A generative modeling framework for statistical link analysis based on sparse data

A Generative Modeling Framework for Statistical Link Analysis Based on Sparse Data

Simon De Ridder, *Student Member, IEEE,* Paolo Manfredi, *Member, IEEE,* Jan De Geest, *Member, IEEE,* Dirk Deschrijver, *Senior Member, IEEE,* Daniël De Zutter, *Fellow, IEEE,* Tom Dhaene, *Senior Member, IEEE,* Dries Vande Ginste, *Senior Member, IEEE.*

***Abstract*—This paper proposes a novel strategy for creating generative models of stochastic link responses starting from limited available data. Whereas state-of-the-art techniques, e.g. based on generalized polynomial chaos expansions, require a considerable amount of (expensive) input data, here we start from a small set of "training" responses. These responses are obtained either from simulations or measurements, to construct a comprehensive stochastic model. Using this model, new response samples can be generated with a distribution as similar as possible to the real data distribution, for use in Monte Carlo-like analyses. The methodology first uses the standard Vector Fitting algorithm to fit the S-parameter data with rational functions having common poles. Then, a generative model for the residues is created by means of Principal Component Analysis and Kernel Density Estimation. An a-posteriori selection of passive samples is performed on the generated data to ensure the new samples are physically consistent. The proposed modeling approach is applied to a commercial connector and to a set of differential striplines. Both are concatenated to produce the stochastic analysis of a complete link. Comparisons on the prediction of time-domain responses are also provided.**

***Index Terms*—Statistical link analysis, stochastic modeling, Principal Component Analysis (PCA), Kernel Density Estimate (KDE), high-speed connectors and links**

I. Introduction

The large manufacturing tolerances in modern electronics have recently prompted a wide interest in stochastic modeling techniques that can accurately predict the effects of component variability during the early design phase. To this end, a large number of statistical samples needs to be generated within a reasonable amount of time. While the Monte Carlo method is recognized to be in most cases prohibitive due to the high computational cost involved, alternative techniques were proposed based on the theoretical framework of the generalized polynomial chaos (gPC). These techniques allow collecting statistical information and/or generating a large number of samples at a smaller cost [1–7].

This work was funded by the Research Foundation Flanders (FWO-Vlaanderen) and the Interuniversity Attraction Poles Programme BESTCOM initiated by the Belgian Science Policy Office. P. Manfredi is currently an FWO Postdoctoral Research Fellow.

S. De Ridder, P. Manfredi, T. Dhaene, D. Deschrijver, D. De Zutter and D. Vande Ginste are with the IBCN/Electromagnetics Group of ID-Lab, Department of Information Technology, Ghent University/imec, 9000 Gent, Belgium (e-mail: simon.deridder@ugent.be, paolo.manfredi@ugent.be, tom.dhaene@ugent.be, dirk.deschrijver@ugent.be, daniel.dezutter@ugent.be, dries.vandeginste@ugent.be).

J. De Geest is with Amphenol-FCI. (e-mail: jan.degeest@fci.com)

Although gPC was demonstrated to be effective in a number of practical applications, a limitation is that it requires sampling the stochastic problem at specific points of the design space. This is at times hard to achieve, for example in the presence of a very large number of varying parameters or when these parameters or their distributions are unknown. In case only measured data can be obtained, efficient and sufficient sampling may even become completely intractable.

In this paper, an alternative, novel and conceptually different black-box approach is put forward to address the problem of generating a large number of samples of the response of a stochastic linear and passive multiport device, starting from a limited set of actual responses. The latter can be obtained either by simulating the original device at a subset of points in the random space (possibly unknown to the model developer) or by measuring several manufactured devices. In this way, a model is constructed that does not depend on, or require the knowledge of, various input parameters or their distributions.

The proposed approach starts by collecting a small amount of S-parameter data (or any other frequency-domain representation), which is then transformed into a pole-residue form by means of the Vector Fitting (VF) algorithm [8–10]. Then, the distribution of the residue matrices is modeled by means of Principal Component Analysis (PCA) [11–13] and Kernel Density Estimation (KDE) [14–16]. This model enables generating new sample values for the residue matrices, yielding corresponding S-parameter responses via the VF model representation. It is essential that the new, generated samples are in good agreement with the 'real data', representing the underlying distribution of an arbitrarily large set of simulated or measured data of which the training samples are a subset. Moreover, particular care is taken in making sure that the generated samples still preserve all physical constraints of stability, causality, passivity, reciprocity, etc., as well as the existing relationships among the different S-parameters in a multiport device. The generated S-parameter samples are readily imported in any SPICE-type simulator to perform, e.g., statistical time-domain link simulations. The proposed technique can be applied to model a complete interconnect link, or possibly to different sub-blocks that are later combined together to form a complex link, thus allowing a Monte-Carlo-like analysis with the desired level of modularity.

It is worth mentioning that this work is based on the idea proposed in [17], where a generative model for the simple case of a pair of coupled microstrips was constructed. In this paper, however, the modeling framework is more carefully

explained and generalized to arbitrary multiports. Moreover, the advocated technique is applied to a commercial connector footprint, an interconnect structure composed of four pairs of differential striplines, and a complete link consisting of two connector footprints and the striplines. Additionally, time-domain results are presented, validating the entire modeling framework and illustrating the appositeness of the technique for signal integrity aware design.

The rest of the paper is organized as follows. In Section II the problem is stated. The new modeling approach is outlined in Section III. A number of application examples and validations are provided in Section IV. Finally, conclusions are drawn in Section V.

II. Goal Statement

The goal of this paper is as follows. A limited number of responses of a stochastic device, typically S-parameters and hereafter called the "training set", is available from simulated or measured data. These responses differ from each other as the result of some (often unknown or ill-defined) source of variability. The samples are therefore understood to be different realizations of the same stochastic process.

From the training set, a generative statistical model is derived that can generate new S-parameter samples whose statistical properties resemble as closely as possible those of the original data. Moreover, the generated samples must preserve physical constraints (e.g., passivity) which constitute the relationships that exist among the different S-parameters of a multiport structure. For example, in a passive 2-port structure, $|S_{11}|^2 + |S_{21}|^2 \leq 1$, where the equality holds for a lossless device. Therefore, the independent statistical modeling of S_{11} and S_{21} would yield unrealistic datasets. This crucial step was not addressed by previous approaches [18, 19], as each parameter was modeled independently and no discussion about physical consistency was provided.

To validate the proposed modeling, a large number of samples is obtained via numerical simulations. Only a small subset is then used to build the generative model. Finally, the new samples generated by the model are statistically compared against the full dataset.

III. Generative Modeling Framework

A flowchart of the new modeling approach is shown in Fig. 1 and involves three main steps: VF, PCA and KDE. Each step is discussed in the following sections.

A. Training Set

The training set consists of a limited set of distinct random samples of the device response. It can be generated via simulation, by varying the random parameters, or by measuring different fabricated samples of a given device. It is even possible to consider several measurements of the same sample, e.g. to model measurement uncertainty and/or poor reproducibility. The size of the training set plays of course an important role. However, as demonstrated in Section IV, accurate statistical models can be obtained with a few tens of training samples. In what follows, we use the term 'sample' to refer to a single realization of the complete matrix characterization of the variable device in the frequency domain.

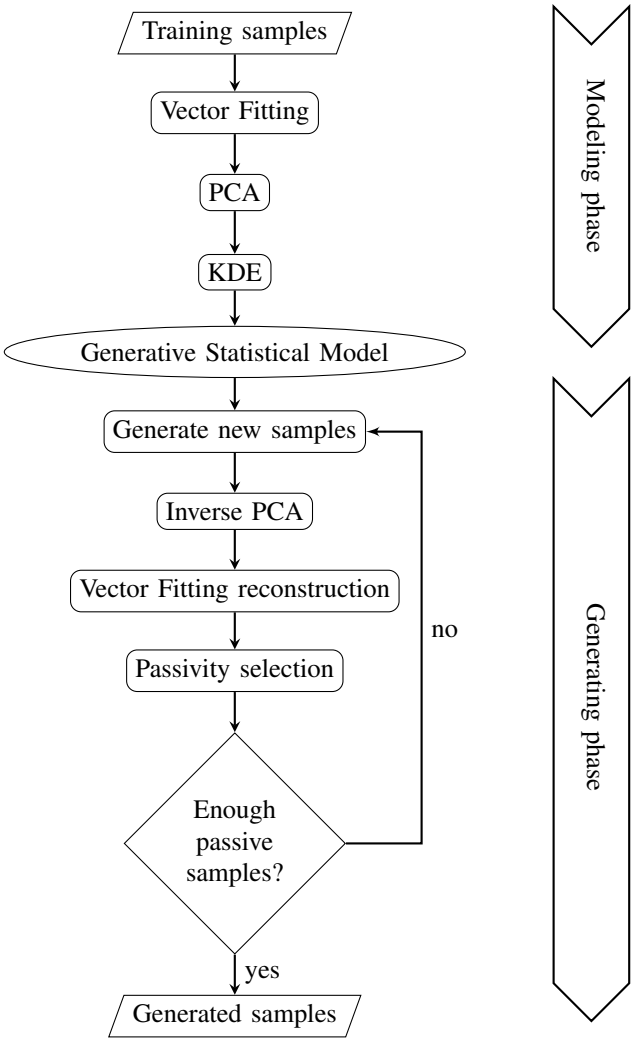

Figure 1. Flowchart of the proposed modeling and analysis framework.

B. Vector Fitting

First of all, the training samples are converted into a pole-residue representation by means of the well-known VF algorithm [8–10]. This allows working with a compact number of frequency-independent model parameters.

By means of measurements or simulations, a limited set of K training samples is available in the form of a generic frequency-domain response (S- or RLGC-parameters, in this paper). Each such response is characterized by a matrix $\overline{\overline{S}}_k(s), k = 1, \ldots, K$, where s denotes the complex frequency. Through use of the VF algorithm, each $\overline{\overline{S}}_k(s)$ is fitted by a rational model as follows:

$$\overline{\overline{S}}_k(s) \approx \sum_{i=1}^{N} \frac{\overline{\overline{R}}_{k,i}}{s - a_i} + \overline{\overline{D}}_k + \overline{\overline{E}}_k s, \quad k = 1, \ldots, K, \qquad (1)$$

where a_i and $\overline{\overline{R}}_{k,i}$ are the poles and the corresponding residue matrices, respectively. The poles and residues are real values

or constitute complex conjugate pairs. For an n-port structure, the S-parameters, and consequently the residues, form a $n \times n$ matrix. Moreover, for stable systems, all poles lie in the left half complex plane (i.e., $Re\{a_i\} < 0$). The quantities $\overline{\overline{D}}_k$ and $\overline{\overline{E}}_k$ are real valued and describe the behavior of $\overline{\overline{S}}_k(s)$ for $s \to 0$ and $s \to \infty$, respectively. They may be omitted in favor of a more homogeneous set of quantities to model. This might require to slightly increase the number of residues N in (1) in order to compensate for the reduced modeling power and still achieve a similar level of accuracy. Without loss of generality, matrices $\overline{\overline{D}}_k$ and $\overline{\overline{E}}_k$ are excluded from the formulation in the following discussion. Note that a generative model could also be constructed from time-domain data, by using Time Domain Vector Fitting (TD-VF) [20].

Furthermore, all S-parameter samples are fitted using a common pole set $\{a_i\}_{i=1}^{N}$. An optimal set of common poles is obtained by fitting all the training samples at once, and this set of poles is later on also used to generate new samples with the statistical model. This avoids any issue with possible outliers appearing in the right half plane and giving rise to unstable samples.

With the above approach, the problem is reduced to building a generative model for the residue matrices only. Capturing the relationship between the different S-parameters requires simultaneous modeling of all residues (i.e. all residues of every S-parameter matrix element).

C. Principal Component Analysis

For each training sample k, a set of N residue matrices $\overline{\overline{R}}_{k,i}$ having dimensions $n \times n$ are obtained. Fortunately, due to reciprocity, it suffices to model only the upper triangular part of $\overline{\overline{R}}_{k,i}$, as the matrices are symmetrical. However, as this still means $Nn(n+1)/2$ possibly complex variables to be modeled, the dimensionality quickly becomes computationally prohibitive. Since the residues are real or pairwise complex conjugate, the total number of elements remains unchanged when they are reshaped into a vector of real quantities. Stacking these vectors for all K training samples yields a matrix $\overline{\overline{X}}$ of size $K \times Nn(n+1)/2$.

To alleviate this high dimensionality strain, a principal component analysis (PCA) [11–13] is performed. This dimensionality reduction technique projects the high-dimensional space spanned by the training set onto a lower-dimensional space with maximal conservation of variance. Before the actual PCA step, each variable is centered and rescaled to unit standard deviation. PCA then projects $\overline{\overline{X}}$ onto the sequentially orthogonalized eigenvectors of $\overline{\overline{X}}^T\overline{\overline{X}}$ having the largest eigenvalues (i.e. with the largest variance along its direction, as $\overline{\overline{X}}^T\overline{\overline{X}}$ is proportional to the covariance matrix of $\overline{\overline{X}}$).

As the number of degrees of freedom is constrained by the size of the training set, K, and one degree of freedom is lost in the centering, the projected space is at most $(K-1)$-dimensional, which is in practice much smaller than $Nn(n+1)/2$. Another effect of this projection is that the new variables are also linearly uncorrelated. After the PCA, the projected training samples are again scaled to unit variance to make the subsequent KDE numerically feasible, as the variances along different axes in the projected space can differ by many orders of magnitude, and KDE is ill-equipped to handle these differences.

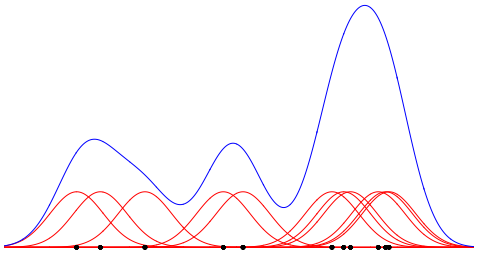

Figure 2. One-dimensional kernel density estimate starting from eleven training points (in black). The (Gaussian) kernels are shown in red. The blue curve is the sum of these kernels and is proportional to the distribution estimate.

D. Kernel Density Estimation

Now that the dimensionality of the variables has been reduced to a more manageable size, the distribution of the elements in the projected space can be more efficiently estimated. Note that due to the nonlinear correlation between the residues, this distribution is in general not spherical. Therefore, it is estimated by means of a multivariate KDE [14–16]. The KDE is a non-parametric method that estimates a probability distribution by placing a 'kernel' or elementary spherical distribution centered on each training point. The sum of these kernels at a particular point is then (up to a normalization factor) the total distribution estimate. One popular choice of such kernels is the Gaussian kernel. Fig. 2 shows a one-dimensional example of a kernel density estimate with Gaussian kernels. Other common kernels are Epanechnikov kernels, which are parabolic and have a finite support.

The difficulty of KDE lies in the estimation of its bandwidth. In the case of a (multivariate) Gaussian kernel, this is equivalent to finding the proper covariance matrix of the kernels. Most methods rely on the (estimated) minimization of the Mean Integrated Squared Error (MISE) or its asymptotic approximation (AMISE). For the results presented in Section IV, we use the AMISE-based approach detailed in [16].

E. Generative Statistical Model

Generating new samples is now straightforward. To sample from the kernel density estimate, it suffices to pick one training sample (with equal probability for each training sample), and then draw a sample from the kernel distribution centered on that training point. The newly generated samples are then rescaled, projected back into the full residue space, followed by a rescaling and a final offset. This reverses the transformations discussed in Section III-C, yielding N possibly complex $n \times n$-matrices for each new sample. Finally, using the same common poles that were used to fit the training data, new S-parameters are created using the rational Vector Fitting model (1).

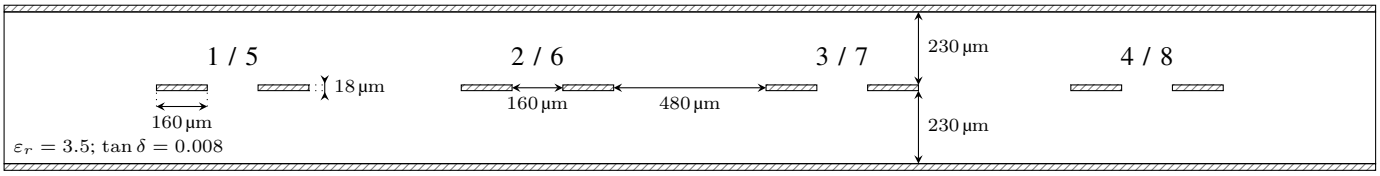

Figure 3. Cross-section of the differential stripline. The lines have a length of 5 cm. Above each conductor pair, the corresponding differential port numbers are shown (near-end / far-end).

F. Passivity Selection

Some of these generated samples may violate the physical constraint of passivity, and must therefore be rejected. If passivity were enforced, e.g. as in [21–23], the distribution of the generated samples would be biased toward samples that are near the passivity boundary. Therefore, we opt to simply reject non-passive samples a posteriori, and keep generating new samples until the desired amount is reached (see Fig. 1). The passivity of the newly generated S-parameters also implies their causality.

IV. Application Examples and Numerical Results

In this section, the proposed methodology is demonstrated on some meaningful examples. A set of differential striplines is considered first. Secondly, a commercial connector footprint is modeled. In a third example, the models from the previous examples are cascaded to form a realistic interconnect link. Finally, the time-domain behavior of the link is evaluated.

A. Stripline interconnect

A stripline interconnect consisting of four differential pairs, with cross-section depicted in Fig. 3, is considered. A set of 1000 per-unit-length (p.u.l.) resistance (R), inductance (L), conductance (G) and capacitance (C) (RLGC) parameters (from 0 to 20 GHz) were generated with an accurate field solver [24] where the relative permittivity ϵ_r of the substrate was varied according to a Gaussian distribution with a standard deviation of 5 % of the mean value of 3.5. $K = 50$ of these samples are used to train the model, while the other 950 will merely serve as a validation for the generated samples. In this example, the method described in Section III is applied to the p.u.l. RLGC parameters directly. This has the advantage of eliminating the distributed effect due to the line length from the VF model. As VF can also accommodate real variables, the proposed method remains unaltered, apart from a conversion from p.u.l. parameters to S-parameters, where a line length of 50 mm was assumed, and a relaxation of the stability enforcement. Fig. 4 shows the training set (red), the samples generated by the model (blue) and the validation set (green) for both the p.u.l. conductance $G_{1,1}$ and the p.u.l. capacitance $C_{1,1}$ of the leftmost stripline. As only the relative permittivity is varied, the R and L parameters are not affected by the variability and are therefore the same for all samples.

The distribution of the generated set of G and C parameters is visually very similar to that of the validation set. This similarity is more objectively apparent in Fig. 5, where the

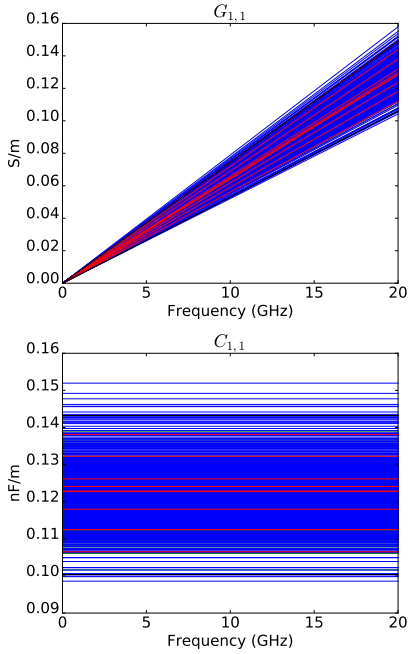

Figure 4. Modeling of the p.u.l. conductance $G_{1,1}$ and the p.u.l. capacitance $C_{1,1}$ of the utmost left stripline. 190 out of the 950 validation samples are shown in the background in green. 200 out of the 1000 generated samples are plotted in blue, and 10 out of the 50 training samples are shown in red on top. The black lines highlight the extreme values of the validation set (but not necessarily the boundaries of the underlying distribution).

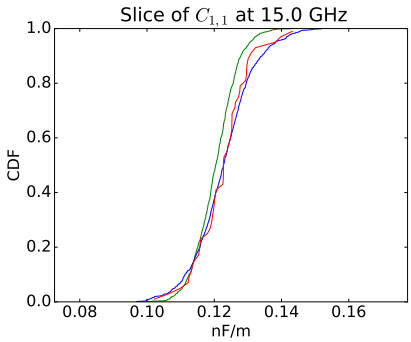

Figure 5. CDF of the training (red), generated (blue) and validation (green) sets for the $C_{1,1}$-parameter.

cumulative distribution functions (CDF) of the $C_{1,1}$-parameter are compared. Note that due to the structure's homogeneity, each C_{ij}-parameter is proportional to ϵ_r and thus also Gaussian distributed. Furthermore, $\overline{\overline{C}}$ is constant over the frequency

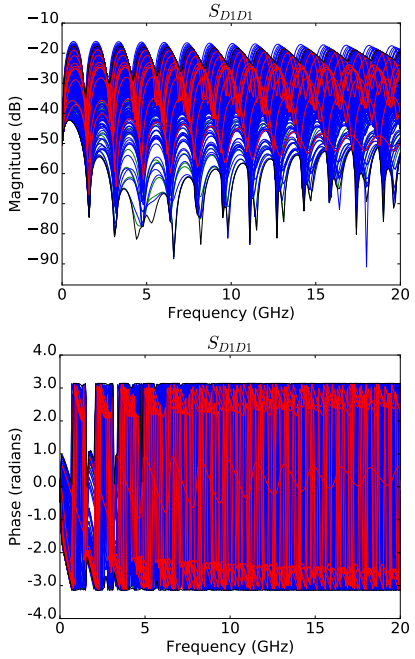

Figure 6. Magnitude and phase of the reflection (S_{D1D1}) S-parameters for the stripline example. Colors are as in Fig. 4.

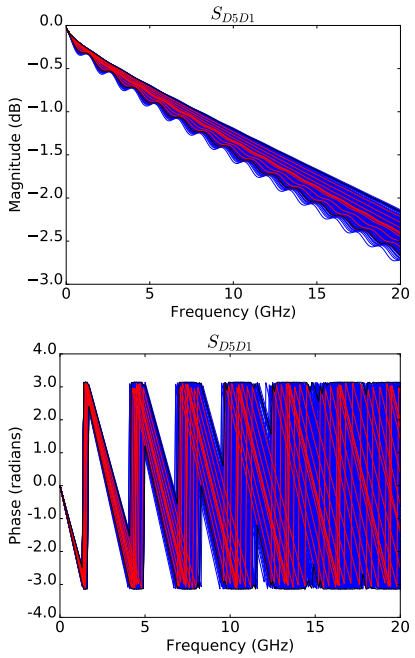

Figure 7. Magnitude and phase of the transmission (S_{D5D1}) S-parameters for the stripline example. Colors are as in Fig. 4.

range, making the CDF the same for every frequency point. The obtained p.u.l. RLGC parameters are converted to mixed-mode S-parameters to demonstrate their validity as a multivariate functional distribution estimate. Figs. 6-8 show the simulated and generated samples of some of the differential S-parameters.

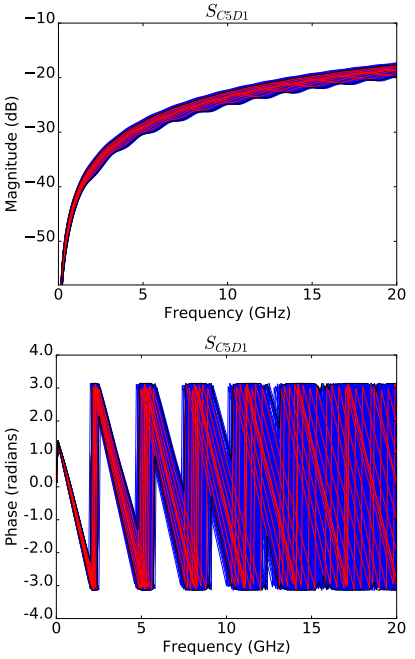

Figure 8. Magnitude and phase of the mode conversion (S_{C5D1}) S-parameters for the stripline example. Colors are as in Fig. 4.

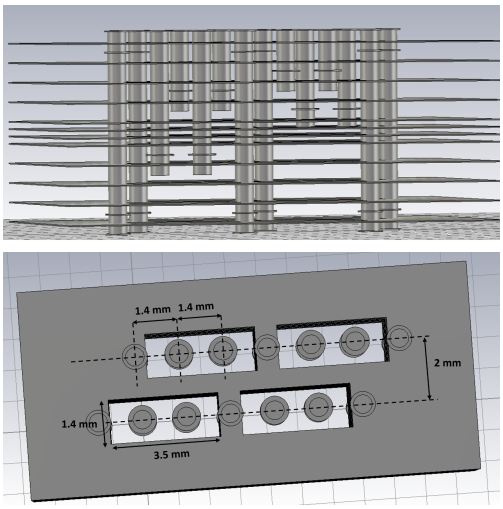

Figure 9. 3D rendering of the commercial connector footprint in a multi-layered printed circuit board. The first image is a side view, the second is a top view detailing some of its dimensions. The nominal pad diameter of each conductor is 900 µm, and their nominal drilled diameter is 600 µm.

B. Connector Footprint

As a second example, the S-parameters of a commercial 16-port connector footprint, depicted in Fig. 9, are considered. For this example, 450 S-parameter matrices were simulated by varying 40 geometrical parameters. Only 50 of these samples were used to train the proposed model. The model was then used to generate another set of 450 samples. Note that gPC based approaches would have difficulty coping with the large amount of independently varying parameters.

Figs. 10-12 portray the validation, generated, and training samples for some of the S-parameter elements.

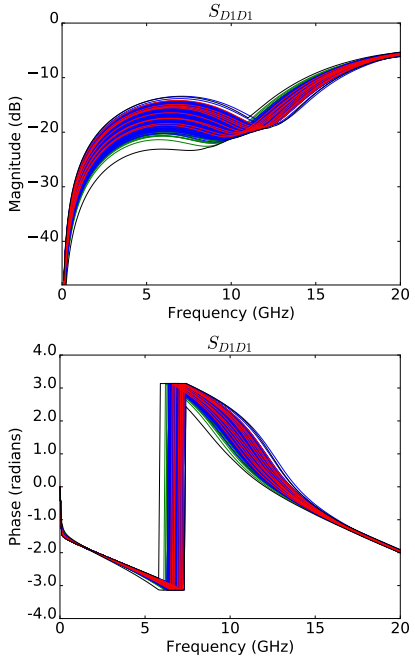

Figure 10. Magnitude and phase of the reflection (S_{D1D1}) S-parameters for the commercial connector footprint. Colors are as in Fig. 4.

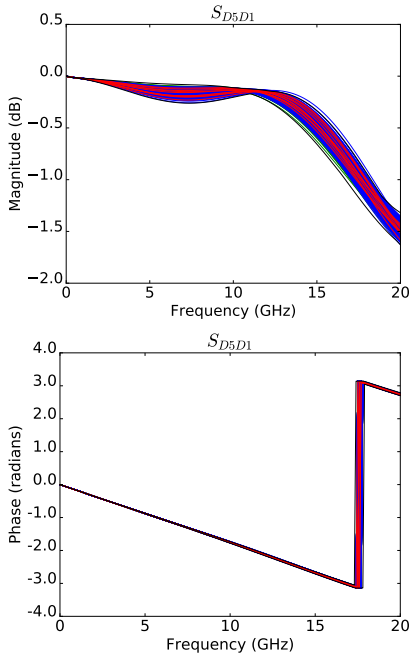

Figure 11. Magnitude and phase of the transmission (S_{D5D1}) S-parameters for the commercial connector footprint. Colors are as in Fig. 4.

From these figures, a good correspondence between the validation samples and the generated samples is apparent. To provide a more quantitative assessment, Fig. 13 shows the cumulative distribution function of S_{C5D1} at a frequency of 17.5 GHz for each of the S-parameter sets.

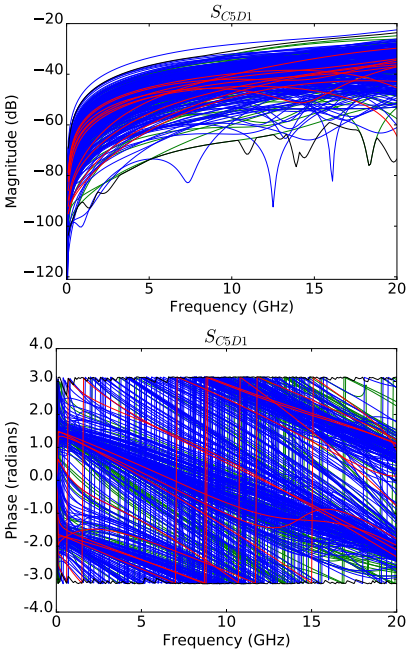

Figure 12. Magnitude and phase of the mode conversion (S_{C5D1}) S-parameters for the commercial connector footprint. Colors are as in Fig. 4.

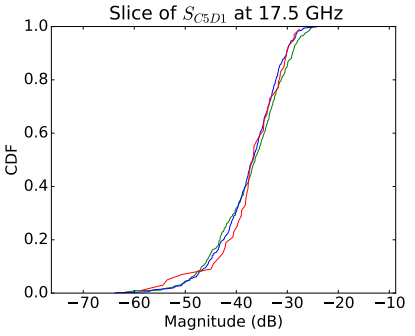

Figure 13. CDF of the validation (green), generated (blue) and training (red) sets for the S_{C5D1}-parameter of the commercial connector footprint example at 17.5 GHz.

C. Cascaded interconnect link

In the third example, the viability of the proposed modeling approach is further ascertained by considering a cascade connection (connector-stripline-connector) of the components modeled in the two previous examples (see Fig. 14). Thereto, for each simulated stripline S-parameter sample, two footprint S-parameter samples are chosen at random and cascaded at either end of the stripline. The same is done for the generated

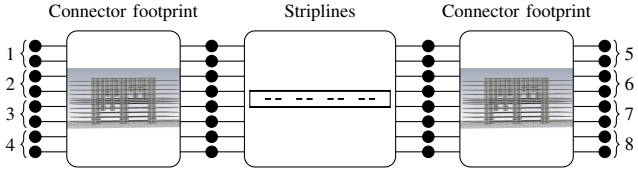

Figure 14. Schematic representation of the complete interconnect link. The numbers on either side denote differential port numbers.

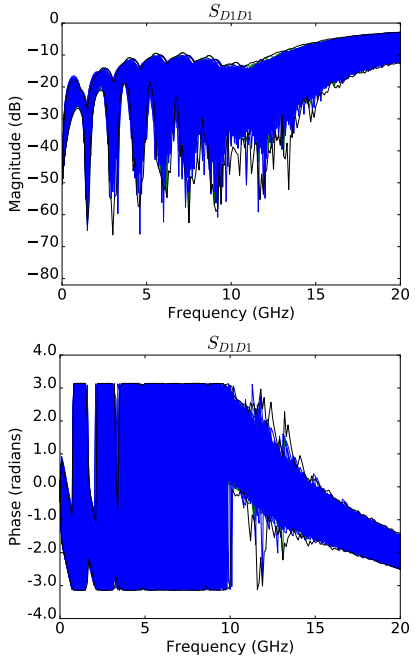

Figure 15. Magnitude and phase of the reflection (S_{D1D1}) S-parameters for the cascade example. Colors are as in Fig. 4, but without the red training samples.

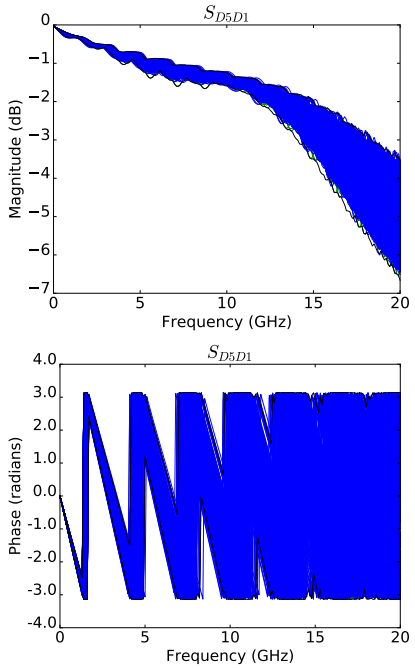

Figure 16. Magnitude and phase of the transmission (S_{D5D1}) S-parameters for the cascade example. Colors are as in Fig. 15.

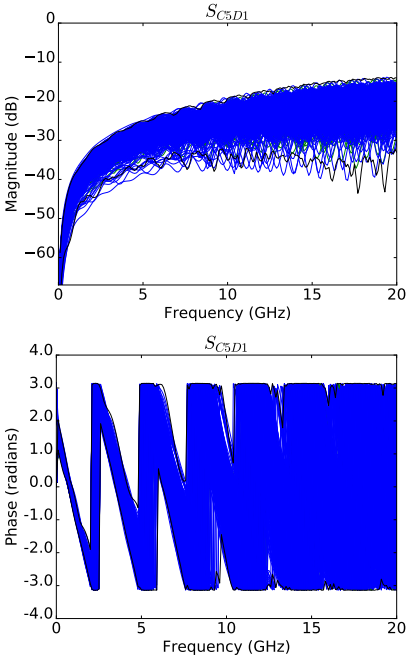

Figure 17. Magnitude and phase of the mode conversion (S_{C5D1}) S-parameters for the cascade example. Colors are as in Fig. 15.

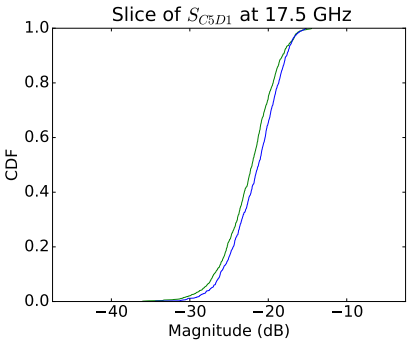

Figure 18. CDF of the validation (green) and generated (blue) sets for the S_{C5D1}-parameter of the cascade example at 17.5 GHz.

S-parameter samples of the footprint and stripline examples. In Fig. 15-17, the simulated (validation) and generated sets are superposed and appear to be in excellent agreement, despite the rather complex behavior of the S-parameters. The CDF at a single frequency, given in Fig. 18, confirms the good agreement between the two distributions.

As can be seen from Fig. 15-18, the variability has a non-negligible impact on the performance of the link. Whereas the return loss (Fig. 15) remains below $-10\,\mathrm{dB}$ from DC up to about $12\,\mathrm{GHz}$, for higher frequencies, the reflection may become quite high for certain link realizations. Consequently, there will be a poor transmission. As observed from Fig. 16, the insertion loss at $20\,\mathrm{GHz}$ ranges between $3\,\mathrm{dB}$ and $7\,\mathrm{dB}$, depending on the sample. The mode conversion (Fig. 17-18) also exhibits a large variation of more than $10\,\mathrm{dB}$, but fortunately, it remains rather low in the entire frequency range.

D. Time Domain Analysis

In a final example, the behavior of the generated samples in the time domain is explored. For this purpose, both the validation and the generated samples from the cascade example were used to predict the voltage response at each port when the leftmost pair is excited with a differential voltage step. The voltage step has an amplitude of $1\,\mathrm{V}$ and a rise time of $250\,\mathrm{ps}$, and it is produced by a generator with an

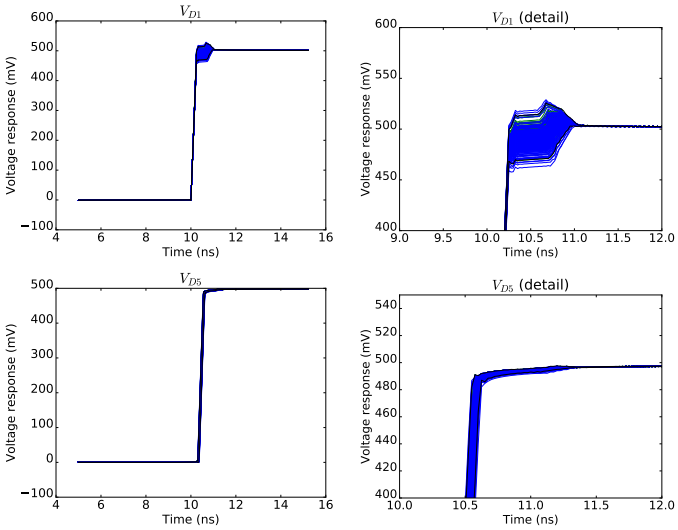

Figure 19. Step voltage response at ports $D1$ (reflection) and $D5$ (transmission) for the cascade example. Colors are as in Fig. 15.

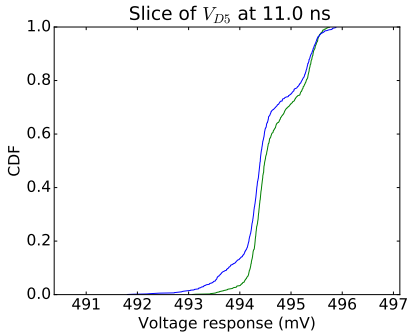

Figure 20. CDF of the validation (green) and generated (blue) sets for the voltage at port $D5$ at 11.0 ns.

internal resistance of $50\,\Omega$. The resulting voltage responses at the input and through-connection differential ports are shown in Fig. 19. Fig. 20 provides a comparison between the CDFs of the generated and validation sets for the through-connection voltage at port $D5$ at 11.0 ns. The shapes of the distributions clearly match, albeit with a slight discrepancy in the lower voltage tail.

As a last verification of the generated samples' validity, the height and width of an eye diagram are calculated for each sample. To this end, a single 1 Gbps pseudo-random bit sequence with rise/fall times of 125 ps was applied to the differential port $D1$ and used to construct the eye diagram at port $D5$. An example of such eye diagram is given in Fig. 21. The height and width of each eye diagram were calculated and they are compared in Fig. 22, showing that the eye diagram features obtained from the generated S-parameters match those of a large set of simulated S-parameters.

The variability observed from the frequency domain results of section IV-C are of course also noticable in the time domain. For example, from Fig. 19, it becomes clear that the reflection mismatch at port $D1$ leads to over- or undershoot. From Fig. 22 it is evident that there is considerable variation of the eye height and width. Together with the voltage waveform V_{D5}

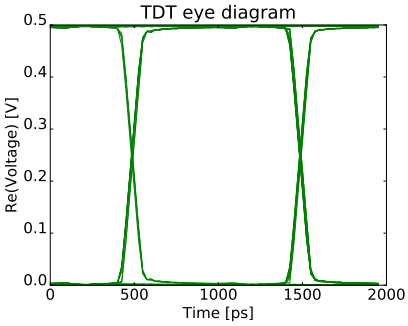

Figure 21. Example of a constructed eye diagram for one of the validation samples.

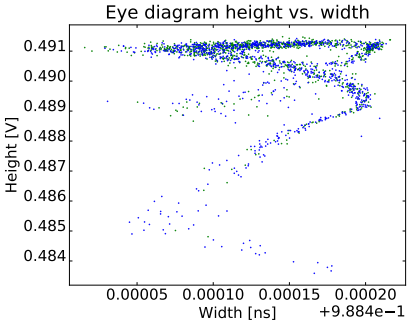

Figure 22. Distribution of height and width of eye diagrams of the validation (green) and generated (blue) sets.

(Figs. 19 and 20), it becomes apparent that the link's variability has a non-negligible impact on potential jitter issues.

E. Training set size

So far, a training set size of 50 samples has been used to build the model. One may wonder what the optimal number of training samples might be. On the one hand, there should be as few as possible in order to minimize the amount of effort that goes into simulating or measuring the training set. On the other hand, the more training samples are available, the better the model can approximate the real data distribution. This trade-off depends on the complexity of the device being modeled.

In order to quantify this trade-off, we introduce a measure called energy distance. This measure quantifies the distance

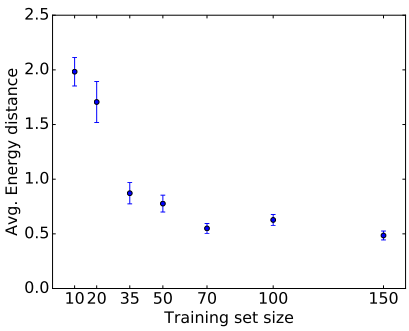

Figure 23. Averaged energy distance for the cascade example.

between two empirical probability distributions and is defined as:

$$E_{n,m}(X,Y) = \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\|x_i - y_j\| - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|x_i - x_j\| - \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\|y_i - y_j\|, \quad (2)$$

where X and Y are vectors of lengths n and m, respectively, collecting all samples in each respective empirical probability distribution function. It can be shown that $E_{n,m}(X,Y) \geq 0$, where the equality holds only if the two distributions perfectly coincide.

Fig. 23 shows an averaged energy distance for the cascade example in function of the training set size. This averaging occurs over frequency points, S-parameter matrix elements, and twelve random training set selections from the complete validation sets of the stripline and footprint examples. The error bars show the estimated mean error arising from the twelve different training sets. In this figure we can see a rapid decrease in both energy distance mean and error, followed by a slight stagnation. As this stagnation sets in around a training set size of 50, we can conclude that — even for such a relatively complex cascade example — a training set size of 50 is sufficient to provide an accurate approximation.

V. Conclusions

In this paper, a novel method is proposed that allows a designer to generate a large set of S-parameter realizations from a small training set of simulated or measured stochastic S-parameters. This method operates by fitting the S-parameters with a rational model in partial-fraction form using the VF technique using a common pole set. For each training sample, the residue matrices corresponding to each pole are identified. This training set of residues is then reduced by means of PCA, and finally its distribution is modeled by a KDE. The generation of new samples is achieved by drawing from the KDE, projecting back to the full residue space using the inverse PCA, and finally reconstructing the S-parameters using the rational VF model with common poles. A post-processing passivity selection ensures the physical consistency of the generated samples.

The validity and aptness of the generated S-parameters are verified in four application examples. The first application is a 16-port differential stripline interconnect that is modeled through its RLGC-parameters. The second example considers a commercial connector footprint. As a third verification, the S-parameters of an interconnect link (cascade of the striplines and connector footprint) are compared to their simulated counterparts. Finally, the generated samples are used to perform time-domain analyses by computing both step voltage responses and eye diagrams. For each of these examples, a very good agreement between the distributions of simulated and generated S-parameters is observed.

One of the goals of our future research is to obtain a validation based on measured instead of simulated data. The measurement validation for the connector example requires a costly time investment as S-parameter measurements for such a multiport device are very painstaking. Nonetheless, as the initial VF step of the model based on a measured training set will produce similar residue distributions, we expect that the proposed methodology will also prove useful for measured data.

References

[1] P. Manfredi, D. Vande Ginste, D. De Zutter, and F. G. Canavero, "Uncertainty assessment of lossy and dispersive lines in SPICE-type environments," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 3, no. 7, pp. 1252–1258, July 2013. [Online]. Available: http://dx.doi.org/10.1109/TCPMT.2013.2259295

[2] Z. Zhang, T. A. El-Moselhy, I. M. Elfadel, and L. Daniel, "Stochastic testing method for transistor-level uncertainty quantification based on generalized polynomial chaos," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 32, no. 10, pp. 1533–1545, Oct 2013. [Online]. Available: http://dx.doi.org/10.1109/TCAD.2013.2263039

[3] J. S. Ochoa and A. C. Cangellaris, "Random-space dimensionality reduction for expedient yield estimation of passive microwave structures," *IEEE Transactions on Microwave Theory and Techniques*, vol. 61, no. 12, pp. 4313–4321, Dec 2013. [Online]. Available: http://dx.doi.org/10.1109/TMTT.2013.2286968

[4] M. R. Rufuie, E. Gad, M. Nakhla, and R. Achar, "Generalized hermite polynomial chaos for variability analysis of macromodels embedded in nonlinear circuits," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 4, no. 4, pp. 673–684, April 2014. [Online]. Available: http://dx.doi.org/10.1109/TCPMT.2013.2285877

[5] P. Manfredi, D. Vande Ginste, D. De Zutter, and F. G. Canavero, "Stochastic modeling of nonlinear circuits via SPICE-compatible spectral equivalents," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 61, no. 7, pp. 2057–2065, July 2014. [Online]. Available: http://dx.doi.org/10.1109/TCSI.2014.2304667

[6] ——, "Generalized decoupled polynomial chaos for nonlinear circuits with many random parameters," *IEEE Microwave and Wireless Components Letters*, vol. 25, no. 8, pp. 505–507, Aug 2015. [Online]. Available: http://dx.doi.org/10.1109/LMWC.2015.2440779

[7] M. Ahadi and S. Roy, "Sparse linear regression (SPLINER) approach for efficient multidimensional uncertainty quantification of high-speed circuits," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. PP, no. 99, pp. 1–1, 2016. [Online]. Available: http://dx.doi.org/10.1109/TCAD.2016.2527711

[8] B. Gustavsen and A. Semlyen, "Rational approximation of frequency domain responses by vector fitting," *IEEE Transactions on Power Delivery*, vol. 14, no. 3, pp. 1052–1061, July 1999. [Online]. Available: http://dx.doi.org/10.1109/61.772353

[9] B. Gustavsen, "Improving the pole relocating properties of vector fitting," *IEEE Transactions on Power Delivery*, vol. 21, no. 3, pp. 1587–1592, July 2006. [Online]. Available: http://dx.doi.org/10.1109/TPWRD.2005.860281

[10] D. Deschrijver, M. Mrozowski, T. Dhaene, and D. De Zutter, "Macromodeling of multiport systems using a fast implementation of the vector fitting method," *IEEE Microwave and Wireless Components Letters*, vol. 18, no. 6, pp. 383–385, June 2008. [Online]. Available: http://dx.doi.org/10.1109/LMWC.2008.922585

[11] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, pp. 417–441, 1933. [Online]. Available: http://dx.doi.org/10.1037/h0071325

[12] S. Wold, K. Esbensen, and P. Geladi, "Proceedings of the multivariate statistical workshop for geologists and geochemists principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1, pp. 37 – 52, 1987. [Online]. Available: http://dx.doi.org/10.1016/0169-7439(87)80084-9

[13] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010. [Online]. Available: http://dx.doi.org/10.1002/wics.101

[14] M. Rosenblatt, "Remarks on some nonparametric estimates of a density function," *Ann. Math. Statist.*, vol. 27, no. 3, pp. 832–837, 09 1956. [Online]. Available: http://dx.doi.org/10.1214/aoms/1177728190

[15] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Statist.*, vol. 33, no. 3, pp. 1065–1076, 09 1962. [Online]. Available: http://dx.doi.org/10.1214/aoms/1177704472

[16] M. Kristan, A. Leonardis, and D. Skoaj, "Multivariate online kernel density estimation with gaussian kernels," *Pattern Recognition*, vol. 44, no. 1011, pp. 2630 – 2642, 2011, semi-Supervised Learning for Visual Content Analysis and Understanding. [Online]. Available: http://dx.doi.org/10.1016/j.patcog.2011.03.019

[17] S. De Ridder, P. Manfredi, J. De Geest, T. Dhaene, D. De Zutter, and D. Vande Ginste, "A novel methodology to create generative statistical models of interconnects," IEEE EDAPS 2016 conference, 3 pages.

[18] L. L. Campbell and J. E. Purviance, "Interpolative modeling of GaAs FET S-parameter data bases for use in Monte Carlo simulations," in *4th NASA Symposium on VLSI Design*, January 1992, pp. 7.4.1–7.4.11. [Online]. Available: https://ntrs.nasa.gov/search.jsp?R=19940017250

[19] ——, "Interpolation modeling of S-parameter data bases for use in Monte Carlo simulations," *International Journal of RF and Microwave Computer-Aided Engineering*, vol. 4, no. 3, pp. 282–296, July 1994. [Online]. Available: http://dx.doi.org/10.1002/mmce.4570040308

[20] S. Grivet-Talocia, "Package macromodeling via time-domain vector fitting," *IEEE Microwave and Wireless Components Letters*, vol. 13, no. 11, pp. 472–474, Nov 2003.

[21] D. Deschrijver and T. Dhaene, "Stability and passivity enforcement of parametric macromodels in time and frequency domain," *IEEE Trans. Microw. Theory Tech*, vol. 56, no. 11, pp. 2435–2441, Nov. 2008. [Online]. Available: http://dx.doi.org/10.1109/TMTT.2008.2005868

[22] T. Dhaene, D. Deschrijver, and N. Stevens, "Efficient algorithm for passivity enforcement of S-parameter-based macromodels," *IEEE Transactions on Microwave Theory and Techniques*, vol. 57, no. 2, pp. 415–420, Feb. 2009. [Online]. Available: http://ieeexplore.ieee.org/abstract/document/4752847/

[23] S. Grivet-Talocia, "Passivity enforcement via perturbation of hamiltonian matrices," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 51, no. 9, pp. 1755–1769, Sept 2004. [Online]. Available: http://dx.doi.org/10.1109/TCSI.2004.834527

[24] T. Demeester and D. De Zutter, "Quasi-TM transmission line parameters of coupled lossy lines based on the dirichlet to neumann boundary operator," *IEEE Transactions on Microwave Theory and Techniques*, vol. 56, no. 7, pp. 1649–1660, July 2008. [Online]. Available: http://ieeexplore.ieee.org/abstract/document/4538249/