On the Optimum Number of Coefficients of Sparse Digital Predistorters: a Bayesian Approach

J. A. Becerra, *Senior Member, IEEE,* M. J. Madero-Ayora, *Senior Member, IEEE,*
R. G. Noguer, and C. Crespo-Cadenas, *Senior Member, IEEE*

Abstract—This work presents insights on the application of the Bayesian information criterion (BIC) to fix the optimum number of coefficients in Volterra series applied to the modeling and linearization of power amplifiers. The BIC is transformed from a rule to be applied after selection techniques to a stopping criterion, that enables the halting of the algorithm when a condition is reached. This study reveals that the BIC is equivalent to allow a certain identification normalized mean square error (NMSE) decrease after the inclusion of a model component. Experimental results of the digital predistortion of a class J power amplifier are provided, demonstrating the proposal applicability in the attaining of the optimum number of coefficients. A comparison is made between the results obtained when the stopping rule is applied to the hill climbing (HC) and the doubly orthogonal matching pursuit (DOMP) algorithms.

Index Terms—Bayesian information criterion, Digital pre-distortion, Doubly orthogonal matching pursuit, Hill climbing algorithm, Order reduction, Power amplifier.

I. INTRODUCTION

In a practical situation of model fitting, the model order is generally unknown. Many order selection methods have been proposed, each one of them following a different approach. The Akaike Information Criterion (AIC) [1] estimates the information lost by a model, providing a means for model selection through the trade-off between the goodness of the fit and the penalty because of the number of estimated parameters. Another popular method is the Bayesian information criterion (BIC) [2] that selects the best model by maximizing the value of the likelihood function. Despite the extensive literature on model selection techniques, the most common order selection criteria remain the AIC and the BIC [3].

Regarding the compressed-sensing recovery of the coefficients of the model, there are a variety of algorithms that have been used in practical applications [4]. There exist multiple scenarios in which the dependence on just a certain number of regressors is desired. This tuning can be achieved by applying model selection techniques, which aim at choosing one model from a set of candidate models.

In the context of behavioral modeling and digital predistortion (DPD) of power amplifiers (PAs), the use of Volterra series as the underlying model poses a serious drawback because of their rapid grow in number of coefficients with the nonlinear order and memory depth. Coefficient selection techniques enable their use by reducing their computational complexity. Recently, the BIC has been applied to fix the optimum number of coefficient [5]–[9] and there exist a vast literature on order reduction techniques that can benefit from its features, such as the hill climbing (HC) [10] algorithm and the doubly orthogonal matching pursuit (DOMP) [11].

In this letter, we focus on the application of the BIC to iterative selection techniques by transforming the BIC cost function to a stopping criterion in order to provide an optimal upper bound for the number of coefficients from information theory perspective. Considering that the BIC was previously used after the DOMP run, here we enable the algorithm halt in the corresponding iteration, which produces remarkable savings in execution time. Also, we give formalism to the ad-hoc rule that was applied to the HC, providing a unique model-selection scheme. The remaining of this communication continues with the notation and prerequisites in Section II. The transformation of the BIC into a stopping rule is provided in Section III. Section IV deals with experimental design and results and Section V concludes this work.

II. NOTATION AND PREREQUISITES

The recovery of unknown model parameters $\mathbf{h} \in \mathbb{C}^N$ after a spatial transformation is usually modeled by the measurement equation

$$\mathbf{y} = \mathbf{X}\mathbf{h} + \mathbf{e}, \tag{1}$$

where $\mathbf{y} \in \mathbb{C}^M$ is the output of the system, $\mathbf{X} \in \mathbb{C}^{M \times N}$ is the measurement matrix that holds one of the N model regressors in each of its columns, $\mathbf{e}$ is the measurement error and M is the number of sample points. A common approach to obtain an estimation $\hat{\mathbf{h}}$ is the minimization of the quadratic error $\|\mathbf{e}\|_2^2$, that leads to the *normal* equation and to the least-squares (LS) solution

$$\hat{\mathbf{h}} = \left(\mathbf{X}^H \mathbf{X}\right)^{-1} \mathbf{X}^H \mathbf{y}, \tag{2}$$

where $\mathbf{X}^H$ stands for the Hermitian transpose of matrix $\mathbf{X}$. The estimation of the output signal is attained with $\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{h}}$. Although this well-established procedure provides a unique solution to the regression, it also finds a dependence of the system output with the whole set of model regressors. The application of compressed-sensing techniques to this scheme allows a sparse recovery of the solution.

Selection techniques aim at obtaining an ordered support set S which contains the indices of the measurement matrix columns sorted by importance. In each iteration the algorithm adds one component —or a set of them, depending on the technique— to the support set, an estimation of $\mathbf{h}$ is performed and the residual error $\mathbf{r}$ is calculated. This loop is repeated until a stopping criterion is achieved. Order selection suppose that there exists a set of candidate models, amongst which the minimum of a cost function will define the optimal model. Since sorting algorithms select one or more components per iteration, that is, they are incremental, it is reasonable to transform the criterion that has to be evaluated after the loop of the algorithm to a stopping rule that breaks the execution once the condition is fulfilled.

III. THE BIC AS A STOPPING RULE

The BIC is defined as the sum of a term that depends on the error and a penalty that is related to the number of model components n [5]

$$\text{BIC} = 2M \ln \hat{\sigma}_{\mathbf{e}}^2 + 2n \ln 2M, \tag{3}$$

where $\hat{\sigma}_{\mathbf{e}}^2$ is the error variance and M is the number of samples used for the model identification. The BIC in (3) acts as a trade-off between the error and the number of components. In an iterative sorting algorithm, the modeling error is commonly known as the residual error $\mathbf{r}$, therefore

$$\hat{\sigma}_{\mathbf{e}}^2 = \text{Var}[\mathbf{e}] = \text{Var}[\mathbf{y} - \hat{\mathbf{y}}] = \text{Var}[\mathbf{r}] = \text{E}[|\mathbf{r}|^2] - |\text{E}[\mathbf{r}]|^2, \tag{4}$$

where $\text{Var}[\cdot]$ represents the variance of its argument, $\text{E}[\cdot]$ is the expected value and $|\cdot|$ stands for its module. Since it is a common practice to center the variables, without any loss of generality $\text{E}[\mathbf{r}] = 0$ can be assumed. Therefore,

$$\text{E}[|\mathbf{r}|^2] = \frac{1}{M} \sum_{i=0}^{M} |r_i|^2 = \frac{1}{M} \|\mathbf{r}\|_2^2, \tag{5}$$

where $\|\cdot\|_2$ stands for the ℓ_2 norm of its argument.

Combining the above results with (3) and adding an explicit dependence with the iteration $k \in \mathbb{N}$, the BIC can be rewritten as

$$\text{BIC}^{(k)} = 2M \ln \left(\frac{1}{M} \|\mathbf{r}^{(k)}\|_2^2 \right) + 2n^{(k)} \ln 2M, \tag{6}$$

where $n^{(k)}$ denotes the number of components at iteration k. This cost function is composed of two terms. The first term is dependent on $\|\mathbf{r}^{(k)}\|_2^2$. Since this magnitude measures the modeling error —or equivalently the amount of output power left to be captured by the model—, as more regressors are added to the model, the basis space in which the output is going to be expressed is richer and the residual decreases. The second term in (6) is a penalty that increases linearly with the number of components and also depends on a fixed number of samples M. Therefore, the BIC is initially decreasing, and a minimum will be found when its increment is not negative.

Assuming that there is no collinearity in the measurement matrix, i.e., all the regressors are different, the BIC is a monotonically decreasing function. As the BIC is initially
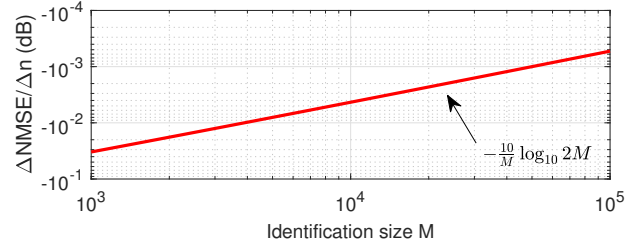

Fig. 1. Evolution of the allowed NMSE increment per iteration and component with the identification size for the BIC.

decreasing, a minimum will be found when it starts to increase, thus $\Delta \text{BIC}^{(k)} > 0$. The BIC increment in one iteration follows

$$\ln \left(\frac{1}{M} \|\mathbf{r}^{(k)}\|_2^2 \right) - \ln \left(\frac{1}{M} \|\mathbf{r}^{(k-1)}\|_2^2 \right) + \Delta n^{(k)} \frac{\ln 2M}{M} > 0, \tag{7}$$

where $\Delta n^{(k)} = n^{(k)} - n^{(k-1)}$ stands for the increment in number of components at iteration k, and rearranging the terms,

$$\frac{\|\mathbf{r}^{(k)}\|_2^2}{\|\mathbf{r}^{(k-1)}\|_2^2} > (2M)^{-\frac{1}{M}\Delta n^{(k)}}. \tag{8}$$

The normalized mean square error (NMSE) at iteration k is defined as $\text{NMSE}^{(k)} = 10 \log_{10} \frac{\|\mathbf{r}^{(k)}\|_2^2}{\|y\|_2^2}$, therefore

$$\frac{\Delta \text{NMSE}^{(k)}}{\Delta n^{(k)}} > -\frac{10}{M} \log_{10} 2M. \tag{9}$$

This relationship is plotted in Fig. 1. The stopping rule in (9) supports the following procedure: "*continue running the algorithm until it is not able to discard a certain amount of error per iteration and component*". This error threshold is dependent on the number of samples used for the identification of the model, which balances the error evolution shape with the number of model components and allows order selection with a reduced risk of overfitting [12].

A. Generalization of the BIC Stopping Rule and Relation with Other Proposals in the Literature

From (8), the definition of a general stopping rule follows

$$\frac{\|\mathbf{r}^{(k)}\|_2^2}{\|\mathbf{r}^{(k-1)}\|_2^2} > \alpha^{\Delta n^{(k)}}, \tag{10}$$

which is expressed as inequality considering that there does not exist regressor repetition. As the stopping rule is expected to be triggered when the model shows good accuracy, we can also state that the residual decrement will be slow, that is, $\|\mathbf{r}^{(k-1)}\|_2^2 \approx \|\mathbf{r}^{(k)}\|_2^2$. From (9) and (10),

$$\frac{\Delta \text{NMSE}^{(k)}}{\Delta n^{(k)}} > 10 \log_{10} \alpha, \tag{11}$$

where $\alpha < 1$, and its value in decibels can be interpreted as the NMSE tolerance per coefficient.

In (6) of [10] —and, equivalently, (10) of [13]— a cost function J was defined as the trade-off between model accuracy and model complexity,

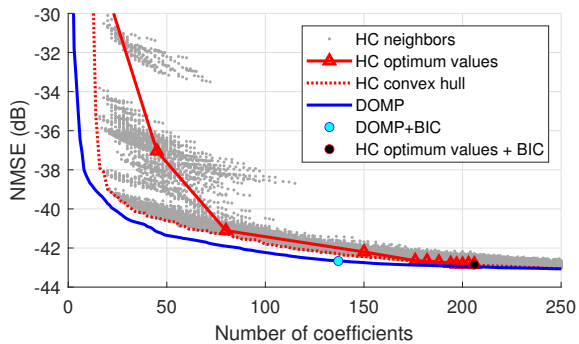$$J^{(k)} = \text{NMSE}^{(k)} + \mu \, n^{(k)}, \tag{12}$$

Fig. 2. Empirical identification NMSE of the HC and DOMP algorithms versus the number of coefficients. The BIC stopping rule indicates the moment in which the error difference decreases below a given threshold.

Fig. 3. Normalized power spectral density of the PA output without DPD and with DPD regressed with the number of coefficients indicated by the BIC.

where μ represents the NMSE tolerance per coefficient, and the minimum of $J^{(k)}$ indicates the optimum model. Considering (13) of [14], which states that

$$n_{\text{BIC}} = \arg\min_n \left[\text{NMSE}^{(k)} + \frac{n^{(k)}}{M} 10 \log_{10} 2M \right], \quad (13)$$

and following the same rationale, the optimum is found when

$$\frac{\Delta \text{NMSE}^{(k)}}{\Delta n^{(k)}} > \mu, \quad (14)$$

therefore $\mu = 10 \log_{10} \alpha$ and $\mu_{\text{BIC}} = -\frac{10}{M} \log_{10} 2M$.

IV. EXPERIMENTAL DESIGN AND RESULTS

In order to validate the proposal, a generalized memory polynomial (GMP) model [15] was regressed for the indirect-learning DPD of a PA working on its nonlinear regime. The testbench was composed by a SMU200A vector signal generator (VSG) from Rohde & Schwarz, a PXA-N9030A vector signal analyzer (VSA) from Keysight Technologies and a dc power supply. The device under test (DUT) was the cascade of two Mini-Circuits TVA-4W-422A+ preamplifiers and the continuous mode class J PA [16] operating at a center frequency of 850 MHz. The output power was set to $+31$ dBm characterized by 3.5 dB of gain compression.

The signal under test was designed according to a 5G-NR scheme with 20 MHz of bandwidth, 30 kHz carrier separation and a sampling rate of 92.16 MHz. The signal segment used for identification was composed of 5500 samples, that through (9) corresponds to a minimum-allowable NMSE decrement of $\mu = -0.0073$ dB or, equivalently, $\alpha = 0.998$.

The proposed BIC stopping rule was applied to two different selection techniques, the HC and the DOMP algorithms, and their results were compared. According to the BIC applied to the HC algorithm, the optimum GMP model presented a set of parameters given by $\mathcal{K}_a = 5$, $\mathcal{L}_a = 7$, $\mathcal{K}_b = 3$, $\mathcal{L}_b = 2$, $\mathcal{M}_b = 2$, $\mathcal{K}_c = 4$, $\mathcal{L}_c = 4$, $\mathcal{M}_c = 7$. For the sake of comparison, the DOMP was applied to the neighbour with maximum number of coefficients in the HC iterations to obtain a set of sorted regressors by importance, to which the BIC defined an optimum set. Fig. 2 shows the identification NMSE evolution with the number of coefficients. Note that the
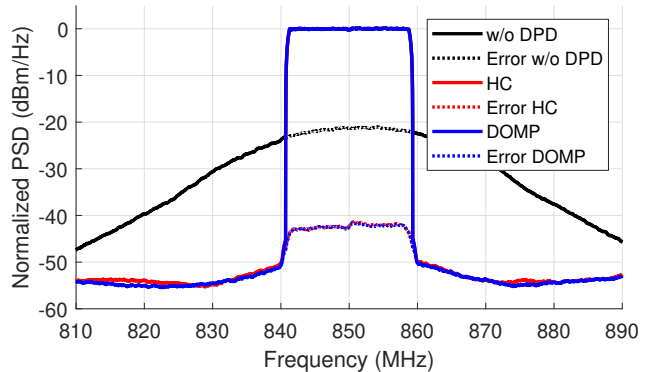
TABLE I
MEASURED PERFORMANCE IN TERMS OF NMSE, ACPR, EVM AND
NUMBER OF COEFFICIENTS FOR THE PA WITH AND WITHOUT DPD

Experiment	NMSE (dB)	ACPR -1/+1 (dBc)	EVM (%)	Number of coef.
w/o DPD	-19.5	-29.1 / -27.0	8.6	-
HC DPD	-41.4	-53.4 / -52.8	1.0	206
DOMP DPD	-41.5	-53.8 / -53.1	1.1	137

BIC was applied to both techniques independently, attaining an optimum number of coefficients of 137 and 206 in the DOMP and the HC techniques, respectively. This difference is justified by the shape of the error evolution, which reaches a steady state sooner in the DOMP case. Note that the DOMP curve does not match with the HC convex hull because the first selects the model coefficients while the latter selects the optimal GMP model parameters, i.e., the HC produces a complete GMP-model regressors matrix while the DOMP produces a subset of it, namely a sparse regressors matrix.

The optimum models for both HC and DOMP were used to generate a predistorted signal, whose linearization performance and power spectral density (PSD) are shown in Table I and Fig. 3, respectively. The effect of both techniques is evidenced by a reduction of the in-band distortion, achieving an enhancement of over 20 dB in NMSE and an error vector magnitude (EVM) of about 1% in the linearized signal. The adjacent channel power ratio (ACPR) improvement is illustrated by the decrease in spectral regrowth, achieving a level of below -52 dB in all the linearized cases.

V. CONCLUSION

A transformation of the BIC into a stopping rule for iterative algorithms has been presented in this letter. This modification allows to extract information about how the optimum model is selected amongst the set of candidates, that follows the rule of stop adding components to the support set once the new selection is not able to discard a certain amount of error. The result has been related to other forms of cost functions in the literature, providing a unique framework. Measurement results show the effect of the error evolution on the optimum number of coefficients and its applicability to techniques such as the HC and the DOMP.

REFERENCES

[1] H. Akaike, *Information Theory and an Extension of the Maximum Likelihood Principle*. Springer New York, 1998.

[2] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, Mar. 1978.

[3] J. Ding, V. Tarokh, and Y. Yang, "Bridging AIC and BIC: A new criterion for autoregression," *IEEE Trans. Inf. Theory*, vol. 64, no. 6, pp. 4024–4043, June 2018.

[4] Y. C. Eldar, *Sampling Theory: Beyond Bandlimited Systems*. Cambridge University Press, 2015.

[5] J. Reina-Tosina, M. Allegue-Martínez, C. Crespo-Cadenas, C. Yu, and S. Cruces, "Behavioral modeling and predistortion of power amplifiers under sparsity hypothesis," *IEEE Trans. Microw. Theory Techn.*, vol. 63, no. 2, pp. 745–753, Feb. 2015.

[6] Q. A. Pham, D. Lopez-Bueno, T. Wang, G. Montoro, and P. L. Gilabert, "Partial least squares identification of multi look-up table digital predistorters for concurrent dual-band envelope tracking power amplifiers," *IEEE Trans. Microw. Theory Techn.*, vol. 66, no. 12, pp. 5143–5150, Dec. 2018.

[7] P. L. Gilabert, D. Lopez-Bueno, and G. Montoro, "Spectral weighting orthogonal matching pursuit algorithm for enhanced out-of-band digital predistortion linearization," *IEEE Trans. Circuits Syst. II*, vol. 66, no. 7, pp. 1277–1281, Jul. 2019.

[8] C. Yu, K. Tang, and Y. Liu, "Adaptive basis direct learning method for predistortion of RF power amplifier," *IEEE Microw. Wireless Compon. Lett.*, vol. 30, no. 1, pp. 98–101, Jan. 2020.

[9] A. Brihuega, M. Abdelaziz, L. Anttila, M. Turunen, M. Allen, T. Eriksson, and M. Valkama, "Piecewise digital predistortion for mmWave active antenna arrays: Algorithms and measurements," *IEEE Trans. Microw. Theory Techn.*, Apr. 2020, doi: 10.1109/TMTT.2020.2994311.

[10] S. Wang, M. A. Hussein, O. Venard, and G. Baudoin, "A novel algorithm for determining the structure of digital predistortion models," *IEEE Trans. Veh. Technol.*, vol. 67, no. 8, pp. 7326–7340, Aug. 2018.

[11] J. A. Becerra, M. J. Madero-Ayora, and C. Crespo-Cadenas, "Comparative analysis of greedy pursuits for the order reduction of wideband digital predistorters," *IEEE Trans. Microw. Theory Techn.*, vol. 67, no. 9, pp. 3575–3585, Sep. 2019.

[12] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *IEEE Signal Proc. Mag.*, vol. 21, no. 4, pp. 36–47, July 2004.

[13] S. Wang, M. A. Hussein, O. Venard, and G. Baudoin, "Optimal sizing of two-stage cascaded sparse memory polynomial model for high power amplifiers linearization," *IEEE Trans. Microw. Theory Techn.*, vol. 66, no. 9, pp. 3958–3965, Sep. 2018.

[14] C. Crespo-Cadenas, M. J. Madero-Ayora, J. Reina-Tosina, and J. A. Becerra, "Transmitter linearization adaptable to power-varying operation," *IEEE Trans. Microw. Theory Techn.*, vol. 65, no. 10, pp. 3624–3632, Oct. 2017.

[15] D. Morgan, Z. Ma, J. Kim, M. Zierdt, and J. Pastalan, "A generalized memory polynomial model for digital predistortion of RF power amplifiers," *IEEE Trans. Signal Process.*, vol. 54, no. 10, pp. 3852–3860, Oct. 2006.

[16] M. J. Madero-Ayora, M. Allegue-Martinez, J. A. Garcia, J. Reina-Tosina, and C. Crespo-Cadenas, "Linearization and EVM enhancement of an efficient class J amplifier for 3G and 4G mobile communication signals," in *2012 Workshop on Integrated Nonlinear Microwave and Millimetre-wave Circuits*, Sep. 2012.