling techniques for assessing the importance of a given system's various dynamics is *characteristic speed*. In particular, given a dynamic system model, these techniques partition it into submodels with comparatively "fast" and "slow" dynamics whose relative importance depends on the given application.

Consider, for instance, the dynamics of a hydraulic car braking system. A full model of such a system may simultaneously capture the dynamics of the car's motion and the dynamics of hydraulic pressure wave propagation. The latter dynamics are typically orders of magnitude faster than the former. A model capturing both sets of dynamics is therefore likely to exhibit significant *numerical stiffness*, defined as a disparity between its different characteristic speeds. Such numerical stiffness may cause the model to be computationally intractable, thereby necessitating a more "proper" technique for modeling this braking system. Such a proper modeling technique may neglect fluid compressibility when the goal is to examine vehicle braking, and conversely neglect vehicle motion when the goal is to examine pressure wave propagation.

This paper refers to all techniques that use characteristic speed as a metric for proper modeling as *frequency-based* techniques. The term "frequency-based", in this context, underscores the congruence between characteristic speeds and eigenvalues in the case of linear systems. Indeed, as the review below shows, frequency-based proper modeling techniques are most often used for linear systems, even though many of them can be generalized to nonlinear systems. This review focuses on eight established classes of frequency-based proper modeling techniques from the literature, namely, *aggregation*, *singular perturbation*, *the model order deduction algorithm (MODA)*, *modal analysis*, *component mode synthesis (CMS)*, *polynomial methods*, *oblique projection*, and *optimal Hankel norm approximation*. It briefly details the fundamental principles behind each technique or class of techniques, in addition to their conceptual similarities and differences.

Aggregation

One of the basic ideas in the model reduction literature is to ignore the small time constants in a system, and keep the large ones, which are assumed to dominate the response. Thus, the earlier model reduction methods were based on retaining the dominant eigenvalues of the system in the reduced model [16-22]. While developing his optimal projection method Mitra showed that Davison's method [16] is a special case of optimal projection [23, 24]. Aoki later developed the more general method of aggregation [25], and it has been shown that Mitra's optimal projection method is a special case of aggregation [26-28].

The basic idea behind the aggregation method can be summarized as follows. Consider the approximation of the *n*-dimensional original system

$$\begin{aligned}\dot{x} &= Ax + Bu \\ y &= Cx + Du\end{aligned} \tag{1}$$

with the *r*-dimensional reduced model

$$\begin{aligned}\dot{x}_r &= A_r x_r + B_r u \\ y &= C_r x_r + Du\end{aligned} \tag{2}$$

Suppose the reduced state vector x_r is related to the original state vector *x* through

$$x_r = Kx \tag{3}$$

where *K* is the $r \times n$ aggregation matrix. It follows that

$$\begin{aligned}A_r K &= KA \\ B_r &= KB \\ C_r K &\approx C\end{aligned} \tag{4}$$

A least-squares solution can be obtained by using the pseudoinverse as

$$\begin{aligned}A_r &= KAK^{\dagger} \\ B_r &= KB \\ C_r &= CK^{\dagger}\end{aligned} \tag{5}$$

It has been shown that a nontrivial aggregation law exists if and only if the A_r retains *r* of the eigenvalues of *A* [28]. Furthermore, *K* can be obtained by

$$K = T\begin{bmatrix} I_r & 0 \end{bmatrix} V^{-1} \tag{6}$$

where T is any nonsingular matrix, and *V* is the modal matrix of *A*.

This basic idea of aggregation has been extended by many researchers. For example, Aoki proposed two ways of relaxing the perfect-aggregation condition [29]. Hickin proposed a method called nonminimal partial realization that combines the ideas of aggregation and moment matching [30]. Siret *et al.* developed a method to chose the arbitrary matrix *T* in Eq. (6) in an optimal way to maximize a performance criterion [27]. It must be noted, however, that even though some of the eigenvalues of *A* are retained, the aggregation method is not *realization-preserving*, because the reduced model uses a different set of state variables than the original one; specifically, a combination of the original state variables. Hence, the intuitive appeal of the original model may not be preserved in the reduced model.

Singular Perturbation Method

As the difference between the large and small time constants in a system increases, or, in other words, as the underlying characteristic speeds become significantly disparate, the system is said to possess multiple time scales and becomes numerically stiff. *Singular perturbation* is a reduction technique particularly suited to this type of models.

Unlike aggregation, singular perturbation is realization-preserving in the sense that it does not necessarily require a coordinate transformation as part of model reduction. This is quite attractive, because it implies that the physical meaning associated with each state in the original model can be preserved in the reduced model.

In its simplest rendition, singular perturbation implicitly assumes *a priori* knowledge of which state variables of a given model correspond to the fast dynamics and which correspond to the slow. Neglecting the influence of the "fast" dynamics on the "slow" states *partitions* the original stiff model into two submodels. The first *driving* submodel captures the slow dynamics and *residualizes* the fast states, while the second *driven* submodel captures the fast dynamics and treats the slow states as input variables. This furnishes a *decoupled* system model that not only mitigates the original model's numerical stiffness but also approaches the original model in accuracy as this stiffness grows.

The origins of the singular perturbation method go back to Prandtl's work on boundary layers in fluid dynamics [31]. Later contributions by Tikhonov [32], Levinson [33], Vasileva [34], Wasow [35] and Kokotovic [36-39] established singular perturbation as a model reduction tool. In its simplest rendition, the singular perturbation method assumes that the dynamics of a system are expressed in state space form, where some derivatives have a small positive number ε as a coefficient, i.e.,

$$\dot{x}_1 = f_1(x_1, x_2, u), \quad x_1 \in \mathbb{R}^n \tag{7}$$

$$\varepsilon \dot{x}_2 = f_2(x_1, x_2, u), \quad x_2 \in \mathbb{R}^m \tag{8}$$

The coefficient ε represents the *disparity* between the characteristic speeds of the fast and slow dynamics. As this coefficient approaches zero, Eq. (8) becomes

$$0 = f_2(\bar{x}_1, \bar{x}_2, \bar{u}) \tag{9}$$

where bars are used to distinguish between this limiting case and the case where ε truly equals zero. Now assume that Eq. (9) can be solved to obtain a distinct real expression for $\bar{x}_2$ in terms of $\bar{x}_1$, i.e.,

$$\bar{x}_2 = \phi(\bar{x}_1, \bar{u}) \tag{10}$$

Substituting this solution into Eq. (7) effectively furnishes a slow submodel that *residualizes* the fast states, i.e.,

$$\bar{x}_1 = f_1(\bar{x}_1, \phi(\bar{x}_1, \bar{u}), \bar{u}) = \bar{f}(\bar{x}_1, \bar{u}) \qquad (11)$$

The reduced model for the fast dynamics can be obtained by introducing a fast time scale τ and fast variables $\tilde{x}_1(\tau)$ and $\tilde{x}_2(\tau)$ defined as follows:

$$\tau = \frac{t}{\varepsilon}, \quad x_j(t) = \bar{x}_j(t) + \tilde{x}_j(\tau), \quad j = 1, 2 \qquad (12)$$

Combining Eq. (7), (8), and (12), and letting $\varepsilon \to 0$, the fast-dynamics model is obtained as

$$\frac{d\tilde{x}_2}{d\tau} = f_2(x_1(t), \bar{x}_2(t) + \tilde{x}_2(\tau), u) \qquad (13)$$

This model uses the slow states as inputs, and is hence *driven* by them.

Equations (7-13) highlight the simplicity with which the singular perturbation method can be applied to a given system. In addition to this simplicity and the method's intuitive appeal, the singular perturbation method furnishes reduced models with attractive mathematical properties in some special cases. In particular, let the original and reduced models be G and G_r, respectively. Furthermore, assume that the full model G is expressed in the time domain using a *balanced realization* (see Section III), then reduced to G_r using the singular perturbation method. Then, the singular perturbation method is equivalent to *balanced residualization*, a projection-based proper modeling technique. Furthermore, the maximum error introduced by singular perturbation, quantified in terms of the $\mathcal{H}_\infty$ norm of the difference $G - G_r$, satisfies:

$$\left\| G - G_r \right\|_\infty \le 2(\sigma_{n+1} + ... + \sigma_{n+m}) \qquad (14)$$

where σ_i, $i = n+1, ..., n+m$ are the Hankel singular values of G corresponding to the fast dynamics [40]. In other words, the $\mathcal{H}_\infty$ norm of the modeling error introduced by singular perturbation cannot exceed twice the sum of the Hankel singular values corresponding to the fast states. Furthermore, this modeling error decreases with the parameter ε, and becomes zero in the limit as ε approaches zero.

Model Order Deduction Algorithm

Like singular perturbation, the *model order deduction algorithm (MODA)* is a realization-preserving technique that deems a model "proper" if it captures only the most relevant characteristic speeds of a given system for a given application. Unlike singular perturbation, however, MODA is a *deduction* algorithm that starts with simple models and increments their complexity until they become proper. Furthermore, MODA does not assume *a priori* knowledge of which states in a system are "fast" and which are "slow". Instead, it explicitly searches for this knowledge as part of its pursuit of proper models.

In its simplest rendition [1], MODA deems a linear system model proper for a given application if the model's *rank* is minimal and its *spectral radius* exceeds a *frequency range of interest (FROI)* desired for the application. The rank of a model, in this context, is the number of components in the model not included in the initial baseline model from which the deduction process proceeds. For instance, a finite-element model of a shaft that uses 30 finite elements has a rank of 23 compared to a baseline finite element model of the same shaft that uses only 7 finite elements. Furthermore, the spectral radius of a linear system is defined as the radius of a closed ball containing all its poles, or equivalently, as the Euclidian norm of its largest poles.

MODA begins with a baseline model and proceeds to increment its rank in a manner that produces the smallest increase in its spectral radius, repeating this process until the spectral radius exceeds the desired FROI [1]. Using this approach, MODA furnishes not only a proper model, but also an understanding of which subsystem dynamics need to be captured accurately to furnish a proper system model. For instance, given a system containing more than one flexible shaft, MODA can determine the number of finite elements needed to model each shaft so that the overall system model is proper. This makes MODA particularly attractive for the automated lumped-parameter modeling of continuous dynamic systems [1].

The literature describes several extensions that enhance the capabilities of MODA. In particular, Ferris *et al.* extend MODA to not only satisfy a given spectral-radius requirement, but also capture system eigenvalues within that spectral radius with a desired level of accuracy [41]. Furthermore, Walker *et al.* modify this algorithm to furnish models that accurately capture the eigenvalues of only the observable and controllable modes of a given system within the desired FROI [42]. Wilson and Taylor modify MODA to seek an accurate representation of a system's frequency response within the desired FROI as opposed to just its eigenvalues [43]. Finally, Taylor and Wilson extend MODA to enable the proper modeling of nonlinear systems over a desired range of input excitation frequencies [44].

MODA is not the only algorithm that adopts the deduction approach to proper modeling. Pirvu *et al.*, for example, propose a bond-graph-model adaptation algorithm that searches for all possible extensions of a given baseline bond-graph model that would result in a desired higher-order transfer function [45]. The baseline model can be extended by adding new interconnections, i.e., 1- and 0-junctions in bond graph terms, or energetic components, i.e., generalized inductors, capacitances or resistors. The transfer-function-matching objective, however, limits this method to linear systems.

Another example of the deduction approach is the bond-graph synthesis using genetic algorithms [46, 47]. Similar to Pirvu's method, this method lets a bond graph evolve from a baseline model. However, the freedom in choosing the fitness function gives this method more flexibility, allowing it to be used not only as a proper modeling tool, but also a conceptual system synthesis tool.

Modal Analysis

In its simplest rendition, *modal analysis* focuses on linear, time-invariant, vector-second-order dynamic systems satisfying the principle of separation of variables (e.g., through proportional damping). Such systems may be finite- or infinite-dimensional. In the latter case, one often approximates the given system's continuous dynamics using a finite-dimensional, lumped-parameter model obtained through a discretization technique (such as finite differences or finite elements). The resulting finite-dimensional model of this vector-second-order system, subject to the assumption of negligible damping, can be expressed as [48, 49]

$$M\ddot{x} + Kx = 0 \tag{15}$$

where *M* and *K* are the effective structural inertia and stiffness matrices, respectively. The modes of such a system can be found by solving the generalized eigenvalue problem

$$Kv = \omega^2 Mv \tag{16}$$

where the natural frequencies are given by the various solutions for ω and the modes shapes are given by the corresponding solutions for *v*. These mode shapes collectively form a basis spanning the complete state space corresponding to Eq. (15). Therefore, the dynamics represented by Eq. (15) can be projected onto the eigenspace given by these mode shapes without loss of information. Such a projection can also be performed on the standard state-space representation of the full model (as opposed to the vector-second-order representation), leading to a new state-space model with a diagonal *A* matrix (with complex entries), as shown below:

$$\dot{\xi} = A\xi + Bu, \;\; y = C\xi$$
$$A = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix}, \; B = \begin{bmatrix} b_1^T \\ b_2^T \\ \dots \\ b_n^T \end{bmatrix}, \; C = \begin{bmatrix} c_1 & c_2 & \dots & c_n \end{bmatrix} \tag{17}$$

Given this new *modal* representation, modal analysis builds on the congruence between the eigenvalues corresponding to a given mode and the characteristic speed of the mode to achieve model reduction. In particular, by eliminating the modes corresponding to the faster eigenvalues from Eq. (17), one can balance the fidelity and complexity of a given model, thereby rendering it proper [48, 49]. Modal analysis is therefore a frequency-based model reduction technique that does not assume *a priori* knowledge of which dynamics of a given system are "fast" and which are "slow". Like singular perturbation, it has the very attractive property of a guaranteed error bound. In particular, the $\mathcal{H}_\infty$ norm of the difference between the original model, *G*, and reduced model, G_r, is bounded by

$$\left\| G - G_r \right\|_\infty \le \sum_{i=k+1}^{n} \bar{\sigma}\left(c_i b_i^T \right) \Big/ \mathrm{Re}\left[\lambda_i \right] \tag{18}$$

where λ_i is the i^{th} eigenvalue, and $\bar{\sigma}$ is the largest singular value of the residues $c_i b_i^T$ [4]. Unlike singular perturbation and MODA, however, modal analysis is not realization-preserving. It expresses the reduced model in terms of modal – rather than physical – coordinates. Consequently, physical insights associated with the original coordinate choice may be lost. Modal analysis shares this property with all *projection-based* proper modeling techniques, and is hence both a frequency-based and projection-based model reduction technique.

The simple rendition of modal analysis presented above only applies to linear finite-dimensional systems. There are several important extensions of this technique, however, that make it applicable to a broader range of problems. First, modal analysis can be applied directly to the partial differential equations governing the dynamics of an infinite-dimensional system: a process that can furnish proper lumped-parameter models of such systems directly. Furthermore, the literature presents many extensions of modal analysis to both linear and nonlinear deterministic and stochastic systems that do not satisfy the assumptions of the above discussion [50-52]. Finally, the literature describes a special extension of modal analysis to modular systems known as *component mode synthesis*. This extension is discussed in further detail below.

Component Mode Synthesis

Component mode synthesis is an extension of modal analysis that is particularly applicable to large, modular systems. It proceeds in two simple steps. First, it uses modal analysis to obtain a proper model of each module in the system separately. Then it assembles these proper module models into a system-level proper model. This two-step approach can be significantly less expensive from a computational standpoint than the direct application of modal analysis to the entire system model, because solving many small eigenvalue problems can be significantly more tractable than solving one large eigenvalue problem. Because of its computational attractiveness, component mode synthesis is widely used in the literature [53-58], particularly in the context of applications involving large modular systems, such as automotive vibration applications [59-61].

Polynomial Approximation Methods

All five proper modeling techniques presented hitherto deem a model proper if it captures the dynamics of a system at either the "fast" or "slow" end of the frequency spectrum accurately and with minimal complexity. It is not uncommon, however, for one to pursue an accurate model of a system over one or more intermediate frequency bands. When modeling automobile noise, vibrations, and harshness (NVH), for instance, one is usually interested in vibration frequencies small enough to be perceptible but large enough to cause potential passenger discomfort or drivability issues.

Padé approximation is a frequency-based model reduction technique particularly suited to this class of problems. Given a complex model, it finds a lower-order approximation of the model by first constructing Laurent series expansions of the frequency responses of both models at one or more interpolation points. It then matches a small number of coefficients of these expansions to parameterize the reduced model.

In particular, let $G(s)$ represent the transfer function of the original – or "full" – model. Then its Laurent series expansion around some $s_0 \in \mathbb{C}$ is given by

$$G(s) = \sum_{k=0}^{\infty} a_k (s - s_0)^k \tag{19}$$

The goal is to find a lower order model with the transfer function

$$G_r(s) = \sum_{k=0}^{\infty} \hat{a}_k (s - s_0)^k \tag{20}$$

such that for a desired number $n \in \mathbb{N}_0$, the equalities $a_k = \hat{a}_k,\ k = 0,1,2,\ldots n,$ are satisfied. The coefficients $a_k, \hat{a}_k,\ k = 0,1,2,\ldots,$ are referred to as moments, and therefore this technique is also known as *moment matching*. When $s_0 = \infty$, the moments become the Markov parameters of the system, in which case the approximation problem can be solved using the Arnoldi procedure [62, 63] or the Lanczos procedure [64, 65]. When s_0 is arbitrary, the rational Krylov method [66, 67] can be used. It is also possible to use multiple interpolation points [65, 67].

Padé approximation is attractive when one seeks a good local approximation of a model around certain interpolation points in the frequency domain at a low computational cost. However, the stability of Padé approximants is, in general, not guaranteed, even if the models being approximated are stable. The literature describes some techniques that address this problem by extending Padé approximation to seek only stable reduced models [68]. Two other important limitations of Padé approximation remain even with these methods. First, there are no global error bounds for Padé approximants. Secondly, Padé approximation, by virtue of its dependence on the Laurent series expansion, is not a realization-preserving technique.

The starting point for Padé approximation is a Laurent series expansion of the frequency response of a given "full" model. If the full model is expressed as a rational polynomial transfer function, one may choose to obtain a proper model by truncating the polynomials in this transfer function directly, rather than expanding it into a Laurent series then performing moment matching. *Continued fraction expansion* is a polynomial approximation technique particularly suited to this scenario [69-73]. In particular, it builds on the fact that a transfer function given by

$$G(s) = \frac{a_{21} + a_{22}s + \ldots + a_{2,n}s^{n-1}}{a_{11} + a_{12}s + \ldots + a_{1,n-1}s^{n}} \tag{21}$$

can be written in the following continuous fraction expansion form

$$G(s) = \cfrac{1}{h_1 + \cfrac{1}{\cfrac{h_2}{s} + \cfrac{1}{h_3 + \cfrac{1}{\cfrac{h_4}{s} + \ddots}}}} \tag{22}$$

with

$$h_i = \frac{a_{i,1}}{a_{i+1,1}}, \quad i = 1,\ldots,2n \tag{23}$$

where the coefficients a_{i1} are the first elements of the rows of the table

$$\begin{array}{llll} a_{11} & a_{12} & a_{13} & \cdots \\ a_{21} & a_{22} & a_{23} & \cdots \\ a_{31} & a_{32} & \cdots & \\ a_{41} & \cdots & & \\ \cdots & & & \end{array}, \quad \begin{array}{c} a_{jk} = a_{j-2,k+1} - \dfrac{a_{j-2,1}a_{j-1,k+1}}{a_{j-1,1}} \\ j = 3,\ldots n+1; \quad k = 1,2,\ldots \end{array} \tag{24}$$

This particular expansion, known as the second Cauer form [73], is just one of the possible forms of continued fraction expansion. Given this expansion, a reduced transfer function of order *r* can be obtained by retaining the first 2*r* coefficients *h* and truncating the rest. This preserves the steady state component of the original transfer function [10]. Other forms that can be used for continued fraction expansion include the first and third Cauer forms and the Stieltjes form [10, 73].

The main drawback of the continued fraction expansion method in general is that, like Padé approximation, unstable reduced models can result from stable original models. The literature addresses this problem by proposing other polynomial approximation methods guaranteed to preserve model stability. One such method is *Routh approximation* [74], which is based on the fact that a transfer function given by

$$G(s) = \frac{b_{11} + b_{12}s + \ldots b_{1n}s^{n-1}}{a_{11} + a_{12}s + \ldots + a_{1,n+1}s^{n}} \tag{25}$$

can be put into a canonical form, known as the alpha-beta expansion, given by

$$G(s) = \beta_1 f_1(s) + \beta_2 f_1(s) f_2(s) + \ldots + \beta_n f_1(s) f_2(s) \ldots f_n(s) \tag{26}$$

where

$$f_1 = \frac{1}{1+\alpha_1 s}$$
$$f_i(s) = \cfrac{1}{\alpha_i s + \cfrac{1}{\alpha_{i+1} s + \cfrac{1}{\begin{matrix}\vdots \\ \alpha_{n-1} s + \cfrac{1}{\alpha_n s}\end{matrix}}}}, \quad i = 2,\dots n \tag{27}$$

and the coefficients α_i and β_i are given by

$$\begin{aligned} &\alpha_i = \frac{a_{i,1}}{a_{i,2}}, \; i = 1,\dots n \\ &\begin{matrix} a_{i,j} = a_{i-1,j+1} & j\ odd \\ a_{i,j} = a_{i-1,j+1} - \alpha_{i-1} a_{i-1,j+2} & j\ even \end{matrix}, \; i = 2,\dots n \end{aligned} \tag{28}$$

$$\begin{aligned} &\beta_i = \frac{b_{i,1}}{a_{i,2}}, \; i = 1,\dots n \\ &\begin{matrix} b_{i,j} = b_{i-1,j+1} & j\ odd \\ b_{i,j} = b_{i-1,j+1} - \beta_{i-1} a_{i-1,j+2} & j\ even \end{matrix}, \; i = 2,\dots n \end{aligned} \tag{29}$$

A reduced model of order *r* can then be obtained by

$$G_r(s) = \frac{1}{s} \hat{G}_r\left(\frac{1}{s}\right) \tag{30}$$

with

$$\hat{G}_r(s) = \beta_1 p_1(s) + \beta_r p_1(s) p_2(s) + \dots + \beta_r p_1(s) p_2(s) \dots p_r(s)$$
$$p_1(s) = f_1(s)$$
$$p_2(s) = \cfrac{1}{\alpha_i s + \cfrac{1}{\begin{matrix}\vdots \\ \alpha_{r-1} s + \cfrac{1}{\alpha_r s}\end{matrix}}}, \quad i = 2,\dots r \tag{31}$$

In addition to preserving stability, the Routh approximant also guarantees that the first *r* coefficients of the Taylor series expansions about $s = 0$ of the original and reduced models match. Furthermore, the impulse-response energies of Routh approximants converge monotonically to those of the original models, and the poles and zeros of the approximants approach the ones of the original model as *r* increases [74].

The literature describes other polynomial approximation methods that preserve stability, such as reduction based on *stability equations* [75]. Furthermore, the literature describes mixed methods that use different methods for approximating the numerator and denominator. These methods aim to resolve the instability problem of the Padé and continued fraction expansion methods, while matching some quantities of the original model. Typically, dominant pole retention or some other stability-preserving polynomial approximation method is used to calculate the denominator of the reduced model, while Padé or continued fraction expansion is used to determine the numerator. Some combinations include dominant pole retention and Padé approximation [16, 18, 19, 21], Routh stability criterion and Padé approximation [76], Routh array and Padé approximation [77, 78], stability equations and Padé approximation [79], and stability equations and continued fraction expansion [80]. Nevertheless, two drawbacks of the polynomial approximation methods in general still remain, namely, that all such methods are limited to linear systems, and they are not realization-preserving.

Oblique Projection

Even though this method is, as its name suggests, a projection-based method, due to its close relationship with the polynomial approximation methods it will be reviewed here. The relationship is in the sense that this method, using the oblique projection approach, gives a unified tool to simultaneously match high and low frequency moments of the transfer function, and high and low power moments of the power spectral density [81].

This method frames the model reduction problem as a projection of the original model

$$\begin{aligned} \dot{x} &= Ax + Bu \\ y &= Cx + Du \end{aligned} \tag{32}$$

into the reduced model

$$\begin{aligned} \dot{x}_r &= A_r x_r + B_r u \\ y &= C_r x_r + Du \end{aligned} \tag{33}$$

by $A_r = LAT$, $B_r = LB$, $C_r = CT$, and $LT = I$. Note that unlike aggregation, it is not required here that $x_r = Lx$ and $A_r L = LA$. Then, if L and T are chosen such that

$$L = \mathcal{O}_{p,q-1}(C);\; T = XL^T (LXL^T)^{-1} \tag{34}$$

where

$$\mathcal{O}_{p,q}(C) \triangleq \begin{bmatrix} CA^{-p} \\ CA^{-p+1} \\ \vdots \\ CA^{q} \end{bmatrix} \tag{35}$$

and X is the controllability Grammian satisfying

$$AX + XA^T + BB^T = 0 \tag{36}$$

then the reduced order model will be asymptotically stable if and only if it is controllable, and it will match *p* low frequency moments

$$M_i(0) = CA^{-i}B,\; i = 1,\dots,p \tag{37}$$

q high frequency moments

$$M_i(\infty) = CA^{i}B,\; i = 0,\dots,q-1 \tag{38}$$

p low frequency power moments

$$R_{ii}(0) = CA^{-i}X(A^T)^{-i}C^T,\ i = 1,\ldots,p \qquad (39)$$

and, *q* high frequency power moments

$$R_{ii}(\infty) = CA^{i}X(A^T)^{i}C^T,\ i = 0,\ldots,q-1 \qquad (40)$$

This basic idea has been extended to controller reduction at selected frequency regions, and also to matching the impulse response at selected time regions [81]. Due to its projection-based approach, this method is not realization-preserving.

Optimal Hankel Norm Approximation

The methods discussed so far deal with *local* approximations of a given system's frequency response. On the one hand, aggregation, singular perturbation, MODA, modal analysis, and component mode synthesis typically aim to approximate the low-frequency behavior of a given system. On the other hand, polynomial approximation methods typically aim to approximate the frequency response of a given system around some frequencies of interest.

Further extending these ideas, one may also seek a good approximation to a system's entire frequency response. Such an approximation may minimize, say, the $\mathcal{H}_\infty$ norm of the error $G - G_r$ between the full and proper models, but the resulting $\mathcal{H}_\infty$ model reduction problem does not have a known analytic solution. If, instead, one uses the Hankel norm of the error $G - G_r$ to quantify the "properness" of the reduced model, then an analytical solution for the optimal proper model does exist, and the resulting proper modeling technique is known as the *optimal Hankel norm approximation* [82-85].

For a given, stable, linear, and time-invariant system, *G*, Hankel norm approximation seeks an optimal reduced model, G_r, whose order, *k*, is specified *a priori* by the modeler. The resulting optimal proper model minimizes the Hankel norm of the error $G - G_r$ over the set of all linear and time-invariant models of the desired order. This highlights the implicit tradeoff between fidelity (measured by the Hankel norm of $G - G_r$) and complexity (measured by the order of G_r) that makes Hankel norm approximation a proper modeling method. Assuming that the state-space description of *G* is given by $\{A, B, C, D\}$, one possible way of finding G_r of order *k* is as follows [86]:

1. Calculate *P* and *Q*, the controllability and observability Grammians of the system *G*, respectively.
2. Calculate $E = QP - \sigma_{k+1}^2 I$, where $\sigma_{k+1} = \sqrt{\lambda_{k+1}(PQ)}$ is the *k*+1st Hankel singular value of *G*.
3. Find the singular value decomposition of *E*,
$$E = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix}$$
4. Apply the transformation
$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix} \left(\sigma_{k+1}^2 A^T + QAP\right) \begin{bmatrix} V_1 & V_2 \end{bmatrix}$$
$$\begin{bmatrix} B_1 \\ B_2 \end{bmatrix} = \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix} \begin{bmatrix} QB & -C^T \end{bmatrix}$$
$$\begin{bmatrix} C_1 & C_2 \end{bmatrix} = \begin{bmatrix} CP \\ -\sigma_{k+1} B^T \end{bmatrix} \begin{bmatrix} V_1 & V_2 \end{bmatrix}$$
5. Form the equivalent model
$$\tilde{G} = \begin{bmatrix} \tilde{A} & \tilde{B} \\ \tilde{C} & \tilde{D} \end{bmatrix} = \begin{bmatrix} \Sigma^{-1}\left(A_{11} - A_{12}A_{22}^{\dagger}A_{21}\right) & \Sigma^{-1}\left(B_1 - A_{12}A_{22}^{\dagger}B_2\right) \\ C_1 - C_2A_{22}^{\dagger}A_{21} & D \end{bmatrix}$$
6. The equivalent model can be decomposed additively into a stable part G_- with *k* stable poles and an anti-stable part G_+ with all poles unstable, i.e., $\tilde{G} = G_- + G_+$. Then, G_- is the *k*-th order optimal Hankel norm approximation of the system *G*, i.e., $G_r = G_-$.

The Hankel norm of the approximation error of any *k*-th order system $\hat{G}_r$ is lower-bounded by $\left\|G - \hat{G}_r\right\|_H \ge \sigma_{k+1}(G)$, and the equality in the error bound is satisfied only by the optimal Hankel norm approximation G_r.

This minimization of error in terms of the Hankel norm comes at the expense of a change in realization due the transformations applied during the calculation of the reduced model. Therefore, the optimal Hankel norm approximation is not a realization-preserving method.

It is worth noting that even though the Hankel norm approximation does not optimize $\mathcal{H}_\infty$ norm of the error, there still exists an $\mathcal{H}_\infty$ error bound, as established first by Glover [85]

$$\left\|G(j\omega) - G_r(j\omega)\right\|_\infty \le 2\sum_{i=k+1}^{n} \sigma_i \qquad (41)$$

It is important to note that the *D* matrix does not affect the Hankel optimality of the approximation, but it does affect the $\mathcal{H}_\infty$ norm of the error. It is possible to choose $\tilde{D}$ in such a way that upper-bound on the $\mathcal{H}_\infty$ norm of the error is cut in half, i.e.,

$$\left\|G(j\omega) - G_r(j\omega) - \tilde{D}\right\|_\infty \le \sum_{i=k+1}^{n} \sigma_i \qquad (42)$$

Please see [85] for the calculation of such a $\tilde{D}$.

The above results for continuous systems have also been extended to discrete-time systems [87-90].

III. PROJECTION-BASED TECHNIQUES

The frequency-based proper modeling techniques discussed hitherto assume, in general, that the salient dynamics of a given system occur over a fairly limited range in the frequency domain. *Projection-based* techniques make a conceptually analogous assumption in the state domain. Specifically, they all assume that the salient dynamics of a given system are limited to a portion of the system's entire state space. They search for this portion – or *subspace* – by searching for the basis vectors spanning it, and they differ in the ways they choose the basis vectors. This section presents three projection-based model reduction techniques, namely, *proper orthogonal decomposition*, *balanced truncation*, and *component cost analysis*. The first two are based on the *Karhunen-Loève expansion*, which we discuss first.

Karhunen-Loève Expansion

The *Karhunen-Loève expansion* [91, 92], also known as principal component analysis [93], the method of empirical orthogonal functions [94], proper orthogonal decomposition [95], singular value decomposition [96], empirical eigenfunction decomposition [97-99], or the method of quasi-harmonic modes [100], is a correlation analysis tool that is a key foundation for most projection-based proper modeling techniques. It can be implemented in a numerically efficient manner using the *method of snapshots* [97-99], and has become widely popular in many fields including fluid dynamics, structural vibrations, image processing, and signal analysis.

Given observation data from either a physical system or its model, the Karhunen-Loéve expansion finds a subspace that captures the dominant dynamics of this system. Specifically, it finds the orthogonal basis that optimally captures the energy of the observation signals, in the least-squares sense. Selecting those basis vectors that capture the most observation signal energy furnishes a subspace that captures the dominant system dynamics. Projecting the system's model onto this subspace using the *Galerkin projection* method then furnishes a reduced model that captures the original system's dominant dynamics. This process leads to a powerful model reduction technique.

For time-invariant finite-dimensional systems, the Karhunen-Loéve expansion method can be applied as follows. Consider a system represented by a state space equation of the form

$$\dot{x} = f(x,u), \quad x \in \mathbb{R}^n \tag{43}$$

Assume that $m \geq n$ observations are made for each state and arranged in matrix form such that

$$A = \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix}_{m\times n} \tag{44}$$

Obtain the singular value decomposition of the matrix *A*, i.e.,

$$A = U\Sigma V^T \tag{45}$$

where $\Sigma = diag(\sigma_1, \sigma_2, \dots, \sigma_n)_{m\times n}$ with $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n \geq 0$. The columns of the orthogonal $n\times n$ matrix *V* form a basis of the state space, and the squares of the singular values provide a measure of how much signal energy is captured by each of these basis vectors. Assume that the last $n-k$ singular values are small, where $k < n$. Then, a reduced order model can be obtained by taking the first *k* columns of the *V* matrix, and projecting the state space onto the subspace spanned by those *k* vectors, i.e.,

$$\dot{x}_r = (V_k)^T f(V_k x_r, u), \quad \overline{x} = V_k x_r \tag{46}$$

where $\overline{x}$ is the approximation to the original state vector *x*. The motivation for using the first *k* columns as a basis for the reduced model is the fact that the rank *k* approximation $A_k = U_k \Sigma_k (V_k)^T$ to the original observation matrix *A* is optimal in a least squares sense. Here U_k and V_k denote the first *k* columns of the matrices *U* and *V*, respectively, and Σ_k denotes the leading $k\times k$ principal minor of the matrix Σ. This optimality is guaranteed for any value of *k*. Furthermore, an error bound exists for the approximation error $A - A_r$, which is given by

$$\left\| A - A_r \right\|_F = \sum_{i=k+1}^{n} \sigma_i^2 \tag{47}$$

where $\left\| \bullet \right\|_F$ denotes the Frobenius norm [101, 102]. Note, however, that the optimality and the error bound are valid only for the approximation to the observation matrix, and not for the reduced order model, i.e., no bound exists for $\left\| x - \overline{x} \right\|$. In fact, unstable reduced models may result from stable original models. Nevertheless, this technique often yields good results and is widely used for model reduction due to its applicability to nonlinear systems as well.

In case the state variable is a function of position and time, $z(x,t)$, which is common in fluid mechanics or in structural vibrations, the same technique can be used to obtain empirical modes, such that the state variable can be approximated as

$$z(x,t) \approx \sum_{i=1}^{M} a_i(x) b_i(t) \tag{48}$$

In this case the observation matrix can be arranged as:

$$A = \begin{bmatrix} z(x_1,t_1) & z(x_2,t_1) & \cdots & z(x_n,t_1) \\ \vdots & \vdots & \vdots & \vdots \\ z(x_1,t_m) & z(x_2,t_m) & \cdots & z(x_n,t_m) \end{bmatrix}_{m\times n} \tag{49}$$

Then, the columns of the *U* matrix in the singular value decomposition in Eq. (45) give the empirical modes known as the proper orthogonal modes and the squares of the diagonal elements of Σ describe how much signal energy is captured by each mode. When used this way, the Karhunen-Loève expansion is similar to the modal analysis technique described in Section II in the approach to obtaining reduced models; namely, by assuming that the total response is a combination of some modal responses and retaining the dominant modes in the reduced model. Note, however, that the modes in the Karhunen-Loève expansion are empirical.

Balanced Truncation

The Karhunen-Loéve expansion can be applied to a wide variety of dynamic system models for the purpose of modeling them properly. This includes linear and nonlinear, time-invariant and time-varying systems. The Karhunen-Loéve expansion can also be applied to the same system for different state and input trajectories. This could ostensibly furnish significantly different proper models, each being "proper" only in the context of the trajectory used for obtaining it.

Balanced truncation is a special model reduction technique that involves applying the Karhunen-Loéve expansion in particular ways to particular classes of systems. Its simplest rendition was originally proposed by Moore [103]. Specifically, Moore suggested the application of the Karhunen-Loéve expansion to the state trajectory of the *balanced realization* of a linear and time-invariant system subjected to a series of impulses. A system's realization is balanced if its observability and controllability Grammians are equal, meaning that each state is as observable as it is controllable. When this is done, one finds that the less observable and less controllable states can be eliminated from the given system's model to furnish a reduced model. This *balanced truncation* process is a very interesting and powerful generalization of the Kalman canonical decomposition, which only eliminates the completely unobservable and completely uncontrollable states from a given system model to furnish a minimal realization of the model [104]. Note, however, that due to balancing the realization of the system changes, and balanced truncation is therefore not realization-preserving.

The balanced truncation technique proceeds mathematically as follows. First, it applies a state transformation to put the original model in a form where each state is equally controllable and observable. In this case, the controllability and observability matrices P and Q become diagonal, with the diagonal elements being the Hankel singular values, i.e., $P = Q = diag(\sigma_1, \sigma_2, \ldots, \sigma_n)$, where $\sigma_i = \sqrt{\lambda_i(PQ)}$ are the Hankel singular values, which give a measure for the controllability and observability of corresponding states. Based on this measure, less controllable and observable states are truncated. There exists a global $\mathcal{H}_\infty$ error bound, which is the same as the $\mathcal{H}_\infty$ error bound in the Hankel norm approximation technique for the case when the $\tilde{D}$ matrix is not optimized, i.e.,

$$\left\| G - G_r \right\|_\infty \le 2 \sum_{i=k+1}^{n} \sigma_i \tag{50}$$

where σ_i are the Hankel singular values of G corresponding to the truncated states. Note, however, that in Hankel norm approximation $\tilde{D}$ can be chosen such that only half of the $\mathcal{H}_\infty$ error bound of balanced truncation is achieved.

It is important to note the norm that is used in Eq. (50), because the singular values may not be as informative for other norms. As first shown by Kabamba, the singular values by themselves are not descriptive enough for the $\mathcal{L}_2$ norm of error [105]. Therefore, Kabamba introduced other invariants of the system, the *balanced gains*, that together with the singular values describe the contribution of each state to the $\mathcal{L}_2$ norm of the impulse response [105].

There is an interesting relationship between balanced truncation and singular perturbation. The *generalized singular perturbation approximation* allows for matching the magnitude of the original model at a desired frequency $s = s_0$, and choosing $s_0 = 0$ corresponds to the singular perturbation as given earlier in the paper, whereas choosing $s_0 = \infty$ corresponds to direct truncation [40]. Thus, assuming the original model is balanced, choosing $s_0 = \infty$ corresponds to balanced truncation, and furthermore, singular perturbation, i.e. choosing $s_0 = 0$, achieves the same error bound as the balanced truncation [40].

The literature describes many extensions of the above balanced truncation technique. These extensions include approximate balancing techniques that can be quite valuable when exact balancing is computationally costly [106-108]. Further extensions extend balanced truncation specifically to stochastic [109-111], passive [109], and bounded real systems [112]. The literature also describes LQG balancing techniques for reduced order controller design [113] and frequency-weighted balanced truncation for reducing the approximation error over a specified frequency range rather than the whole spectrum [114-118]. Significant research has also pursued the balanced truncation of nonlinear systems [119-123]. This literature highlights the importance of balanced truncation, both as a powerful model reduction technique and as the basis for very extensive ongoing research, both theoretical and applied.

Component Cost Analysis

Another method that can be reviewed under the projection-based techniques category is component cost analysis [124-128]. In this approach, a cost function for the linear stable system

$$\begin{aligned} \dot{x} &= Ax + Bu, \quad && x \in \mathbb{R}^n, u \in \mathbb{R}^m \\ y &= Cx, && y \in \mathbb{R}^p \end{aligned} \tag{51}$$

is defined as

$$V \triangleq \lim_{t \to \infty} E[\left\| y \right\|^2]; \quad \left\| y \right\|^2 = y^T y \tag{52}$$

This cost function satisfies the cost decomposition property

$$V = \sum_{i=1}^{n} V_i \tag{53}$$

where V_i is the contribution of the ith state, x_i, to the system cost, and is given by

$$V_i = \left[XC^T C \right]_{ii} \tag{54}$$

where X is the controllability Grammian, satisfying

$$XA^T + AX + BB^T = 0 \tag{55}$$

The reduced model is then obtained by truncating the low-cost states based on the rationale that the system cost should be perturbed minimally. However, it is important to know that deleting x_i, in general, does not necessarily cause a change of ΔV_i in V.

Note that the component cost analysis in this most basic form does not require a state transformation. Nevertheless, if the system is transformed into cost-decoupled coordinates, where $XC^T C$ is diagonal, the component costs also quantify the amount by which the system cost will change if the corresponding states were truncated from the model. Furthermore, in these coordinates $n - r_C$ components will have zero component costs, where r_C is the rank of the matrix C. Therefore, in these coordinates a reduced model can be obtained that preserves the system cost. Cost decoupled coordinates are not unique, and one possible transformation into the cost-decoupled coordinates is given by

$$\begin{gathered} x = Tz \\ T = T_1 U;\ X = T_1 T_1^T;\ T_1^T C^T C T_1 = U \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} U^T \end{gathered} \tag{56}$$

There is a close connection between component cost analysis and the idea of balanced gains introduced by Kabamba [105]. Specifically, if component cost analysis is applied to the balanced coordinates, the component costs exactly match Kabamba's results [126].

Furthermore, a very interesting relationship exists between balanced realization and cost-decoupled coordinates [128]. A generalization of the basic component cost analysis defines

$$\begin{aligned} y' \triangleq \begin{bmatrix} y \\ \dot{y} \\ \ddot{y} \\ \vdots \\ y^{(q-1)} \end{bmatrix} &= \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{q-1} \end{bmatrix} x + \begin{bmatrix} 0 & 0 & \cdots & 0 \\ CB & 0 & & \vdots \\ CAB & CB & & \vdots \\ \vdots & & \ddots & \vdots \\ CA^{q-2}B & \cdots & CB & 0 \end{bmatrix} \begin{bmatrix} u \\ \dot{u} \\ \ddot{u} \\ \vdots \\ \vdots \end{bmatrix} \\ &= \tilde{C}x + \tilde{D}u' \end{aligned} \tag{57}$$

and considers the system

$$\begin{aligned} \dot{x} &= Ax + Bu \\ y' &= \tilde{C}x + \tilde{D}u' \end{aligned} \tag{58}$$

with the cost function

$$V \triangleq \sum_{k=1}^{m} \int_0^{\infty} \bar{y}^T(k,t) Q \bar{y}(k,t) dt \tag{59}$$

where $\bar{y}(k,t)$ is the response of the system for an impulse at the k^{th} input channel while all other inputs being zero, and Q is a weight matrix. Then, the cost-decoupling transformation

$$\tilde{T} = T_1 U \begin{bmatrix} \Sigma^{-1/4} & 0 \\ 0 & I \end{bmatrix} \tag{60}$$

yields the balanced coordinates, if

$$q = n,\ Q = \int_0^{\infty} \begin{bmatrix} \alpha_0(t) \\ \vdots \\ \alpha_{q-1}(t) \end{bmatrix} I_p \left[\alpha_0(t) I \quad \cdots \quad \alpha_{q-1}(t) I \right] dt \tag{61}$$

where $e^{At} = \sum_{i=0}^{n-1} \alpha_i(t) A^i$. These results imply that balanced coordinates are a special case of the generalized cost-decoupled coordinates, and thus the component cost analysis is a generalization of the balanced truncation.

IV. OPTIMIZATION-BASED TECHNIQUES

The frameworks of both frequency- and projection-based proper modeling techniques are based on the same goal: to identify and isolate the dominant characteristics of a given model. For frequency-based methods these characteristics lie in the frequency domain, and for projection-based methods they are in the state space.

In addition to this rather intuitive and practical motivation of retaining the model's dominant characteristics, one may also seek to formally achieve a minimal difference between the predictions of the full and reduced models subject to a complexity constraint. Such techniques are referred to as *optimization-based* proper modeling techniques in this paper.

Optimal Hankel norm approximation, for instance, is an optimization-based proper modeling technique, because it seeks to minimize the Hankel norm of the difference between a full model and a reduced model, subject to a bound on the reduced model's order. The fact that optimal Hankel norm approximation is also a frequency-based proper modeling technique underscores the fact that our classification of proper modeling techniques, while intuitively appealing, is certainly not strict. Interestingly, optimal Hankel norm approximation is also a projection-based proper modeling technique. This raises an important question, namely, whether one can formulate a "unified" model reduction problem: one that simultaneously seeks optimality in the frequency and state space domains.

The above question was partly answered by Hyland and Bernstein's seminal work on the *optimal projection equations* [129]. In this work, Hyland and Bernstein formulated the proper modeling problem as a problem of minimizing a quadratic measure of the error between a full model and its proper counterpart, subject to implicit rank constraints on the proper counterpart. This furnished a set of first-order necessary conditions for optimality of the reduced proper model, which Hyland and Bernstein expressed as a coupled system of two Lyapunov equations. Hyland and Bernstein then studied balanced truncation in the context of these necessary conditions for proper model optimality. They found that balanced truncation furnished reduced models that deviate significantly from quadratic optimality: a conclusion also supported by earlier research by Kabamba [105]. The significance of this finding cannot be overemphasized. It highlights the fact that a "proper" model developed using one metric (e.g., the relative observability and controllability of different states) can be far from being "proper" in the context of a different metric (e.g., quadratic optimality). In other words, there is no universal proper modeling algorithm applicable to all systems under all circumstances. Rather, different proper modeling algorithms are better suited to different problems, and one should carefully select the proper modeling metric ideally suited for the problem at hand.

Optimization-based proper modeling techniques typically seek to minimize the $\mathcal{L}_2$, $\mathcal{H}_2$, or $\mathcal{H}_\infty$ norm of the difference between a given "full" model and its proper counterpart, subject to a constraint on the order (i.e., "complexity") of the proper counterpart. Wilson, for instance, was the first to address the minimization of the $\mathcal{L}_2$ norm of error in model reduction [130]. Howitt and Luus, give another example in which they optimize the pole and zero locations of a reduced model to minimize the integral square error of the difference between the impulse responses of the full and reduced models [131]. Similarly, Luus optimizes a reduced model to minimize the deviation of its frequency response from that of the corresponding full model [132]. The proper modeling problems resulting from such formulations often do not have analytic solutions, and must hence be solved numerically.

As a result, much of the optimization-based proper modeling literature focuses on the development of numerically efficient optimization algorithms, with special attention to the convergence properties of these algorithms. Gouda *et al.*, for instance, obtain a proper model of a building's thermal response using sequential quadratic programming [133]. Similarly, Hachtel *et al.* propose an interactive optimization technique incorporating linear programming as a tool for nonlinear model reduction [134]. Both linear and sequential quadratic programming are local search techniques that may not be able to find globally optimal proper models. With this in mind, Assunção and Peres propose a branch-and-bound algorithm for the solution of the optimal $\mathcal{H}_2$-norm-based proper modeling problem [135]. Finally, Chen and Fang [136], Spanos *et al.* [137], and Ferrante *et al.* [138] propose reduced model optimization algorithms that have attractive mathematical guarantees of convergence.

Optimization-based approaches may or may not be realization-preserving, depending on whether they fix the given system's realization during the search for an optimal reduced model or allow it to vary. While most optimization-based approaches in the literature are not realization-preserving, it is certainly possible to construct ones that are.

V. ENERGY-BASED TECHNIQUES

Energy-based proper modeling techniques are built on the intuitive fundamental premise that in an energetic system, the most important components to model accurately are those characterized by the largest magnitudes of energy (or power) flow. Therefore, these algorithms simplify a given model by eliminating less energetic components, while trying to minimize the effect of the elimination on the overall energy flow. The well-known Rayleigh-Ritz method exemplifies this perspective on model reduction [49]. Other energy-based model reduction algorithms include statistical energy analysis [139] and the power-based model reduction algorithm by Rosenberg and Zhou [140, 141].

Rosenberg and Zhou's model reduction algorithm [140, 141] is based on the intuitive notion that in an energetic dynamic system those components characterized by higher mean-square energies should be more important to model than those characterized by lower mean-square energies. This leads to a simple, intuitive, realization-preserving, and powerful model reduction technique with no theoretical proof for convergence, reduced model stability, or "optimality".

Louca *et al.* extend Rosenberg and Zhou's algorithm by proposing a new energy-based model reduction metric called *activity* [142]. The activity of an energetic element is defined as the time integral of the absolute value of the power flowing through it over a particular time-window for a particular input. In a bond-graph setting, where the flow through an element *i* and the effort across it are denoted as f_i and e_i, respectively, the element's activity is defined as

$$A_i \triangleq \int_0^T \left| f_i(t) e_i(t) \right| dt \tag{62}$$

where *T* is the width of the desired time-window. The activity of an element can, hence, be physically interpreted as the total energy flow through the element within a specified time-window for a specific input. It can also be interpreted as the $\mathcal{L}_1$ norm of the power flow through the element, multiplied by the width of the time window used to compute that norm.

Louca *et al.* conjectured that in an energetic system, the more active elements are more important to model than the less active elements. An *element*, in this context, is any component in the system's bond-graph representation, including generalized resistors, capacitors, and inductors. Based on this conjecture, Louca *et al.* proposed an activity-based realization-preserving *model order reduction algorithm (MORA)* [142], and developed techniques for physically interpreting the reduced models generated by this algorithm [143].

The fundamental premise behind MORA, namely, that activity can be used as a proper modeling metric, is mostly intuitive. However, it is supported by some important application studies [144-146]. Furthermore, recent work by Fathy and Stein has unveiled *fundamental concordances* between MORA and balanced truncation [147]. These concordances are special cases where the two algorithms are mathematically guaranteed to furnish identical reduced models. While these concordances do not provide a general mathematical foundation for MORA, they do lend credence to MORA as a mathematical model reduction algorithm, at least in the special cases covered by the concordances [147].

Beyond its viability as a model reduction metric, activity has also proven viable as a model *partitioning* metric. Specifically, Rideout *et al.* use activity to quantitatively and systematically look for decoupling among the elements of a model and to partition the model based on the discovered decoupling [148]. Once the partitions are obtained, the simulation can be carried out either by simulating the driving partition first and using its output as an input to the driven system, or, in case only the driving partition is of interest, by completely eliminating the driven partition and keeping only the driving partition.

VI. DISCUSSION AND CONCLUSIONS

The process of modeling a dynamic system invariably entails a tradeoff between model accuracy and simplicity. Simpler models can be easier to simulate, analyze, comprehend, and control than more complex ones, but this often comes at the expense of accuracy and, hence, potential viability. Recognizing this fundamental tradeoff, the literature deems a model “proper” if it balances the needs for accuracy and simplicity.

The formal definition of “proper” models may be relatively new [1], but its underlying emphasis on the need for balancing model fidelity and complexity has been recognized for many decades. In fact, the literature presents many techniques for *reducing* complex models until they become proper, or *deducing* proper system models from simpler subsystem models. This paper briefly surveys these techniques and classifies them into *frequency-*, *projection-*, *optimization-*, and *energy-based* depending on their underlying metrics for assessing the relative importance of a model’s different dynamics and subsystems. This classification is neither well-established nor strict, as evident from the fact that a given proper modeling algorithm often belongs to more than one of these categories. However, the authors have found it convenient for both presentation and pedagogy, and hence adopt it herein.

A careful examination of the different proper modeling techniques in the literature leads to the fundamentally important conclusion that *there is no universal proper modeling technique suitable for all modeling problems and all applications*. Rather, different proper modeling techniques are often better suited to different problem spaces, and the authors hope that this review may be used as a guide in selecting the appropriate method.

Despite the richness of the proper modeling literature, many important problems remain to be addressed. In particular, in many circumstances, it may be possible to simplify a given model and thus make it proper not only by *reducing* or eliminating its various submodels but also by simplifying the interconnections between these submodels. Such *model structure simplification* includes simplifying a model by lumping its coupled inertias, partitioning its weakly coupled subsystems, or simplifying its mathematical representation without loss of accuracy. This paper touches briefly on one of these aspects of model structure simplification, namely, model partitioning. For brevity, however, it does not explore the complete model structure simplification area and the significant ongoing research pertaining to it.

For simplicity, the paper also focuses mostly on the deterministic proper modeling problem. The notion of a “proper model” becomes particularly powerful in the context of systems with significant uncertainties. In particular, when modeling a stochastic system, one may legitimately ask: which of the system’s various uncertainties are more important to model, and which are negligible? This leads to the notion of a *stochastic proper model*: one capturing only the most salient dynamics *and uncertainties* of a given system. Significant research exists, and continues, in the area of stochastic proper modeling, but this paper focuses on deterministic proper modeling for brevity.

Finally, it is important to note that proper models of dynamic systems are often a means to an important practical end. In particular, the ultimate goal of any proper system modeling exercise is often to not only better understand the system’s behavior, but also to use this understanding as a means towards better system designs and controls. This implies that a proper model must, therefore, be both *scalable* and *control-oriented*. A system model is *scalable* if it captures not only the dynamics of a given system, but also how these dynamics change with system design parameters. Furthermore, a system model is *control-oriented* if it accurately captures those dynamics that are most important for the effective control of the given system. Both scalable and control-oriented modeling are rapidly becoming active research topics, and a thorough discussion of these topics is omitted from this paper for brevity.

REFERENCES

[1] Wilson, B. H. and Stein, J. L., 1992, "An Algorithm for Obtaining Minimum-Order Models of Distributed and Discrete Systems", *Proceedings of Winter Annual Meeting of the American Society of Mechanical Engineering, Nov 8-13 1992*, Publ by ASME, New York, NY, USA, **41**, pp. 47-58.

[2] Assanis, D., Bryzik, W., Chalhoub, N., Filipi, Z., Henein, N., Jung, D., Liu, X., Louca, L., Moskwa, J., Munns, S., Overholt, J., Papalambros, P., Riley, S., Rubin, Z., Sendur, P., Stein, J., and Zhang, G., 1999, "Integration and Use of Diesel Engine, Driveline and Vehicle Dynamics Models for Heavy Duty Truck Simulation", 1999-01-0970, *Proceedings of 1999 SAE Congress*, SAE.

[3] Kozaki, T., Mori, H., Fathy, H. K., and Gopalswamy, S., 2004, "Balancing the Speed and Fidelity of Automotive Powertrain Models through Surrogation", *Proceedings of 2004 ASME International Mechanical Engineering Congress and Exposition*, ASME, New York, NY, **73**, pp. 249-258.

[4] Skogestad, S. and Postlethwaite, I., 1996, *Multivariable Feedback Control: Analysis and Design*, John Wiley and Sons.

[5] Elrazaz, Z. and Sinha, N. K., 1981, "A Review of Some Model Reduction Techniques", Canadian Electrical Engineering Journal, **6**(1), pp. 34-40.

[6] Lamba, S. S. and Mahmoud, M. S., 1982, "Model Simplification - an Overview", *Proceedings of Theory and Application of Digital Control, Proceedings of the IFAC Symposium.*, IFAC by Pergamon Press, Oxford, Engl, pp. 479-487.

[7] Saksena, V. R., O'Reilly, J., and Kokotovic, P. V., 1984, "Singular Perturbations and Time-Scale Methods in Control Theory: Survey 1976-83", Automatica, **20**(3), pp. 273-293.

[8] Bertrand, P., Duc, G., and Michailesco, G., 1985, "Développements Récents Sur La Réduction De Modèles", Automatique-Productique Informatique Industrielle, **19**(2), pp. 131-146.

[9] Sugimoto, S., Kaino, E., and Mori, Y., 1985, "Comparative Studies for Several Model Reduction Algorithms", *Proceedings of International Conference - Control 85 (IEE Conf. Publ. No. 252)*, IEE, pp. 673-678.

[10] Fortuna, L., Nunnari, G., and Gallo, A., 1992, *Model Order Reduction Techniques with Applications in Electrical Engineering*, Springer-Verlag, London.

[11] Gugercin, S. and Antoulas, A. C., 2000, "A Comparative Study of 7 Algorithms for Model Reduction", *Proceedings of 39th IEEE Conference on Decision and Control*, Institute of Electrical and Electronics Engineers Inc., **3**, pp. 2367-2372.

[12] Antoulas, A. C., Sorensen, D. C., and Gugercin, S., 2001, "A Survey of Model Reduction Methods for Large Scale Systems", Contemporary Mathematics, **280**, pp. 193-219.

[13] Antoulas, A. C. and Gugercin, S., 2002, "A New Approach to Model Reduction Which Preserves Stability and Passivity: An Overview", *Proceedings of IEEE Conference on Decision and Control*, Institute of Electrical and Electronics Engineers Inc., **3**, pp. 2544-2545.

[14] Gugercin, S. and Antoulas, A. C., 2004, "A Survey of Model Reduction by Balanced Truncation and Some New Results", International Journal of Control, **77**(8), pp. 748-766.

[15] Qu, Z.-Q., 2004, *Model Order Reduction Techniques : With Applications in Finite Element Analysis*, Springer, London ; New York.

[16] Davison, E. J., 1966, "A Method for Simplifying Linear Dynamic Systems", IEEE Transactions on Automatic Control, **AC-11**(1), pp. 93-101.

[17] Marshall, S. A., 1966, "An Approximate Method for Reducing the Order of a Linear System", Control, **10**, pp. 642-643.

[18] Chidambara, M. R., 1967, "On "A Method for Simplifying Linear Dynamic Systems"", IEEE Transactions on Automatic Control, **AC-12**(6), pp. 119-121.

[19] Chidambara, M. R., 1967, "Further Comments On "A Method for Simplifying Linear Dynamic Systems"", IEEE Transactions on Automatic Control, **AC-12**(6), pp. 799-800.

[20] Chidambara, M. R. and Davison, E. J., 1967, "Further Remarks on Simplifying Linear Dynamic Systems", IEEE Transactions on Automatic Control, **AC-12**(2), pp. 213-214.

[21] Davison, E. J., 1968, "A New Method for Simplifying Large Linear Dynamic Systems", IEEE Transactions on Automatic Control, **AC-13**(2), pp. 215-216.

[22] Fossard, A., 1970, "On a Method for Simplifying Linear Dynamic Systems", IEEE Transactions on Automatic Control, **AC-15**(2), pp. 261-262.

[23] Mitra, D., 1967, "On the Reduction of Complexity of Linear Dynamic Models", AEEW-R520, United Kingdom Atomic Energy Authority, Winfrith.

[24] Mitra, D., 1969, "The Reduction of Complexity of Linear Time Invariant Dynamical Systems", *Proceedings of IFAC World Congress*, pp. 19-33.

[25] Aoki, M., 1968, "Control of Large-Scale Dynamic Systems by Aggregation", IEEE Transactions on Automatic Control, **AC-13**(3), pp. 246-253.

[26] Hickin, J. and Sinha, N. K., 1975, "Aggregation Matrices for a Class of Low-Order Models for Large-Scale Systems", Electronics Letters, **11**(9), pp. 186.

[27] Siret, J. M., Michailesco, G., and Bertrand, P., 1977, "Representation of Linear Dynamical Systems by Aggregated Models", International Journal of Control, **26**(1), pp. 121-128.

[28] Hickin, J., 1978, "A Unified Theory of Model Reduction for Linear Time Invariant Dynamical Systems", Ph.D. Dissertation, McMaster University, Hamilton, Ontario, Canada.

[29] Aoki, M., 1978, "Some Approximation Methods for Estimation and Control of Large Scale Systems", IEEE Transactions on Automatic Control, **AC-23**(2), pp. 173-182.

[30] Hickin, J. and Sinha, N. K., 1980, "Model Reduction for Linear Multivariable Systems", IEEE Transactions on Automatic Control, **AC-25**(6), pp. 1121-1127.

[31] Prandtl, L., 1904, "Über Flüssigkeitsbewegung Bei Sehr Kleiner Reibung", *Proceedings of Verhandlung des III. Internationalen Mathematiker Kongresses*.

[32] Tikhonov, A., 1948, "On the Dependence of the Solutions of Differential Equations on a Small Parameter", Mat. Sb., **22**, pp. 193-204.

[33] Levinson, N., 1950, "Perturbations of Discontinuous Solutions of Non-Linear Systems of Differential Equations", Acta Mathematica, **82**, pp. 71-106.

[34] Vasileva, A. B., 1963, "Asymptotic Behavior of Solutions to Certain Problems Involving Nonlinear Differential Equations Containing a Small Parameter Multiplying the Highest Derivatives", Russian Mathematical Surveys, **18**, pp. 13-81.

[35] Wasow, W., 1965, *Asymptotic Expansions for Ordinary Differential Equations*, Wiley-Interscience, New York.

[36] Kokotovic, P. and Sannuti, P., 1968, "Singular Perturbation Method for Reducing Model Order in Optimal Control Design", IEEE Transactions on Automatic Control, **AC-13**(4), pp. 377-384.

[37] Kokotovic, P. V., O'Malley, R. E., Jr., and Sannuti, P., 1976, "Singular Perturbations and Order Reduction in Control Theory-an Overview", Automatica, **12**(2), pp. 123-132.

[38] Kokotovic, P. V., 1984, "Applications of Singular Perturbation Techniques to Control-Problems", Siam Review, **26**(4), pp. 501-550.

[39] Kokotovic, P., Khalil, H. K., and O'Reilly, J., 1986, *Singular Perturbation Methods in Control: Analysis and Design*, Academic Press, London.

[40] Liu, Y. and Anderson, B. D. O., 1989, "Singular Perturbation Approximation of Balanced Systems", International Journal of Control, **50**(4), pp. 1379-405.

[41] Ferris, J. B., Stein, J. L., and Bernitsas, M. M., 1994, "Development of Proper Models of Hybrid Systems", *Proceedings of ASME Winter Annual Meeting, Symposium on Automated Modeling*, pp. 629-636.

[42] Walker, D. G., Stein, J. L., and Ulsoy, A. G., 2000, "An Input-Output Criterion for Linear Model Deduction", Journal of Dynamic Systems, Measurement, and Control, **122**(3), pp. 507-513.

[43] Wilson, B. H. and Taylor, J. H., 1998, "A Frequency Domain Model-Order-Deduction Algorithm for Linear Systems", Transactions of

the ASME. Journal of Dynamic Systems, Measurement and Control, **120**(2), pp. 185-192.

[44] Taylor, J. H. and Wilson, B. H., 1995, "A Frequency-Domain Model-Order-Deduction Algorithm for Nonlinear Systems", *Proceedings of International Conference on Control Applications*, IEEE, pp. 1053-1058.

[45] Pirvu, A.-M., Dauphin-Tanguy, G., and Kubiak, P., 2007, "Automatic System for Bond Graph Model Adaptation - Application to an Electro-Hydrostatic Actuator", *Proceedings of 2007 International Conference on Bond Graph Modeling and Simulation (ICBGM '07)*, J. Granda and F. Cellier, eds., SCS, pp. 35-41.

[46] Kisung, S., Zhun, F., Jianjun, H., Goodman, E. D., and Rosenberg, R. C., 2003, "Toward a Unified and Automated Design Methodology for Multi-Domain Dynamic Systems Using Bond Graphs and Genetic Programming", Mechatronics, **13**(8-9), pp. 851-885.

[47] Fan, Z., Seo, K., Hu, J., Goodman, E. D., and Rosenberg, R. C., 2004, "A Novel Evolutionary Engineering Design Approach for Mixed-Domain Systems", Engineering Optimization, **36**(2), pp. 127-147.

[48] Margolis, D. L. and Young, G. E., 1977, "Reduction of Models of Large Scale Lumped Structures Using Normal Modes and Bond Graphs", Journal of the Franklin Institute, **304**(1), pp. 65-79.

[49] Fertis, D. G., 1995, *Mechanical and Structural Vibrations*, John Wiley & Sohns, New York.

[50] Mohanty, P. and Rixen, D. J., 2004, "Operational Modal Analysis in the Presence of Harmonic Excitation", Journal of Sound and Vibration, **270**(1-2), pp. 93-109.

[51] Vukazich, S. M., Mish, K. D., and Romstad, K. M., 1996, "Nonlinear Dynamic Response of Frames Using Lanczos Modal Analysis", Journal of Structural Engineering, **122**(12), pp. 1418-1426.

[52] Li, A. and Dowell, E. H., 2004, "Asymptotic Modal Analysis of Dynamical Systems: The Effect of Modal Cross-Correlation", Journal of Sound and Vibration, **276**(1-2), pp. 65-80.

[53] Hurty, W. C., 1960, "Vibrations of Structural Systems by Component Mode Synthesis", American Society of Civil Engineers -- Proceedings-- Journal of the Engineering Mechanics Division, **86**(EM4, Part 1), pp. 51-69.

[54] Hurty, W. C., 1965, "Dynamic Analysis of Structural Systems Using Component Modes", AIAA Journal, **3**(4), pp. 678-685.

[55] Craig, J., R.R. and Bampton, M. C. C., 1968, "Coupling of Substructures for Dynamic Analyses", AIAA Journal, **6**(7), pp. 1313-1319.

[56] Aoyama, Y. and Yagawa, G., 2001, "Large-Scale Eigenvalue Analysis of Structures Using Component Mode Synthesis", JSME International Journal, Series A: Solid Mechanics and Material Engineering, **44**(4), pp. 631-640.

[57] Apiwattanalunggarn, P., Shaw, S. W., and Pierre, C., 2005, "Component Mode Synthesis Using Nonlinear Normal Modes", Nonlinear Dynamics, **41**(1-3), pp. 17-46.

[58] Moon, J. D. and Cho, D. W., 1992, "A Component Mode Synthesis Applied to Mechanisms for an Investigation of Vibration", Journal of Sound and Vibration, **157**(1), pp. 67-79.

[59] Sung, S. H. and Nefske, D. J., 1986, "Component Mode Synthesis of a Vehicle Structural-Acoustic System Model", AIAA Journal, **24**(6), pp. 1021-1026.

[60] Verros, G. and Natsiavas, S., 2002, "Ride Dynamics of Nonlinear Vehicle Models Using Component Mode Synthesis", Journal of Vibration and Acoustics, Transactions of the ASME, **124**(3), pp. 427-434.

[61] Karpel, M., Moulin, B., and Feldgun, V., 2003, "Component Mode Synthesis of a Vehicle System Model Using the Fictitious Mass Method", *Proceedings of 44th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference, Apr 7-10 2003*, American Inst. Aeronautics and Astronautics Inc., **3**, pp. 1836-1845.

[62] Arnoldi, W. E., 1951, "The Principle of Minimized Iterations in the Solution of the Matrix Eigenvalue Problem", Quarterly of Applied Mathematics, **9**(1), pp. 17-29.

[63] Sorensen, D. C., 1992, "Implicit Application of Polynomial Filters in a K-Step Arnoldi Method", Siam Journal on Matrix Analysis and Applications, **13**(1), pp. 357-385.

[64] Lanczos, C., 1950, "An Iteration Method for the Solution of the Eigenvalue Problem of Linear Differential and Integral Operators", Journal of Research of the National Bureau of Standards, **45**(4), pp. 255-282.

[65] Gallivan, K., Grimme, E., and Van Dooren, P., 1994, "Pade Approximation of Large-Scale Dynamic Systems with Lanczos Methods", *Proceedings of 33rd IEEE Conference on Decision and Control. Part 1 (of 4), Dec 14-16 1994*, IEEE, Piscataway, NJ, USA, **1**, pp. 443-448.

[66] Ruhe, A., 1994, "Rational Krylov Algorithms for Nonsymmetric Eigenvalue Problems Ii: Matrix Pairs", Linear Algebra and Its Applications, **197**, pp. 283-295.

[67] Grimme, E. J., 1997, "Krylov Projection Methods for Model Reduction", Ph.D. Dissertation, University of Illinois, Urbana-Champaign.

[68] Alexandro, F. J., Jr., 1984, "Stable Partial Padé Approximations for Reduced-Order Transfer Functions", IEEE Transactions on Automatic Control, **AC-29**(2), pp. 159-162.

[69] Goldman, M. J., Porras, W. J., and Leondes, C. T., 1981, "Multivariable Systems Reduction Via Cauer Forms", International Journal of Control, **34**(4), pp. 623-650.

[70] Shieh, L.-S., Chang, F.-R., and Yates, R. E., 1986, "The Generalized Matrix Continued-Fraction Descriptions and Their Application to Model Simplification", International Journal of Systems Science, **17**(1), pp. 1-19.

[71] Chen, C. F., 1974, "Model Reduction of Multivariable Control Systems by Means of Matrix Continued Fractions", International Journal of Control, **20**(2), pp. 225-238.

[72] Chen, C. F. and Shieh, L. S., 1968, "A Novel Approach to Linear Model Simplification", International Journal of Control, **8**(6), pp. 561-570.

[73] Chen, C. F. and Shieh, L. S., 1969, "Continued Fraction Inversion by Routh's Algorithm", IEEE Transactions on Circuit Theory, **CT-16**(2), pp. 197-202.

[74] Hutton, M. F. and Friedland, B., 1975, "Routh Approximations for Reducing Order of Linear, Time-Invariant Systems", IEEE Transactions on Automatic Control, **AC-20**(3), pp. 329-337.

[75] Chen, T. C., Han, K. W., and Chang, C. Y., 1979, "Reduction of Transfer Functions by the Stability-Equation Method", Journal of the Franklin Institute, **308**(4), pp. 389-404.

[76] Shamash, Y., 1975, "Model Reduction Using the Routh Stability Criterion and the Pade Approximation Technique", International Journal of Control, **21**(3), pp. 475-484.

[77] Pal, J., 1979, "Stable Reduced-Order Padé Approximation Using the Routh-Hurwitz Array", Electronics Letters, **15**(8), pp. 225-226.

[78] Pal, J. and Ray, L. M., 1980, "Stable Padé Approximants to Multivariable Systems Using a Mixed Method", Proceedings of the IEEE, **68**(1), pp. 176-178.

[79] Chen, T. C., Chang, C. Y., and Han, K. W., 1980, "Model Reduction Using the Stability-Equation Method and the Padé Approximation Method", Journal of the Franklin Institute, **309**(6), pp. 473-490.

[80] Chen, T. C., Chang, C. Y., and Han, K. W., 1980, "Model Reduction Using the Stability-Equation Method and the Continued-Fraction Method", International Journal of Control, **32**(1), pp. 81-94.

[81] de Villemagne, C. and Skelton, R. E., 1987, "Model Reductions Using a Projection Formulation", International Journal of Control, **46**(6), pp. 2141-2169.

[82] Adamjan, V. M., Arov, D. V., and Krein, M. G., 1971, "Analytic Properties of Schmidt Pairs for a Hankel Operator and the Generalized Schmidt-Takagi Problem", Math. USSR Sbornik, **15**(1).

[83] Kung, S., 1978, "A New Identification and Model Reduction Algorithm Via Singular Value Decompositions", *Proceedings of Twelfth Asilomar Conference on Circuits, Systems and Computers, 6-8 Nov. 1978*, IEEE, pp. 705-714.

[84] Kung, S.-Y. and Lin, D. W., 1981, "Optimal Hankel-Norm Model Reductions: Multivariable Systems", IEEE Transactions on Automatic Control, **AC-26**(4), pp. 832-852.

[85] Glover, K., 1984, "All Optimal Hankel-Norm Approximations of Linear Multivariable Systems and Their L-Infinity Error Bounds", International Journal of Control, **39**(6), pp. 1115-1193.

[86] Safonov, M. G., Chiang, R. Y., and Limebeer, D. J. N., 1990, "Optimal Hankel Model Reduction for Nonminimal Systems", IEEE Transactions on Automatic Control, **35**(4), pp. 496-502.

[87] Ionescu, V. and Oara, C., 2001, "The Four-Block Adamjan-Arov-Krein Problem for Discrete-Time Systems", Linear Algebra and Its Applications, **328**(1-3), pp. 95-119.

[88] Gu, G., 2005, "All Optimal Hankel-Norm Approximations and Their L-Infinity Error Bounds in Discrete-Time", International Journal of Control, **78**(6), pp. 408-423.

[89] Ball, J. A. and Ran, A. C. M., 1987, "Optimal Hankel Norm Model Reductions and Wiener-Hopf Factorization. I: The Canonical Case", SIAM Journal on Control and Optimization, **25**(2), pp. 362-382.

[90] Ionescu, V. and Oara, C., 1996, "Class of Suboptimal and Optimal Solutions to the Discrete Nehari Problem: Characterization and Computation", International Journal of Control, **64**(3), pp. 483-509.

[91] Loève, M. M., 1955, *Probability Theory*, Princeton, N.J., VanNostrand.

[92] Karhunen, K., 1946, "Zur Spektraltheorie Stochastischer Prozesse", Annales Academiae Scientarum Fennicae, **37**.

[93] Hotelling, H., 1933, "Analysis of a Complex of Statistical Variables into Principal Components", Journal of Educational Psychology, **24**, pp. 417-441, 498-520.

[94] Lorenz, E. N., 1956, "Empirical Orthogonal Functions and Statistical Weather Prediction", 1, MIT, Department of Meteorology, Cambridge.

[95] Lumley, J. L., "The Structure of Inhomogeneous Turbulent Flow", in *Atmospheric Turbulence and Radio Wave Propagation*, A. M. Yaglom and V. I. Tatarski, Eds. Nauka, Moscow, 1967, pp. 166-178.

[96] Golub, G. H. and Van Loan, C. F., 1983, *Matrix Computations*, North Oxford Academic, Oxford.

[97] Sirovich, L., 1987, "Turbulence and the Dynamics of Coherent Structures. I. Coherent Structures", Quarterly of Applied Mathematics, **45**(3), pp. 561-570.

[98] Sirovich, L., 1987, "Turbulence and the Dynamics of Coherent Structures. Ii. Symmetries and Transformations", Quarterly of Applied Mathematics, **45**(3), pp. 573-582.

[99] Sirovich, L., 1987, "Turbulence and the Dynamics of Coherent Structures. Iii. Dynamics and Scaling", Quarterly of Applied Mathematics, **45**(3), pp. 583-590.

[100] Brooks, C. L., Karplus, M., and Pettitt, B. M., 1988, *Proteins: A Theoretical Perspective of Dynamics, Structure and Thermodynamics*, Wiley, New York.

[101] Berkooz, G., Holmes, P., and Lumley, J. L., 1993, "Proper Orthogonal, Decomposition in the Analysis of Turbulent Flows", Annual Review of Fluid Mechanics, **25**, pp. 537.

[102] Wu, G. G., Liang, Y. C., Lin, W. Z., Lee, H. P., and Lim, S. P., 2003, "A Note on Equivalence of Proper Orthogonal Decomposition Methods", Journal of Sound and Vibration, **265**(5), pp. 1103-1110.

[103] Moore, B. C., 1981, "Principal Component Analysis in Linear Systems: Controllability, Observability, and Model Reduction", IEEE Transactions on Automatic Control, **26**(1), pp. 17-32.

[104] Kalman, R. E., 1965, "Irreducible Realizations and Degree of a Rational Matrix", Journal of the Society for Industrial and Applied Mathematics, **13**(2), pp. 520-544.

[105] Kabamba, P. T., 1985, "Balanced Gains and Their Significance for L2 Model Reduction", IEEE Transactions on Automatic Control, **AC-30**(7), pp. 690-693.

[106] Penzl, T., 1999, "Cyclic Low-Rank Smith Method for Large Sparse Lyapunov Equations", SIAM Journal of Scientific Computing, **21**(4), pp. 1401-1418.

[107] Sorensen, D. C. and Antoulas, A. C., 2002, "The Sylvester Equation and Approximate Balanced Reduction", Linear Algebra and Its Applications, **351-352**, pp. 671-700.

[108] Gugercin, S., Sorensen, D. C., and Antoulas, A. C., 2003, "A Modified Low-Rank Smith Method for Large-Scale Lyapunov Equations", Numerical Algorithms, **32**(1), pp. 27-55.

[109] Desai, U. B. and Pal, D., 1984, "A Transformation Approach to Stochastic Model Reduction", IEEE Transactions on Automatic Control, **AC-29**(12), pp. 1097-1100.

[110] Green, M., 1988, "A Relative Error Bound for Balanced Stochastic Truncation", IEEE Transactions on Automatic Control, **33**(10), pp. 961-965.

[111] Green, M., 1988, "Balanced Stochastic Realizations", Linear Algebra and Its Applications, **98**, pp. 211-247.

[112] Opdenacker, P. C. and Jonckheere, E. A., 1988, "A Contraction Mapping Preserving Balanced Reduction Scheme and Its Infinity Norm Error Bounds", IEEE Transactions on Circuits and Systems, **35**(2), pp. 184-189.

[113] Jonckheere, E. A. and Silverman, L. M., 1983, "A New Set of Invariants for Linear Systems - Application to Reduced Order Compensator Design", IEEE Transactions on Automatic Control, **AC-28**(10), pp. 953-964.

[114] Enns, D. F., 1984, "Model Reduction with Balanced Realizations: An Error Bound and a Frequency Weighted Generalization", *Proceedings of 23rd IEEE Conference on Decision and Control*, IEEE, New York, NY, USA, pp. 127-132.

[115] Lin, C.-A. and Chiu, T.-Y., 1992, "Model Reduction Via Frequency Weighted Balanced Realization", Control Theory and Advanced Technology, **8**(2), pp. 341-351.

[116] Sreeram, V. and Anderson, B. D. O., 1995, "Frequency Weighted Balanced Reduction Technique: A Generalization and an Error Bound", *Proceedings of 34th IEEE Conference on Decision and Control. Part 4 (of 4), Dec 13-15 1995*, IEEE, Piscataway, NJ, USA, **4**, pp. 3576-3581.

[117] Wang, G., Sreeram, V., and Liu, W. Q., 1999, "New Frequency-Weighted Balanced Truncation Method and an Error Bound", IEEE Transactions on Automatic Control, **44**(9), pp. 1734-1737.

[118] Zhou, K., 1993, "Frequency-Weighted Model Reduction with L Infinity Error Bound", Systems & Control Letters, **21**(2), pp. 115-125.

[119] Hahn, J. and Edgar, T. F., 2002, "Balancing Approach to Minimal Realization and Model Reduction of Stable Nonlinear Systems", Industrial and Engineering Chemistry Research, **41**(9), pp. 2204-2212.

[120] Scherpen, J. M. A., 1993, "Balancing for Nonlinear Systems", Systems & Control Letters, **21**(2), pp. 143-153.

[121] Hahn, J. and Edgar, T. F., 2002, "An Improved Method for Nonlinear Model Reduction Using Balancing of Empirical Gramians", Computers and Chemical Engineering, **26**(10), pp. 1379-1397.

[122] Verriest, E. I. and Gray, W. S., 2001, "Nonlinear Balanced Realizations", *Proceedings of 40th IEEE Conference on Decision and Control*, Institute of Electrical and Electronics Engineers Inc., **4**, pp. 3250-3251.

[123] Ma, X. and De Abreu-Garcia, J. A., 1988, "On the Computation of Reduced Order Models of Nonlinear Systems Using Balancing Technique", *Proceedings of 27th IEEE Conference on Decision and Control*, IEEE, pp. 1165-1166.

[124] Skelton, R. E., 1980, "Cost Decomposition of Linear Systems with Application to Model Reduction", International Journal of Control, **32**(6), pp. 1031-1055.

[125] Skelton, R. E. and Yousuff, A., 1983, "Component Cost Analysis of Large Scale Systems", International Journal of Control, **37**(2), pp. 285-304.

[126] Skelton, R. E. and Kabamba, P., 1986, "Comments on 'Balanced Gains and Their Significance for L2 Model Reduction' by P. T. Kabamba", IEEE Transactions on Automatic Control, **AC-31**(8), pp. 796-797.

[127] Skelton, R. E., 1988, *Dynamic Systems Control: Linear Systems Analysis and Synthesis*, Wiley, New York.

[128] Skelton, R. E., 1989, "The Explicit Relation between Component Cost Analysis and Balanced Coordinates", *Proceedings of 28th IEEE Conference on Decision and Control*, IEEE, Piscataway, NJ, USA, **2**, pp. 1326-1330.

[129] Hyland, D. C. and Bernstein, D. S., 1985, "The Optimal Projection Equations for Model Reduction and the Relationships among the

Methods of Wilson, Skelton, and Moore", IEEE Transactions on Automatic Control, **AC-30**(12), pp. 1201-1211.

[130] Wilson, D. A., 1970, "Optimum Solution of Model-Reduction Problem", *Proceedings of Institution of Electrical Engineers*, **117**, pp. 1161-1165.

[131] Howitt, G. D. and Luus, R., 1990, "Model Reduction by Minimization of Integral Square Error Performance Indexes", Journal of the Franklin Institute-Engineering and Applied Mathematics, **327**(3), pp. 343-357.

[132] Luus, R., 1980, "Optimization in Model Reduction", International Journal of Control, **32**(5), pp. 741-747.

[133] Gouda, M. M., Danaher, S., and Underwood, C. P., 2002, "Building Thermal Model Reduction Using Nonlinear Constrained Optimization", Building and Environment, **37**(12), pp. 1255-1265.

[134] Hachtel, G., Sangiovanni-Vincentelli, A., and Visvanathan, V., 1981, "An Optimization-Based Approach to Model Simplification", *Proceedings of IEEE International Symposium on Circuits and Systems*, IEEE, Piscataway, NJ, **3**, pp. 995-1000.

[135] Assunção, E. and Peres, P. L. D., 1999, "A Global Optimization Approach for the H2-Norm Model Reduction Problem", *Proceedings of 38th IEEE Conference on Decision and Control*, IEEE, Piscataway, NJ, USA, **2**, pp. 1857-1862.

[136] Chen, H.-F. and Fang, H.-T., 2002, "Nonconvex Stochastic Optimization for Model Reduction", Journal of Global Optimization, **23**(3-4), pp. 359-372.

[137] Spanos, J. T., Milman, M. H., and Mingori, D. L., 1992, "A New Algorithm for L2 Optimal Model Reduction", Automatica, **28**(5), pp. 897-909.

[138] Ferrante, A., Krajewski, W., Lepschy, A., and Viaro, U., 1999, "Convergent Algorithm for L2 Model Reduction", Automatica, **35**(1), pp. 75-79.

[139] Ungar, E. E., 1997, "Statistical Energy Analysis", Journal of Sound and Vibration, **31**(10), pp. 28-32.

[140] Rosenberg, R. C. and Zhou, T., 1988, "Power-Based Simplification of Dynamic System Models", *Proceedings of Advances in Design Automation, Sep 25-28 1988*, American Soc of Mechanical Engineers (ASME), New York, NY, USA, **14**, pp. 487-492.

[141] Rosenberg, R. C. and Zhou, T., 1988, "Power-Based Model Insight", *Proceedings of Automated Modeling for Design, Nov 27-Dec 2 1988*, American Soc of Mechanical Engineers (ASME), New York, NY, USA, **8**, pp. 61-67.

[142] Louca, L. S., Stein, J. L., Hulbert, G. M., and Sprague, J., 1997, "Proper Model Generation: An Energy-Based Methodology", *Proceedings of 1997 International Conference on Bond Graph Modeling*, SCS, **29**, pp. 44-49.

[143] Louca, L. S. and Stein, J. L., 1998, "Physical Interpretation of Reduced Bond Graphs", *Proceedings of 2nd IMACS International Multiconference: Computational Engineering in Systems and Applications (CESA'98)*.

[144] Louca, L. S., Stein, J. L., and Hulbert, G. M., 1998, "A Physical-Based Model Reduction Metric with an Application to Vehicle Dynamics", *Proceedings of 4th IFAC Nonlinear Control Systems Design Symposium (NOLCOS 98)*, **3**, pp. 607-612.

[145] Louca, L. S. and Yildir, U. B., 2003, "Modeling and Reduction Techniques for Studies of Integrated Hybrid Vehicle Systems", *Proceedings of 4th International Symposium on Mathematical Modeling*.

[146] Rideout, D. G., Stein, J. L., and Louca, L. S., 2004, "Systematic Model Decoupling through Assessment of Power-Conserving Constraints - an Engine Dynamics Case Study", *Proceedings of 2004 ASME International Mechanical Engineering Congress and Exposition*, American Society of Mechanical Engineers, New York, NY, **2**.

[147] Fathy, H. K. and Stein, J. L., 2005, "Fundamental Concordances between Balanced Truncation and Activity-Based Model Reduction", *Proceedings of IMAACA '05, Bond Graph Methods for Dynamical Systems*, A. Bruzzone, G. Dauphin-Tanguy, C. Frydman, and S. Junco, eds., pp. 109-116.

[148] Rideout, D. G., Stein, J. L., and Louca, L. S., 2004, "System Partitioning and Physical- Domain Model Reduction through Assessment of Bond Graph Junction Structure", *Proceedings of IMAACA '04, Bond Graph Techniques for Modeling Dynamic Systems*, SCS.