Outstanding Issues in Model Order Reduction*

João M. S. Silva[1], Jorge Fernández Villena[1], Paulo Flores[1], and L. Miguel Silveira[1,2]

[1] INESC ID / Instituto Superior Técnico
Technical University of Lisbon
Rua Alves Redol, 9
1000-029 Lisboa, Portugal
{jmss,jorge,lms}@algos.inesc-id.pt
[2] Cadence Laboratories
Cadence Design Systems

Summary. With roots dating back to many years ago and applications in a wide variety of areas, model order reduction has emerged in the last few decades as a crucial step in the simulation, control, and optimization of complex physical systems. Reducing the order or dimension of models of such systems, is paramount to enabling their simulation and verification. While much progress has been achieved in the last few years regarding the robustness, efficiency and applicability of these techniques, certain problems of relevance still pose difficulties or renewed challenges that are not satisfactorily solved with the existing approaches. Furthermore, new applications for which dimension reduction is crucial, are becoming increasingly relevant, raising new issues in the quest for increased performance.

Keywords—Model order reduction, massively coupled systems, orthogonal projection, parametric systems, circuit simulation.

1 Introduction

Model reduction algorithms are standard techniques nowadays in many areas, including the microelectronics design community. The goal of model order reduction is to replace a large-scale model of a physical system by a model of lower dimension which exhibits similar behavior, typically measured in terms of its frequency or time response characteristics. Such techniques are commonly used for analysis, approximation, and simulation of models arising from electromagnetic formulation of physical structures. The need to accurately account for all relevant physical effects implies that the mathematical formulation used to describe such structures often results in very large models. Reducing the order or dimension of these models is crucial to enabling the simulation and verification of such systems [2, 1].
An area to which extensive research has been devoted in the last few years is the problem of order reduction of nonlinear systems [20, 18, 4]. A discussion of such methods is however beyond the scope of this paper. Due to space constraints we will restrict the discussion to issues arising from linear systems reduction. Nevertheless this discussion is still relevant in the nonlinear case as most existing nonlinear reduction algorithms are based on extensions of linear methods or the solution of carefully selected sequences of linear problems. While enormous progress has been achieved in the last decades in this field, both from a theoretical as

* Invited Paper at SCEE-2006

well as a practical standpoint, still greater challenges lie ahead as new and exciting applications are being researched for which order reduction is again a crucial step.

Existing methods for linear model reduction can be broadly characterized into two types: those that are based on projection methods, and those based on balancing techniques (sometimes also referred to as SVD[3]-based [1]). Among the first, Krylov subspace projection methods such as PVL [6] and PRIMA [15] have been the most widely studied over the past decade. They are very appealing because of their simplicity and performance in terms of efficiency and accuracy, despite the fact that they exhibit several known shortcomings. The lack of a general strategy for error control and order selection, as well as a dependence on the original model's structure if passivity is to be guaranteed after the reduction are among the more obvious such shortcomings. The alternative methods, those in the truncated balanced realization (TBR) family [14], perform reduction based on the concept of controllability and observability of the system states and are purported to produce nearly optimal models and have easy to compute *a-posteriori* error bounds. However, they are awkward to implement and expensive to apply, which limits their applicability to small and medium sized problems. Hybrid techniques that combine some of the features of each type of methods have also been presented [11, 9, 10]. Recently, a new technique was also proposed that attempts to establish a bridge between the two techniques. The Poor Man's TBR [19] is based on a projection scheme where the projection matrix approximately spans the dominant eigenspaces of the controllability and observability matrices and provides an interesting platform for bridging between the two types of techniques. Still the technique is not without drawbacks, as it relies on proper choice of sampling points, a non-trivial task in general.

In spite of their shortcomings, all of the mentioned methods are in widespread use nowadays. Still, there are situations that challenge the existing knowledge in the field. For instance, consider the problem of reducing systems with a large number of ports, also known as massively coupled systems. Such systems typically occur in substrate, power grid and package parasitic networks. Furthermore, the trend to nano-scale dimensions together with the increasing frequencies of operation implies that non-neglectable electromagnetic effects have to be accounted for in the models, which will also give rise to these massively coupled problems. Projection-based algorithms are inefficient for such systems as they rely on block iterations, where the size of the block equals the number of ports. Therefore, each block iteration increases the size of the model by an amount equal to the number of ports, leading to large models even for moderate reduction order. This trend is particularly troublesome when simulation with such models is necessary. TBR is intrinsically somewhat less sensitive to the number of input ports. Unfortunately such systems are typically very large, which makes reduction based on balancing techniques impractical.

Additionally, new challenges are being posed that require further research. As an example, consider the problem of order reduction of parametrized systems. Parameter-based descriptions are now starting to be used as the basis for variability-aware design models. For high frequencies, at nano-scale feature sizes, process variability effects, as well as dependence on operating conditions become extremely relevant and should be accounted for in the models. Existing techniques for handling such systems are, for the most part, straightforward extensions of the basic order reduction algorithms [3, 12]. Projection-based techniques match Taylor-series coefficients, which in parameter-based descriptions are multidimensional moments. Unfortunately this technique has exponential cost increase with the number of parameters and is thus expensive except for small size and small number of parameters. Building a projection space assuming small perturbations around the nominal operating point is also problematic: it is hard to do anything beyond first-order and thus it is not clear how to dial in accuracy. Sampling the parameter space also presents a challenge, as it is not clear where to place sample point in such a multidimensional space. Still if some information regarding the statistical distribution of the parameter values is available, this can be used to guide the sampling and to build the model accordingly.

[3] SVD – Singular value decomposition.

In this paper we review some of these current and future challenges for which much research is still needed in model order reduction. In Section 2 we discuss the problem of reducing massively coupled problems, and in Section 3 we discuss the reduction of parametrized systems, a recent topic of much research work. Finally in Section 4, we present some conclusions.

2 Massively Coupled Systems

As an illustration of the problems pertaining to massively coupled systems, results from the study of the reduction of power distribution networks, also known as power grids, will be presented. Power grids are fairly regular structures which must cover the whole area of the chip for power delivery purposes. Since all devices, wells and substrate plugs, are connected to the power grid, the total number of ports of such circuits can be as high as hundreds of thousands, or millions. This unfortunately brings added difficulty to the reduction process.

2.1 Background

Modeling a power grid as an RC network and using the nodal analysis formulation leads to:

$$\begin{aligned} C\dot{v} + Gv &= Mu \\ y &= N^T v \end{aligned} \tag{1}$$

where $C, G \in \mathbb{R}^{n\times n}$ are the capacitance and conductance matrices, respectively, $M \in \mathbb{R}^{n\times p}$ is a matrix that relates the inputs $u \in \mathbb{R}^p$ to the states $v \in \mathbb{R}^n$ that describe the node voltages, $N \in \mathbb{R}^{n\times q}$ being its counterpart with respect to the outputs $y \in \mathbb{R}^q$, n is the number of states, p the number of inputs and q the number of outputs. The $p \times q$ matrix transfer function of the network is then given by $H(s) = N^T(G + sC)^{-1}M$. Typically, matrices C and G are very sparse but also very large. For a typical power grid, the number of nodes will be in the order of several millions but the number of ports, input and output, is also quite large. Solving Eqn. (1) directly or using it inside a circuit simulator is therefore too expensive.
The goal of model-order reduction is, generically, to determine a reduced model,

$$H_k(s) = \hat{N}^T(\hat{G} + s\hat{C})^{-1}\hat{M} \tag{2}$$

of size $k \ll n$, that closely matches the input-output behavior of the original model, and where the state description is given by $z = V^T v \in \mathbb{R}^k$. However, even if $k \ll n$, the reduced-order model may fail to provide relevant compression. This may happen because, for large networks, the matrices C and G are sparse, having a number of non-zeros entries of order $\mathcal{O}(n)$. If the number of non-zero entries in the reduced-order model increases with the number of ports, the benefits of reduction may vanish with increasingly large p and q.

Projection-based framework

Projection-based Krylov subspace algorithms, such as PRIMA [15], provide a general-purpose, rigorous framework for deriving interconnect modeling algorithms and have been shown to produce excellent compression in many scenarios involving on- and off-chip interconnect and packaging structures. In its simplest form, they can be used to compute individual approximations to each of the $p \times q$ matrix transfer function entries. However, more commonly, they are used to generate a single approximation to the full system transfer function. The PRIMA algorithm [15], for instance, reduces a state-space model in the form of (1) by use of a projection matrix V, through the operations:

$$\hat{G} = V^T G V, \quad \hat{M} = V^T M, \quad \hat{C} = V^T C V, \quad \hat{N} = V^T N \tag{3}$$

to obtain a reduced model in the form of (2). In the standard approach, the projection matrix V is chosen as an orthogonal basis of a block Krylov subspace, $\mathcal{K}_m(A, b) = span\{b, Ab, \ldots, A^{m-1}b\}$, a typical choice being $A = G^{-1}C$ and $b = G^{-1}M$. The construction of the projection matrix V is done iteratively by blocks, with each block being generated through a back-orthogonalizing procedure. When the projection matrix is constructed in this way, the moments of the reduced model can be shown to match the moments of the original model to some order. Consequently, the reduced model size is proportional to the number of matched moments multiplied by the number of ports. Furthermore, the reduced system matrices will be dense. Therefore, these methods present two problems when dealing with networks with a large number of ports. First, the cost associated with model computation is directly proportional to the number of inputs, p, i.e. to the number of columns in the matrices defining the inputs. This is easy to see by noting that the number of columns in the projection matrix V in (3) is directly proportional to p (a direct result of the block construction procedure described). This implies that model construction for systems with large number of ports is costly. Second, the size of the reduced model is also proportional to p, as was discussed earlier and can directly be seen from (3). While the cost of model construction can perhaps be amortized in later simulations, the large size of the model is more problematic since it directly affects simulation cost.

Truncated balanced realizations

An alternative class of reduction algorithms are based on Truncated Balanced Realization (TBR). The TBR algorithm first computes the observability and controllability Gramians, X and Y, by solving the Lyapunov equations:

$$GXC^T + CXG^T = MM^T, \tag{4}$$

$$G^TYC + C^TYG = N^TN \tag{5}$$

and then reduces the model by projection onto the space associated with the dominant eigenvalues of the product XY [14]. Model size selection and error control in TBR is based on the eigenvalues of XY, also known as the Hankel singular values. In the proper case, there is an *a-posteriori* theoretical bound on the frequency-domain error for the TBR model given by [14]:

$$\|H - H_k\| \leq 2 \sum_{i=k+1}^{n} \sigma_i \tag{6}$$

The existence of such an error bound is an important advantage of the TBR class of algorithms as there is no counterpart in the projection-based algorithms. Theoretically, the model selection criteria, and therefore the size of the generated model, can be done independently of the number of inputs. However, there is an indirect dependence in most problems and in particular for networks such as power grids, that exhibit a large number of inputs (see [19] for additional discussion on the topic). In this case, useful reductions are not achievable. Furthermore, the solution of the Lyapunov equations required to obtain X and Y is computationally intensive for large systems and as such the technique is only of theoretical interest in this context. A variety of approximate methods have been proposed that attempt to circumvent this problem (see [19] and references therein).

2.2 Methods

As stated previously, the difficulty with standard projection algorithms like PRIMA or multipoint projection schemes, is that the models produced have size proportional to the number of ports. This limits their applicability to problems such as power grids, where the number of network ports is likely to be very large. An interesting question that might be raised is whether this restriction is inherent to the system, given the number of ports, or an artifact of

the computation scheme chosen. In other words, one might ask whether accurate modeling and analysis of a power grid, modeled as a large RC mesh, does indeed require so much dynamic information. This question is all the more relevant as there is a common popular belief that only a few poles are required to accurately model an RC circuit. It is now widely accepted that in certain settings that is indeed the case, but this conclusion is emphatically not general (see [22]).
In the following, two recently proposed methods for overcoming the difficulties faced by standard MOR methods are presented. The first method is based on the analysis of singular values of the system moments while the second one is a "cheaper" version of a TBR class method previously mentioned [19], also based on projection.

Singular Value Decomposition MOR (SVDMOR)

The SVDMOR [5] algorithm was developed to address the reduction of systems with a large number of ports, like power grids. While the size of a reduced model produced via PRIMA is directly proportional to the number of ports in the circuit, SVDMOR theoretically overcomes this problem using singular value decomposition (SVD) analysis in order to truncate the system to any desired order.
The main idea behind SVDMOR is to assume that there is a large degree of correlation between the various inputs and outputs. SVDMOR further assumes that such input-output correlation can be captured from observation of structural system properties, evidenced in matrices M and N. The method can, for instance, use an input-output correlation matrix, like the one given by the zero order moment matrix $S_{DC} = N^T G^{-1} M$, which contains only DC information. Alternatively, more complicated response correlations can be used such as frequency, s_j-shifted moments, $S_{DC}^{(s_j)} = N^T (G + s_j C)^{-1} M$, a more generic k-order moment, $S_k = N^T (G^{-1} C)^k G^{-1} M$, or even combinations of these. Let K be the appropriate correlation matrix. If the basic correlation hypothesis holds true, then K can be approximated by a low-rank matrix. This low rank property can be revealed by computing the SVD of K, $K = U \Sigma W^T$, where U and W are orthogonal matrices and Σ is the diagonal matrix containing the ordered singular values. Assuming correlation, there will be only a small number, $m \ll p + q$, of dominant singular values. Therefore, we can approximate $K \approx U_m \Sigma_m V_m^T$, where truncation is performed keeping the m most significant singular values. The method further approximates:

$$\begin{aligned} M &\approx b_M V_m^T = M V_m (V_m^T V_m)^{-1} V_m^T \\ N &\approx b_N U_m^T = N U_m (U_m^T U_m)^{-1} U_m^T \end{aligned} \tag{7}$$

where b_M and b_N are obtained using the Moore-Penrose pseudo-inverse, resulting in:

$$H(s) \approx U_m \underbrace{b_N^T (G + sC)^{-1} b_M}_{H_m(s)} V_m^T \tag{8}$$

Standard MOR methods, like PVL or PRIMA, can now be applied to $H_m(s)$, leading to $\tilde{H}_m(s)$, an r-th order model, from which a final model approximation $H(s) \approx H_r(s) = U_m \tilde{H}_m(s) V_m^T$ is computed. The reduced system is $p \times q$ with a number of nonzero elements of order $\mathcal{O}(r^2)$.

Input-Correlated Poor Man's TBR (PMTBR)

The PMTBR algorithm [19, 22] was motivated by a connection between frequency-domain projection methods and approximation to truncated balanced realization. The method is less expensive in terms of computation, but tends to TBR when the order of the approximation increases. The actual mechanics of the algorithm are akin to multi-point projection. In a multi-point rational approximation the projection matrix columns are computed by sampling at several frequency points along a desired frequency interval. The samples are given by

$z_i = (G + s_i C)^{-1} M$, where $s_i = j\omega_i$ (with $i = 1, 2, \ldots, P$) are P frequency sample points. The frequency-sampled matrix thus obtained can then be used to project the original system in order to obtain a reduced model. In the PMTBR algorithm, a similar procedure is used. The connection to TBR methods is made by noting that an approximation $\hat{X}$ to the Gramian X can be can be computed as:

$$\hat{X} = \sum_i w_i z_i z_i^H \tag{9}$$

where the ω_i which defines each sample, and the w_i can be interpreted as nodes and weights of a quadrature scheme applied to a frequency-domain interpretation of the Gramian matrix (see [19] for details). Let Z be a matrix whose columns are the z_i, and W the diagonal matrix of the square root of the weights. Eqn. (9) can be written more compactly as:

$$\hat{X} = ZW^2 Z^H \tag{10}$$

If the quadrature rule applied is accurate, $\hat{X}$ will converge to X, which implies the dominant eigenspace of $\hat{X}$ converges to the dominant eigenspace of X. Computing the singular value decomposition of ZW, $ZW = V_Z S_Z U_Z$ (with S_Z real diagonal, and V_Z, U_Z unitary matrices), it is easy to see that V_Z converges to the eigenspaces of X, and the Hankel singular values are obtained directly from the entries of S_Z. V_Z can then be used as the projection matrix in a model order reduction scheme. The method was shown to perform quite well in a wide variety of settings [19].

An interesting additional interpretation was more recently presented [22] which is of relevance in our context. It has been shown that if further information revealing time-domain correlation between the ports is available, a variant of PMTBR can be used that can lead to significant efficiency improvement. This idea is akin to the basic assumptions in SVDMOR and relate to exploiting correlation between the inputs. Unlike SVDMOR, however, it is assumed that the correlation information is not contained in the circuit information directly, but rather in its inputs. In this variant of PMTBR, a correlation matrix K is formed by columns which are samples of port values along the time-steps of some interval. Those samples should characterize as well as possible the values expected at the inputs of the system, i.e. K should be a suitably representative model of the possible inputs. An SVD is then performed over K in order to retain only the r most significant components of the input correlation information, $K \approx U_r \Sigma_r V_r^T$. With this additional correlation information, the samples relative to multipoint approximation become $z_i = (G + s_i C)^{-1} M U_r \Sigma_r$. Using these z_i as columns of the Z matrix in (10), leads to the input-correlated TBR algorithm (ICTBR). See [19] for more details and a more thorough description of the probabilistic interpretation of both PMTBR as well as ICTBR.

2.3 Results

Both the standard model order reduction as well as the methods described in the previous section can be applied to massively coupled systems. Methods like SVDMOR are reported to provide significant advantages over the standard algorithms if certain conditions are met, namely that significant port correlation exists and can be ascertained in a practical way. PMTBR is a more general algorithm for model reduction, which can nonetheless be applied to large systems, given its reduced computational complexity.

In this section, results are presented for two types of topologies: a first mesh, grid A, with voltage inputs on the left side and current outputs on the right one, and a second mesh, grid B, with voltage ports along the left side and current ports randomly distributed over the remaining nodes. For practical reasons, we have kept the mesh sizes smaller than they would be in realistic applications but scaling of all appropriate dimensions and sizes would produce qualitatively the same results. There are two main differences between the two setups described. The first one concerns formulation. While in grid A matrices M and N in Eqn. (1) are distinct (M yields input information and N yields output information), in grid B, $M = N$, thus all

ports are controllable and observable. The second main difference consists in the separation between ports. In grid A the separation between inputs and outputs is maximal, while in grid B not only every port is both input and output, but also the geometric proximity between ports is reduced. Grid A is thus expected to be fairly compressible, but smaller reductions are expected for grid B. Grid A is similar to the one used in [5], while grid B was created in order to illustrate a more realistic setup. The electrical model of all grids is as follows: every connection between nodes is purely resistive and at every node there is a capacitance to ground. While this is not necessary, it simplifies the ensuing description (furthermore, a parasitic capacitance is usually extracted at all nodes). Resistance and capacitance values were randomly generated in the interval $(0.9, 1.1)$. In the following set of experiments the size of the reduced model is the same for all methods and was pre-determined. The correlation matrix of SVDMOR is the DC moment matrix. For this method, after computing the SVD and choosing how many singular values to keep, a number of PRIMA iterations is performed in order to generate a model of the required size. The number of frequency samples of PMTBR was set such that a model of the same size can be drawn from matrix Z. Samples were chosen uniformly in the frequency range shown in the plots, with an additional sample added at DC.

Highly-correlated ports

The previously discussed methods were first used to reduce grid A. The Bode plot of an arbitrarily selected transfer function is presented in Figure 1 (left). The number of retained states was forced at $r = 1200$. In the case of SVDMOR, 15 singular values were kept and 80 PRIMA iterations were run, yielding the reduced model of $15 \times 80 = 1200$ states. One observes that SVDMOR shows good results, better than PRIMA and PMTBR. In order to understand the reason for these results the plot of the singular values of SVDMOR and PMTBR methods is presented in Figure 1 (right). The singular values (s.v.) of the DC moment, used by SVDMOR to guide the reduction, decay quite fast. Therefore keeping just the first 15 yields a good approximation. On the other hand the PMTBR s.v. decay very slowly. Table 1 shows the

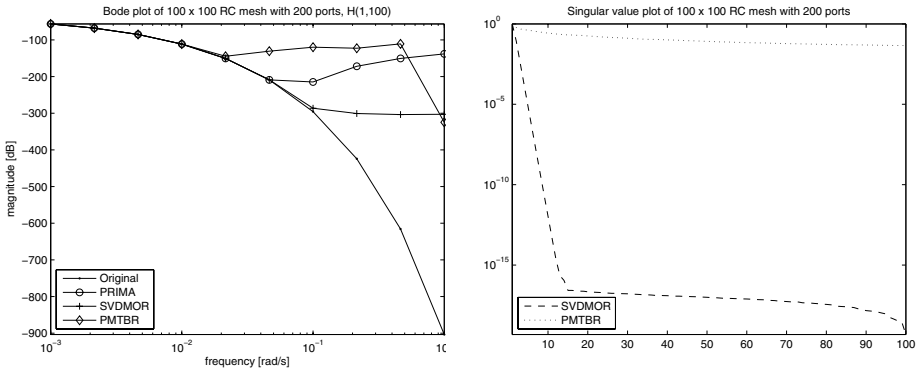

Fig. 1: Results for grid A ($r = 1200$): Bode plot of arbitrarily selected entry of 100×100 transfer function matrix (left); normalized plot of singular values: SVDMOR moment matrix and PMTBR samples matrix (right).

maximum absolute error of the transfer matrix, $max\{|H(s) - H_r(s)|\}$. Analysis of the table indicates that in the overall model, SVDMOR shows the smallest error as expected for this grid setup.

Table 1: Maximum absolute error of $|H(s) - H_r(s)|$ for 100×100 mesh with 100 inputs on the left side and 100 outputs on the right side. SVDMOR used 15 singular values.

$r = 1200$	PRIMA	SVDMOR	PMTBR
$max\{\|H - H_r\|\}$	1.443×10^{-6}	1.406×10^{-7}	1.160×10^{-5}

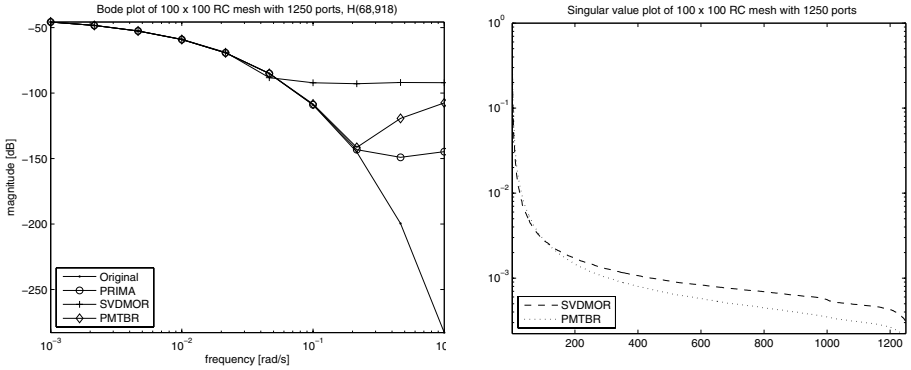

Fig. 2: Results for grid B ($r = 2500$): Bode plot of arbitrarily selected entry of 100×100 transfer function matrix (left); normalized plot of singular values: SVDMOR moment matrix and PMTBR samples matrix (right).

Table 2: Maximum absolute error of $|H - H_r|$ for 100×100 mesh with 100 ports on the left side and 1150 randomly distributed ports over the mesh.

$r = 2500$	PRIMA	SVDMOR	PMTBR
$max\{\|H - H_r\|\}$	$1.284e \times 10^{-2}$	2.533×10^{-1}	1.545×10^{-3}

Weakly-correlated ports

In grid B the objective was to emulate a more realistic situation whereby potentially many devices, modeled as current sources, are attached to the power grid and can draw or sink current from/to it when switching. The number of current sources was chosen to be $1/8$ of the number of nodes. There are 1150 current sources and 100 voltage sources (for a total of 10000 nodes). This is a harder problem to reduce, due to port proximity, and thus interaction, and the results show it. Again the Bode plot of an arbitrarily selected transfer function is presented in Figure 2 (left). The number of retained states was now forced at $r = 2500$ already showing smaller reduction than for grid A. In this case, the approximation produced by SVDMOR is less accurate. This is expected from inspection of Figure 2 (right), where one observes that the s.v. of SVDMOR decay slower than in the previous case. Clearly, the assumption of highly correlated ports is not valid here. The results concerning the error of the transfer matrix are in Table 2. PMTBR produces the most accurate model, while PRIMA shows a reasonable approximation.

Note that while the Bode plots show large errors for higher (normalized) frequencies, concerning to higher order moments which are harder to match, these frequencies are uninteresting in practical simulations. Note also that the matrices in the reduced models for all methods in both experiences are full, which has drastic consequences for usage of these models in a simulation environment.

3 Parametrized System Descriptions

In any manufacturing process there is always a certain degree of uncertainty involved given our limited control over the environment and other physical conditions. For the most part this uncertainty was previously ignored when analyzing or simulating systems, but as we step towards the nano-scale and higher frequency eras, such environmental, geometrical and electromagnetic fluctuations become more significant. Nowadays, parameter variability can no longer be disregarded, and its effect must be accounted for in early design stages so that unwanted consequences can be minimized. This leads to parametric descriptions of systems, including the effects of the manufacturing variability, which further increases the complexity of such models. When model reduction is required, these parametric representations must be addressed and the resulting reduced models must retain the ability to model the effects of small random fluctuations, in order to accurately predict behavior and optimize designs. This is the aim of the Parametric Model Order Reduction (pMOR).

3.1 Background

Actual fabrication of physical devices is prone to the variation of certain circuit parameters due to deliberate adjustment of the process or from random deviations inherent to this manufacturing. This variability leads to a dependence of the extracted circuit elements on several parameters, of electrical or geometrical origin. This dependence results in a parametric state-space system representation, which in descriptor form can be written as

$$\begin{aligned} &C(\lambda_1,\ldots,\lambda_L)\dot{v}(\lambda_1,\ldots,\lambda_L) + G(\lambda_1,\ldots,\lambda_L)v(\lambda_1,\ldots,\lambda_L) = Mu \\ &y = N^T v(\lambda_1,\ldots,\lambda_L) \end{aligned} \tag{11}$$

where $C, G \in \mathbb{R}^{n\times n}$ are again, respectively, the capacitance and conductance matrices, $M \in \mathbb{R}^{n\times p}$ is the matrix that relates the input vector $u \in \mathbb{R}^p$ to the inner states $v \in \mathbb{R}^n$ and $N \in \mathbb{R}^{n\times q}$ is the matrix that links those inner states to the outputs $y \in \mathbb{R}^q$. The elements of the matrices C and G, as well as the states of the system v, depend on a set of L parameters $\lambda = [\lambda_1, \lambda_2, \ldots, \lambda_L]$ which model the effects of the mentioned uncertainty. Usually the system is formulated so that the matrices related to the inputs and outputs (M and N) do not depend on the parameters. This time-domain descriptor yields a parametric dependent frequency response modeled via the transfer function

$$H(s,\lambda_1,\ldots,\lambda_L) = N^T(sC(\lambda_1,\ldots,\lambda_L) + G(\lambda_1,\ldots,\lambda_L))^{-1}M \tag{12}$$

for which we seek to generate a reduced order approximation, able to accurately capture the input-output behavior of the system for any point in the parameter space.

$$\hat{H}(s,\lambda_1,\ldots,\lambda_L) = \hat{N}^T(s\hat{C}(\lambda_1,\ldots,\lambda_L) + \hat{G}(\lambda_1,\ldots,\lambda_L))^{-1}\hat{M} \tag{13}$$

In general, one attempts to generate a reduced order model whose structure is, as much as possible, similar to the original, i.e. exhibiting a similar parametric dependence.

3.2 Methods

In the following we summarize the main methods presented for dealing with this problem.

Perturbation-Based Techniques

One of the earliest attempts to address this variational issue was to combine perturbation theory with moment matching MOR algorithms [13]. To model the variational effects of the

interconnects, an affine model can be built for the capacitance and conductance matrices, so that

$$\begin{aligned} G(\lambda_1, \ldots, \lambda_L) &= G_0 + \lambda_1 G_1 + \ldots + \lambda_L G_L \\ C(\lambda_1, \ldots, \lambda_L) &= C_0 + \lambda_1 C_1 + \ldots + \lambda_L C_L \end{aligned} \tag{14}$$

where now C_0 and G_0 are the nominal matrix values, i.e. the value of the matrices under no parameter variation, and C_i and G_i, $i = 1, \cdots, L$, are its sensitivities with respect to those parameters. For small parameter variations, the projection matrix obtained via a moment-matching type algorithm such as PRIMA also suffers small perturbations. Therefore, the idea was to draw several samples in the parameter space for the system matrices $G(\lambda_1, \ldots, \lambda_L)$ and $C(\lambda_1, \ldots, \lambda_L)$, and for each sample PRIMA was applied so a projection matrix is obtained. Fitting is later applied over all the computed projectors in order to determine the coefficients of a parameter dependent projection matrix

$$V(\lambda_1, \ldots, \lambda_L) = V_0 + \lambda_1 V_1 + \ldots + \lambda_L V_L \tag{15}$$

which is in turn applied in a congruence-like transformation to the parametric system in (11), yielding a reduced system parametrized with respect to the set $[\lambda_1, \lambda_2, \ldots, \lambda_L]$.
Another approach also based on perturbation theory arguments was applied to the Truncate Balanced Realization (TBR) [14, 17] framework, so that a theoretically based perturbation matrix was obtained starting from the affine models shown in (14) [8]. This matrix was then applied via a congruence transformation over the Gramians to address the variability, and yield the perturbed Gramians. These in turn were used inside a balancing truncation procedure. As with most TBR-inspired methods, this one is also expensive to compute and hard to implement.
The above methods have obvious drawbacks, perhaps the most glaring of which is the heavy computation cost required for obtaining the reduced models and the limitation that comes from first order approximations possibly leading to inaccuracy in certain cases.

Multi-Dimensional Moment Matching

These techniques appear as extensions to nominal moment-matching techniques [15, 6, 21]. Moment matching algorithms have gained a well deserved fame in nominal MOR due to their simplicity and efficiency. The extensions of these techniques to the parametric case are usually based in the implicit or explicit moment matching of the parametric transfer function (12).
This type of algorithms assumes small fluctuations of the parameters, so that a model based on the Taylor Series expansion can be used for approximating the behavior of the conductance and capacitance, $G(\lambda)$ and $C(\lambda)$, expressed as a function of the parameters

$$\begin{aligned} G(\lambda_1, \ldots, \lambda_L) &= \textstyle\sum_{i_1=0}^{\infty} \cdots \sum_{i_L=0}^{\infty} G_{i_1,\ldots,i_L} \lambda_1^{i_1} \ldots \lambda_L^{i_L} \\ C(\lambda_1, \ldots, \lambda_L) &= \textstyle\sum_{i_1=0}^{\infty} \cdots \sum_{i_L=0}^{\infty} C_{i_1,\ldots,i_L} \lambda_1^{i_1} \ldots \lambda_L^{i_L} \end{aligned} \tag{16}$$

where $G_0, C_0, G_{i_1,\ldots,i_L}$ and $C_{i_1,\ldots,i_L}$ are the multidimensional Taylor series coefficients. This Taylor series can be extended up to the desired (or required) order, including cross derivatives, for the sake of accuracy. If this formulation is used, the structure for parameter dependence may be maintained if the projection is not only applied to the nominal matrices, but to the sensitivities as well.
The Multi-Parameter Moment Matching method is a single-point expansion of the transfer function (12) in the joint space of the frequency s and the parameters $\lambda_i, i = 1, \cdots, L$, in order to obtain a power series in several variables $s, \lambda_1, \ldots, \lambda_L$ [3],

$$v(s, \lambda_1, \ldots, \lambda_L) = \sum_{k=0}^{\infty} \sum_{k_s=0}^{k} \sum_{k_1=0}^{k-k_s} \cdots \sum_{k_L=0}^{k-k_s-k_1\ldots.-k_{L-1}} M_{k,k_s,k_1,\ldots,k_L} s^{k_s} \lambda_1^{k_1} \ldots \lambda_L^{k_L} \tag{17}$$

where $M_{k,k_s,k_1,\ldots,k_L}$ is a k-th ($k = k_s + k_1 + \ldots + k_L$) order multi-parameter moment corresponding to the coefficient term $s^{k_s}\lambda_1^{k_1}\ldots\lambda_L^{k_L}$. Following the same idea used in the nominal moment matching techniques, a basis for the subspace formed from these moments can be built and the resulting matrix V can be used as a projection matrix for reducing the original system. It has been shown that this parametrized reduced model matches up to the k-th order multi-parameter moment of the original system. The main inefficiency of this method is that process parameters fluctuate in a small range around their nominal value, whereas the frequency range is much larger, and a higher number of moments are necessary in order to capture the global response for the whole frequency range. For this reason, the reduced model size grows exponentially with the number of parameters and the moments to match. A similar idea but more efficient, is to rely in a two-step moment matching scheme [12]. In this method, one first matches in an explicit way the multi-parameter moments for the process variability parameters (by expanding the state space vector v and the matrices G and C in its Taylor Series only w.r.t. the parameters), and in a second stage implicitly match moments with respect to the frequency via Krylov projection. This two-step approach avoids the exponential growth of model size with the number of moments matched, suffered by the multi-parameter moment matching. This method allows a certain degree of flexibility as the number of moments matched with respect to the frequency and to the parameters can be different. In principle, in spite of the larger size of the augmented model, the order of the reduced system can be much smaller than in the previous cases. On the other hand, the structure of the dependence with respect to the parameters is lost since the parametric dependence is shifted to the later projected output related N matrix.
A different multi-dimensional moment matching approach was also presented [7], which relies on the computation of several subspaces, built separately for each dimension, i.e. the frequency s and the parameter set λ. So given a parametric system (11), the first step of the algorithm is to obtain the k_s block moments of the transfer function with respect to the frequency when the parameters take their nominal value (for example, via PRIMA). This block moments will be denoted as Q_s. The next step is to obtain the subspaces which match k_{λ_i} block moments of v with respect to each of the parameter λ_i, and will be denoted by Q_{λ_i}. Once all the subspaces have been computed, an orthonormal basis can be obtained so that its columns spans the joint of all subspaces. Applying the resulting matrix in a projection scheme ensures that the parametric ROM[4] matches k_s moments of the original system with respect to the frequency, and k_{λ_i} moments with respect to the parameter λ_i. If the cross-term moments are needed for accuracy reasons, the subspace that spans these moments can be also included by following the same scheme.

Variational PMTBR

A novel approach was recently proposed that extends the PMTBR algorithm to include variability [16]. This approach is based on the statistical interpretation of the algorithm (see [19] for details) and enhances its applicability. In this interpretation, the Gramian is seen as a covariance matrix for a Gaussian variable, $v(0)$, obtained by exciting the (presumed stable) system with white noise. Rewriting the Gramian as

$$X_\lambda = \int_{S_\lambda}\int_{-\infty}^{\infty}(sC_\lambda + G_\lambda)^{-1}MM^T(sC_\lambda + G_\lambda)^{-H}p(\lambda)dwd\lambda \tag{18}$$

where $p(\lambda)$ is the probability density of λ in the parameter space, S_λ. Just as in PMTBR, a quadrature rule can be applied in the overall parameter plus frequency space to approximate the Gramian via numerical computation. But in this case the weights are chosen taking into account the PDF[5] of λ_i and the frequency constraints. This can be generalized to a set of

[4] Reduced Order Model
[5] PDF – Probability density function.

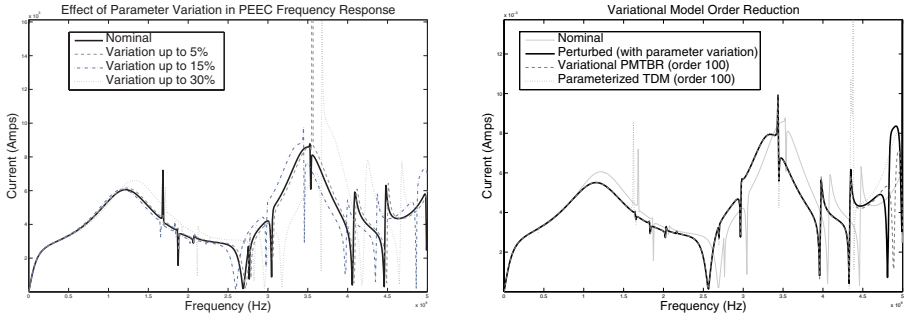

Fig. 3: Variational PEEC: effects on the frequency response (left) and performance of parametric MOR methods (right).

parameters, where a joint PDF of all the parameters can be applied to the joint parameter space, or the individual PDF of each parameter can be used. The ability to do this represents an interesting advantage, since *a-priori* knowledge of the parameters and the frequency can be included in order to constrain the sampling and yield a more accurate reduced model. As in the deterministic case, an error analysis and control can be included, via the eigenvalues of the SVD, but in this variational case only an expected error bound can be given:

$$E\{\|\hat{v}_0 - v_0\|_2^2\} \leq \sum_{i=r+1}^{n} \sigma_i^2 \tag{19}$$

where r is the reduced order and n the original number of states. In this method, the issue of sample selection, already an important one in the deterministic version, becomes even more relevant, since the sampling must now be done in a potentially much higher-dimensional space.

3.3 Results

To illustrate (for a qualitative analysis mostly) the effect of parameter variability on the response of a circuit we resort to a simple example of a partial equivalent electric circuit (PEEC) model. The system under analysis is an RLC model of a connector of order 304. In this example we consider the effect of five geometric parameters, each having a different effect on the conductance and capacitance matrices. Figure 3(left) shows the effect of random variations on each parameter up to a limit of $5\%, 15\%$ and 30%. It can be seen that even small range variations in the parameters can result in large deviations from nominal. An important effect of the parameter variation is that those deviations not only can change the overall shape of the frequency response but also cause frequency shifts in the pole location. Figure 3(right) shows a comparison of the reduction of the variational system with two different methods: variational PMTBR and parametrized time-domain macromodels [7], all of the same order, versus the nominal response and the system response under parameter variation (Perturbed). As can be seen, the parametric MOR algorithms are able to maintain an acceptable accuracy up to high frequencies in the presence of strong variations.

4 Conclusions

Model order reduction is a crucial enabling technique for simulation, control, and optimization of complex physical systems. In this paper we discussed how, in spite of the progress achieved in the area in the last few years, certain types of problems such as those derived

from massively coupled systems, still pose difficulties to the existing approaches. We also discussed new challenges in the field, brought by new applications such as the reduction of parametric systems, that are becoming increasingly relevant, raising new issues in the quest for increased performance. Clearly, we have but scratched the surface of the relevant issues facing us. Other challenging problems exist, like the reduction on nonlinear systems, which has also been subject to extensive research.

Acknowledgements

Jorge Fernández Villena, Paulo Flores and L. Miguel Silveira acknowledge the financial support from the FP6/IST/027378 Chameleon-RF project (`http://www.chameleon-rf.org`). João M. S. Silva was supported by a PhD fellowship from Fundação para a Ciência e Tecnologia (FCT), Portugal, with the reference SFRH/BD/10586/2002.

References

1. A. C. Antoulas. *Approximation of Large-Scale Dynamical Systems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2005.
2. P. Benner, V. Mehrmann, and D. Sorensen, editors. *Dimension Reduction of Large-Scale Systems*, volume 45 of *Lecture Notes in Computational Science and Engineering*. Springer-Verlag, Berlin/Heidelberg, Germany, 2005.
3. L. Daniel, O. C. Siong, S. C. Low, K. H. Lee, and J. K. White. A multiparameter moment-matching model-reduction approach for generating geometrically parametrized interconnect performance models. *IEEE Trans. Computer-Aided Design*, 23:678–693, May 2004.
4. N. Dong and J. Roychowdhury. Piecewise polynomial nonlinear model reduction. In 40^{th} *ACM/IEEE Design Automation Conference*, pages 484–489, Anaheim, CA, June 2003.
5. P. Feldmann. Model order reduction techniques for linear systems with large number of terminals. In *DATE'2004 - Design, Automation and Test in Europe, Exhibition and Conference*, volume 2, pages 944–947, Paris, France, February 2004.
6. P. Feldmann and R. W. Freund. Efficient linear circuit analysis by Padé approximation via the Lanczos process. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 14(5):639–649, May 1995.
7. P. Gunupudi, R. Khazaka, M. Nakhla, T. Smy, and D. Celo. Passive parameterized time-domain macromodels for high-speed transmission-line networks. *IEEE Trans. On Microwave Theory and Techniques*, 51(12):2347–2354, December 2003.
8. P. Heydari and M. Pedram. Model reduction of variable-geometry interconnects using variational spectrally-weighted balanced truncation. In *International Conference on Computer Aided-Design*, pages 586–591, San Jose, CA, USA, November 2001.
9. I. M. Jaimoukha and E. M. Kasenally. Krylov subspace methods for solving large Lyapunov equations. *SIAM Journal on Numerical Analysis*, 31:227–251, 1994.
10. M. Kamon, F. Wang, and J. White. Generating nearly optimally compact models from Krylov-subspace based reduced-order models. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 47(4):239–248, April 2000.
11. J.-R. Li, F. Wang, and J. White. Efficient model reduction of interconnect via approximate system grammians. In *International Conference on Computer Aided-Design*, pages 380–383, San Jose, CA, November 1999.
12. X. Li, P. Li, and L. Pileggi. Parameterized interconnect order reduction with Explicit-and-Implicit multi-Parameter moment matching for Inter/Intra-Die variations. In *International Conference on Computer Aided-Design*, pages 806–812, San Jose, CA, November 2005.
13. Y. Liu, L. T. Pileggi, and A. J. Strojwas. Model order reduction of RC(L) interconnect including variational analysis. In 36^{th} *ACM/IEEE Design Automation Conference*, pages 201–206, June 1999.

14. B. Moore. Principal Component Analysis in Linear Systems: Controllability, Observability, and Model Reduction. *IEEE Transactions on Automatic Control*, AC-26(1):17–32, February 1981.
15. A. Odabasioglu, M. Celik, and L. T. Pileggi. PRIMA: passive reduced-order interconnect macromodeling algorithm. *IEEE Trans. Computer-Aided Design*, 17(8):645–654, August 1998.
16. J. Phillips. Variational interconnect analysis via PMTBR. In *International Conference on Computer Aided-Design*, pages 872–879, San Jose, CA, USA, November 2004.
17. J. Phillips, L. Daniel, and L. M. Silveira. Guaranteed passive balancing transformations for model order reduction. *IEEE Trans. Computer-Aided Design*, 22(8):1027–1041, August 2003.
18. J. R. Phillips. Projection-based approaches for model reduction of weakly nonlinear, time-varying systems. *IEEE Trans. Computer-Aided Design*, 22:171–187, 2003.
19. J. R. Phillips and L. M. Silveira. Poor Man's TBR: A simple model reduction scheme. *IEEE Trans. Computer-Aided Design*, 24(1):43–55, Jan. 2005.
20. M. Rewienski and J. White. A trajectory piecewise-linear approach to model order reduction and fast simulation of nonlinear circuits and micromachined devices. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 22(2):155–170, Feb. 2003.
21. L. M. Silveira, M. Kamon, I. Elfadel, and J. K. White. A coordinate-transformed Arnoldi algorithm for generating guaranteed stable reduced-order models of rlc circuits. In *International Conference on Computer Aided-Design*, pages 288–294, San Jose, California, November 1996.
22. L. M. Silveira and J. Phillips. Exploiting input information in a model reduction algorithm for massively coupled parasitic networks. In 41^{st} *ACM/IEEE Design Automation Conference*, pages 385–388, San Diego, CA, USA, June 2004.