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
ls louca@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

\[
\dot{x} = Ax + Bu
\]

\[
y = Cx + Du
\]

with the \( r \)-dimensional reduced model

\[
\dot{x}_r = A_r x_r + B_r u
\]

\[
y = C_r x_r + Du
\]

Suppose the reduced state vector \( x_r \) is related to the original state vector \( x \) through

\[
x_r = K x
\]

where \( K \) is the \( r \times n \) aggregation matrix. It follows that

\[
A_r K = KA
\]

\[
B_r = KB
\]

\[
C_r K \approx C
\]

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

\[
A_r = K A K^\dagger
\]

\[
B_r = KB
\]

\[
C_r = CK^\dagger
\]

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

\[
K = T \left[I_r \ 0\right] V^{-1}
\]

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 \( \varepsilon \) as a coefficient, i.e.,

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

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

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

\[
0 = f_2(\bar{x}_1, \bar{x}_2, \bar{u})
\]

where bars are used to distinguish between this limiting case and the case where \( \varepsilon \) 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.,

\[
\bar{x}_2 = \phi(\bar{x}_1, \bar{u})
\]

Substituting this solution into Eq. (7) effectively furnishes a slow submodel that residualizes the fast states, i.e.,
\[
\bar{x}_i = f_i(\bar{x}, \phi(\bar{x}, \bar{u}), \bar{u}) = \bar{f}(\bar{x}, \bar{u})
\] (11)

The reduced model for the fast dynamics can be obtained by introducing a fast time scale \( \tau \) and fast variables \( \bar{x}_f(\tau) \) and \( \bar{x}_s(\tau) \) defined as follows:

\[
\tau = \frac{t}{\epsilon}, \quad x_j(t) = \bar{x}_f(t) + \bar{x}_j(\tau), \quad j = 1, 2 \] (12)

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

\[
\frac{d\bar{x}_f}{d\tau} = f_2(x(t), \bar{x}_f(t) + \bar{x}_f(\tau), u)
\] (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:

\[
\|G - G_r\| \leq 2(\sigma_{n+1} + \cdots + \sigma_{n+m})
\] (14)

where \( \sigma_i, i = n+1, \ldots, n+m \) 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 \( \epsilon \), and becomes zero in the limit as \( \epsilon \) 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]

\[ M\ddot{x} + Kx = 0 \]  

(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

\[ K\nu = \omega^2 M\nu \]

(16)

where the natural frequencies are given by the various solutions for \( \omega \) and the modes shapes are given by the corresponding solutions for \( \nu \). 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 matrix

\[ A = \begin{bmatrix} 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_n \end{bmatrix} \]

(17)

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].

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. It 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

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

(19)

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

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

(20)

such that for a desired number $n \in \mathbb{N}_0$, the equalities $a_k = \hat{a}_k$, $k = 0, 1, 2, ..., n$, are referred to as moments. The coefficients $a_k, \hat{a}_k$, $k = 0, 1, 2, ...$, 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

$$G(s) = \frac{a_2 s^2 + \cdots + a_{2n}s^{2n-1}}{a_1 s + a_2 s^2 + \cdots + a_{n+1}s^n}$$

(21)

can be written in the following continuous fraction expansion form

$$G(s) = \frac{1}{h_1 + \frac{1}{\frac{1}{h_2} + \frac{1}{\frac{1}{h_3 + \frac{1}{\frac{1}{h_4 + \cdots}}}}}}$$

(22)

with

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

(23)

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

\[
\begin{align*}
 a_{11} & a_{12} & a_{13} & \cdots \\
 a_{21} & a_{22} & a_{23} & \cdots \\
 a_{31} & a_{32} & \cdots & j = 3, \ldots, n+1; \quad k = 1, 2, \ldots \\
 a_{41} & \cdots & \cdots & \cdots \\
 \end{align*}
\]

(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 2r 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

$$G(s) = \frac{b_1 s + b_2 s^2 + \cdots + b_{2n}s^{2n-1}}{a_1 s + a_2 s^2 + \cdots + a_{n+1}s^n}$$

(25)

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

$$G(s) = \beta_1 f_1(s) + \beta_2 f_1(s)f_2(s) + \cdots + \beta_{n+1} f_1(s)f_2(s) \cdots f_{n}(s)$$

(26)

where
\[ f_1 = \frac{1}{1+\alpha_1 s} \]

\[ f_i(s) = \frac{1}{\alpha_i s + \frac{1}{\alpha_i+1 s + \frac{1}{\alpha_i+2 s + \frac{1}{\ddots \frac{1}{\alpha_{n-i} s + \frac{1}{\alpha_{n-i+1} s + \frac{1}{\alpha_{n-i+2} s + \frac{1}{\ddots \frac{1}{\alpha_{n} s}}}}}}}} }{i = 2, \ldots, n} \tag{27} \]

and the coefficients \(\alpha_i\) and \(\beta_i\) are given by

\[ \alpha_i = \frac{a_{i,1}}{a_{i,2}}, \quad i = 1, \ldots, n \tag{28} \]

\[ a_{i,j} = a_{i-1,j+1} \quad j \text{ odd}, \quad i = 2, \ldots, n \]

\[ a_{i,j} = a_{i-1,j+1} - \alpha_{i-1} a_{i-1,j+2} \quad j \text{ even}, \quad i = 2, \ldots, n \]

\[ \beta_i = \frac{b_{i,1}}{a_{i,2}}, \quad i = 1, \ldots, n \tag{29} \]

\[ b_{i,j} = b_{i-1,j+1} \quad j \text{ odd} \]

\[ b_{i,j} = b_{i-1,j+1} - \beta_{i-1} a_{i-1,j+2} \quad j \text{ even}, \quad i = 2, \ldots, n \]

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

\[ \hat{G}_r(s) = \beta_1 p_1(s) + \beta_r p_r(s) \]

\[ p_1(s) = f_1(s) \]

\[ p_2(s) = \frac{1}{\alpha_2 s + \frac{1}{\ddots \frac{1}{\alpha_r s + \frac{1}{\alpha_{r-1} s + \frac{1}{\alpha_{r-2} s + \frac{1}{\ddots \frac{1}{\alpha_1 s}}}}}}}} }{i = 2, \ldots, r} \tag{31} \]

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

\[ \dot{x} = Ax + Bu \]

\[ y = Cx + Du \tag{32} \]

into the reduced model

\[ \dot{x}_r = A_r x_r + B_r u \]

\[ y = C_r x_r + D_r \tag{33} \]

by \(A_r = LA_L, \quad B_r = LB_L, \quad C_r = CT_L, \quad \text{and} \quad LT = I\). 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

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

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

\[ M_p(0) = CA^i B, \quad i = 1, \ldots, p \tag{37} \]

\(q\) high frequency moments

\[ M_p(\infty) = CA^i B, \quad i = 0, \ldots, q - 1 \tag{38} \]
p low frequency power moments

\[ R_i(0) = CA^{-i}X(A^T)^{-i}C^T, \quad i = 1, \ldots, p \]

and, q high frequency power moments

\[ R_i(\infty) = CA^iX(A^T)^iC^T, \quad i = 0, \ldots, q - 1 \]

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 approximation to a system’s entire frequency response. Such an approximate solution. If, instead, one uses the Hankel norm of the error \( \|G - \hat{G}\| \), and \( \|G - \hat{G}\| \) is lower-bounded by \( \sum_{k=1}^{n} \sigma_k(G) \), and the equality in the error bound is satisfied only by the optimal Hankel norm approximation of the system \( \hat{G} \).

This minimization of error in terms of the Hankel norm comes at the expense of a change in realizability due to 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]

\[ \|G(j\omega) - \hat{G}(j\omega)\|_\infty \leq 2 \sum_{i=k+1}^{n} \sigma_i \]  

(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 \( \hat{D} \) in such a way that upper-bound on the \( \mathcal{H}_\infty \) norm of the error is cut in half, i.e.,

\[ \|G(j\omega) - \hat{G}(j\omega) - \hat{D}\|_\infty \leq \sum_{i=k+1}^{n} \sigma_i \]

(42)

Please see [85] for the calculation of such a \( \hat{D} \).

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-Loève 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-Loève expansion method can be applied as follows. Consider a system represented by a state space equation of the form

\[ \dot{x} = f(x,u), \quad x \in \mathbb{R}^n \]

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

\[ A = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}_{mn \times n} \]

Obtain the singular value decomposition of the matrix \( A \), i.e.,

\[ A = U \Sigma V^T \]  \hspace{1cm} (45)

where \( \Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)_{mn \times n} \) with \( \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq 0 \).

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.,

\[ \dot{x}_r = (V_k)^T f(V_k x_r, u), \quad \bar{x} = V_k x_r \]  \hspace{1cm} (46)

where \( \bar{x} \) 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 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 \( \Sigma_k \) denotes the leading \( k \times k \) principal minor of the matrix \( \Sigma \). 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

\[ \|A - A_r\|_F = \sum_{i=k+1}^{n} \sigma_i^2 \]  \hspace{1cm} (47)

where \( \| \|_F \) 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 \( \|x - \bar{x}\| \). 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

\[ z(x,t) \approx \sum_{i=1}^{M} a_i(x) h_i(t) \]

In this case the observation matrix can be arranged as:

\[ A = \begin{bmatrix} z(x_1,t_1) & z(x_2,t_1) & \cdots & z(x_n,t_1) \\ \vdots & \vdots & \ddots & \vdots \\ z(x_1,t_m) & z(x_2,t_m) & \cdots & z(x_n,t_m) \end{bmatrix}_{mn \times n} \]  \hspace{1cm} (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 \( \Sigma \) 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-Loève 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-Loève 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-Loève 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-Loève 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 [112]. 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.,

\[
P = Q = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n),
\]

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 \( \tilde{\mathcal{D}} \) matrix is not optimized, i.e.,

\[
\|G - G_r\|_\infty \leq 2 \sum_{i=k+1}^{\infty} \sigma_i \tag{50}
\]

where \( \sigma_i \) are the Hankel singular values of \( G \) corresponding to the truncated states. Note, however, that in Hankel norm approximation \( \tilde{\mathcal{D}} \) 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

\[
\begin{align*}
\dot{x} &= Ax + Bu, & x \in \mathbb{R}^n, u \in \mathbb{R}^m \\
y &= Cx, & y \in \mathbb{R}^p
\end{align*} \tag{51}
\]

is defined as

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

This cost function satisfies the cost decomposition property
\[ V = \sum_{i=1}^{n} V_i \]  

(53)

where \( V_i \) is the contribution of the \( i^{th} \) state, \( x_i \), to the system cost, and is given by

\[ V_i = \left[ XC^T C \right]_i \]  

(54)

where \( X \) is the controllability Grammian, satisfying

\[ XA^T + AX + BB^T = 0 \]  

(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 \( \Delta V_i \) 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

\[ x = Tz \]

\[ T = T_0 U; \quad X = T_0^T \bar{X} T_0; \quad T_0^T C^T C T_0 = U \left[ \begin{array}{cc} \Sigma & 0 \\ 0 & 0 \end{array} \right] U^T \]  

(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

\[ y' = Cx + \bar{D}u' \]

\[ y'(q-1) \]

\[ = \hat{C}x + \hat{D}u' \]

and considers the system

\[ \dot{x} = Ax + Bu \]

\[ y' = \hat{C}x + \hat{D}u' \]  

(58)

with the cost function

\[ V = \sum_{k=1}^{m} \int_{0}^{\infty} \bar{y}^T(k,t)Q\bar{y}(k,t)dt \]  

(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

\[ \hat{T} = T_0 U \left[ \begin{array}{cc} \Sigma^{-1/4} & 0 \\ 0 & I \end{array} \right] \]  

(60)

yields the balanced coordinates, if

\[ q = n, \quad Q = \int_{0}^{\infty} \left[ \begin{array}{c} \alpha_0(t) \\ \vdots \\ \alpha_{q-1}(t) \end{array} \right] I_p \left[ \begin{array}{ccc} \alpha_0(t)I & \cdots & \alpha_{q-1}(t)I \end{array} \right] dt \]

(61)

where \( e^{\Delta t} = \sum_{i=0}^{n-1} \alpha_i(t) A^i \). 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 \( \mathcal{H}_\infty \)-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 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

\[
A_i = \int_0^T |f_i(t)|e_i(t)\,dt
\]  \hspace{1cm} (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
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


