A Sparse-Bayesian Approach for the Design of Robust Digital Predistorters Under Power-Varying Operation

Carlos Crespo-Cadenas, *Life Senior Member, IEEE*, María J. Madero-Ayora, *Senior Member, IEEE*, Juan A. Becerra, *Senior Member, IEEE*, and Sergio Cruces, *Senior Member, IEEE*

***Abstract*— In this article, a sparse-Bayesian treatment is proposed to solve the crucial questions posed by power amplifier (PA) and digital predistorter (DPD) modeling. To learn a model, the advanced Bayesian framework includes a group of specific processes that maximize the likelihood of the measured data: regressor pursuit and identification, coefficient estimation, stopping criterion, and regressor deselection. The relevance vector machine (RVM) method is reformulated theoretically to be implemented in complex-valued linear regression. In essence, given an initial set of candidate regressors, the result of this sparse-Bayesian learning approach is the most likely model. Experimental results are provided for the linearization of class AB and class J PAs driven by a 30-MHz fifth-generation new radio signal for a fixed average power, where the evolution of the figures of merit versus the number of active coefficients is examined for the proposed sparse-Bayesian pursuit (SBP) algorithm in comparison to other greedy algorithms. The SBP presents a good performance in terms of linearization capabilities and computational cost. Furthermore, the proposed Bayesian framework enabled the design of a DPD model structure, deselect regressors, and readjust coefficients in a direct learning architecture, demonstrating the robustness to changes in the power level over a 10-dB range.**

***Index Terms*— Behavioral modeling, digital predistortion, nonlinear model identification, power amplifier (PA), Volterra series.**

I. Introduction

MODERN wireless communications front ends are subject to demanding design requirements, generally sharing a common trend toward efficiency. On the one hand, diverse techniques have arisen to strive for better spectral efficiency, being orthogonal frequency-division multiplexing (OFDM) and massive multiple-input/multiple-output (MIMO) technologies among them. The fifth-generation new radio (5G-NR) standard makes use of both. On the other hand, the seek for energy efficiency is also a core objective of 5G. Radio transmitters can constitute a bottleneck in this search for energy efficiency [1], with the power amplifier (PA) being the most critical subsystem in terms of power consumption. Furthermore, the PA output power is traded for linearity. Considering the regulatory requirements to reduce out-of-band emissions, the use of linearization techniques can help to optimize the overall energy efficiency. Digital predistortion is the most widely used form of linearization [2], being nowadays successfully supported by machine learning and data science techniques [3]–[5].

Pioneering papers on PA linearization use the memory polynomial (MP) [6], the generalized memory polynomial (GMP) [7], and the dynamic deviation reduction (DDR) [8] models. The reason of this choice is twofold: first, the high performance provided by beneficial terms, and second, the simpler model structure of these models compared to the huge full Volterra model. Once the digital predistorter (DPD) designer has selected the nonlinear order and the memory length, the number of regressors is fixed, and therefore, the models with fewer regressors are preferred.

A different approach has been presented to prune PA behavioral models and DPDs based on the compressed sampling theory [9], [10]. In this perspective, pursuit algorithms were implemented to identify the active regressors in sparse systems and, unlike the previous approaches, models with a very large and richer set of regressors are more favorable.

Conventional pursuit algorithms proposed to prune the models were orthogonal matching pursuit (OMP) and doubly orthogonal matching pursuit (DOMP) [10], [11], but other approaches are possible. In the case of real-valued sparse systems, the relevance vector machine (RVM) [12]–[14] is an elegant Bayesian technique for regression obtained by linearly weighting a small number of active regressors from a large set of potential candidates. The essence of the RVM is the adoption of a particular strategy for maximization of the marginal likelihood.

The sparse-Bayesian technique in [12] can be applied without modification to the design of a DPD after the complex envelope of the communication signal is transformed into the real domain, as it was proposed by Peng *et al.* [15]. In that

Manuscript received September 15, 2021; revised December 14, 2021 and February 22, 2022; accepted February 23, 2022. This work was supported by Grant TEC2017-82807-P funded by MCIN/AEI/10.13039/501100011033/ and by "ERDF A way of making Europe". The work of Sergio Cruces was supported in part by the Junta de Andalucía through the FEDER-Andalucian Research Project under Grant US-1264994. This article is an expanded version from the 2021 IEEE MTT-S International Wireless Symposium [DOI: 10.1109/IWS52775.2021.9499382]. *(Corresponding author: María J. Madero-Ayora.)*

The authors are with the Departamento de Teoría de la Señal y Comunicaciones, Escuela Técnica Superior de Ingeniería, Universidad de Sevilla, 41092 Seville, Spain (e-mail: mjmadero@us.es).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TMTT.2022.3157586.

Digital Object Identifier 10.1109/TMTT.2022.3157586

paper, they reported a lower efficiency compared to the pruning procedure [10] when the size of the data matrix is small and a computational cost comparable to that technique when the size is large enough. Since work [11] was a later publication, a comparison with the results for this pruning algorithm was not provided.

In this present article, a complex-valued Bayesian treatment for pruning baseband PA models, a limited summary of which was advanced in the conference communication [16], is further developed in Section II. In this article, the real-valued treatment in [12] and the sequential method [14] are adapted to complex-valued linear regression. The proposal includes a sparse-Bayesian pursuit (SBP) based on a simpler algorithm to search the active regressors, speeding up the identification process. The proposed method has been extended with new routines to remove nonactive regressors and reestimate the model coefficients in a scenario of power-level variation. After advancing some experimental results for the behavioral modeling of a PA, Section III is devoted to show the linearization performance achieved by the SBP technique. First, the proposed pursuit approach is compared to the greedy pursuit algorithms OMP and DOMP together with the Bayesian information criterion (BIC) rule, and afterward, the robustness of the reduced complexity DPD model structure that the SBP technique provides is examined against power-level changes. Finally, some concluding remarks are presented in Section IV.

II. Bayesian Approach to PA Modeling

The Volterra series can be used to represent a PA and the complex envelope of the output signal is described with the corresponding discrete-time baseband model. In that case, the PA output is expressed by a linear combination of basis functions, denoted here as Volterra regressors, given by monomials resulting from the multiplication of delayed samples of the input complex envelope $x(k)$ and its conjugate $x^*(k)$. Gathering M samples of the input signal to arrange the column vector $\mathbf{x} = [x(1)\ x(2)\ \cdots\ x(M)]^T$ and defining in a similar way the Volterra regressor vectors $\boldsymbol{\phi}_i$, normalized and arranged in an ordered fashion, the output can be expressed in the matrix form as

$$\begin{aligned}\mathbf{y}_d &= \sum_{i=1}^{N} h_i \boldsymbol{\phi}_i \\ &= \mathbf{X}\mathbf{h}\end{aligned} \tag{1}$$

where N is the number of regressors, h_i are the regression coefficients arranged in the column vector $\mathbf{h}$, and $\mathbf{X} = [\boldsymbol{\phi}_1\ \boldsymbol{\phi}_2\ \cdots\ \boldsymbol{\phi}_N]$ is the $M \times N$ regressors matrix.

In the presence of experimental noise, a given acquisition can be written as

$$\mathbf{y} = \mathbf{X}\mathbf{h} + \mathbf{e} \tag{2}$$

where $\mathbf{y}$ is a column vector with M acquired samples of the output complex envelope. The regression vector $\mathbf{h}$ and the additive noise vector $\mathbf{e}$ are assumed to be drawn from zero-mean random vectors, whose elements have respective variances α_i^{-1}, $i = 1, \ldots, N$, and σ^2.

An oversized regressors stock of the model is a discouraging drawback if the coefficients' estimation is accomplished directly by the least-squares (LS) procedure because it involves the inversion of $\mathbf{X}$ and a high computational cost. However, the richness of the regressor set is advantageous if it is combined with a pursuit method to identify the active regressors of a sparse system.

A suitable pruning procedure was presented in [10] and [11], with the application of greedy algorithms (OMP and DOMP) in the search of the most significant regressors among the complete pool of the Volterra model and the use of the BIC to stop the pursuit of regressors and avoid overfitting.

The present proposal is a Bayesian treatment of a generalized linear model [12] based on the application of the RVM to the problem of PA linearization. The SBP algorithm included in the proposal to pursuit the active regressors can be highlighted and its comparison with the other mentioned greedy algorithms. One of the most popular Bayesian estimators for $\mathbf{h}$ is given by the minimum of the mean square error (mse) solution. This solution is given by the conditional mean of the regression coefficients given the observed measurement vector $\mathbf{y}$. However, the analytical evaluation of this expectation is only known for a few specific distributions, and one of such favorable cases happens when the conditional density is Gaussian. Instead of directly assuming the Gaussian density *ad hoc*, in Appendix I, we theoretically justify its choice as the robust density estimate with the maximum degree of randomness given the current variance estimates of σ^2 and α_i^{-1}, $i = 1, \ldots, N$. This robust density has the following desirable properties.

1) The signal component $\mathbf{y}_d$ and noise component $\mathbf{e}$ of the measurement vector $\mathbf{y}$ are mutually independent.
2) The noise samples are independent and drawn from a complex proper Gaussian density of zero mean and σ^2 variance.
3) The regression coefficients $h_1, \ldots, h_N$ are mutually independent and drawn from complex proper Gaussian densities of zero mean and respective variances α_i^{-1}, $i = 1, \ldots, N$.

These three properties match well with those to be expected from measurements in communications applications. In particular, the elementwise distributions of the measured samples and noise are guaranteed to be circularly symmetric, i.e., invariant under planar rotation. This places a relevant difference with the sparse-Bayesian learning approach mentioned in [15], which relies on the independence of the real and imaginary components of h_i, since this property alone is not sufficient to guarantee the circularity of their distributions. In this sense, the model we consider here seems to capture better relevant properties of the signals involved while requiring the estimation of a smaller number of real parameters. Note that the number of parameters α_i is doubled in [15] due to the decoupling of the real and imaginary components. Therefore, the descriptive efficiency of our proposed representation for this specific scenario of application will, in general, result in a more reliable Bayesian estimation of the regression coefficients.

A. RVM Applied to Complex-Valued PA Modeling

Since the measure of the PA model parameters is different for different experiments, the statistical approach mentioned in [12] is developed by assuming here a joint complex-valued Gaussian distribution for the coefficients $\mathbf{h}$ and the vector of measurements $\mathbf{y}$. Interested readers can find a justification for this assumption in Appendix I. Assuming independence of the samples, the likelihood of the complete dataset can be written as

$$p(\mathbf{y}|\mathbf{h},\sigma^2) = \frac{1}{(\pi\sigma^2)^M} e^{-\frac{1}{\sigma^2}||\mathbf{y}-\mathbf{X}\mathbf{h}||^2} \tag{3}$$

where $\mathbf{y} = [y_1, \ldots, y_M]^T$.

A Bayesian perspective to avoid overfitting is adopted and the parameters are constrained by defining an explicit prior probability distribution over them and adding a complexity penalty term to (3). The likelihood function is complemented by a zero-mean Gaussian prior distribution over the coefficients $\mathbf{h}$

$$p(\mathbf{h}|\boldsymbol{\alpha}) = \prod_{i=1}^{N} \frac{1}{\pi\alpha_i^{-1}} e^{-\alpha_i|h_i|^2} \tag{4}$$

with $\boldsymbol{\alpha}$ a vector of N hyperparameters. Therefore,

$$p(\mathbf{h}|\boldsymbol{\alpha}) = \frac{1}{\pi^N|\mathbf{A}|^{-1}} e^{-\mathbf{h}^H\mathbf{A}\mathbf{h}} \tag{5}$$

where $\mathbf{A} = \mathrm{diag}(\alpha_1, \ldots, \alpha_N)$ is the diagonal precision matrix of the hyperparameters vector $\boldsymbol{\alpha}$. The posterior distribution over the coefficients is deduced from the Bayes rule

$$p(\mathbf{h}|\mathbf{y},\boldsymbol{\alpha},\sigma^2) = \frac{1}{\pi^{N+1}\det\boldsymbol{\Sigma}} e^{-(\mathbf{h}-\boldsymbol{\mu})^H\boldsymbol{\Sigma}^{-1}(\mathbf{h}-\boldsymbol{\mu})} \tag{6}$$

where the posterior precision and mean of the coefficients are, respectively, given by

$$\boldsymbol{\Sigma}^{-1} = \beta\mathbf{X}^H\mathbf{X} + \mathbf{A} \tag{7}$$

and

$$\begin{aligned}\hat{\mathbf{h}} &\equiv \boldsymbol{\mu} \\ &= \beta\boldsymbol{\Sigma}\mathbf{X}^H\mathbf{y}\end{aligned} \tag{8}$$

with $\beta = \sigma^{-2}$. Note that the posterior precision matrix $\boldsymbol{\Sigma}^{-1}$ optimally updates in an additive manner the *a priori* precision $\mathbf{A}$ of the coefficients with the increase in precision $\beta\mathbf{X}^H\mathbf{X}$ provided from the extra information of the measurements. On the one hand, for a reduced number of samples, the posterior precision matrix of the estimate is dominated by the a priori precision $\mathbf{A}$. On the other hand, with a vague prior $\mathbf{A} \to \mathbf{0}$, the posterior precision is dominated by the contribution of the measurements and, therefore, tends to the one of the LS estimator, i.e., $\boldsymbol{\Sigma}^{-1} \to \beta\mathbf{X}^H\mathbf{X}$.

The new formulation of the RVM approach [12] in the case of complex-valued systems involves the maximization over $\boldsymbol{\alpha}$ of the marginal likelihood, which is given by

$$p(\mathbf{y}|\boldsymbol{\alpha},\sigma^2) = \frac{1}{(\pi)^{N_s}|\mathbf{C}|} e^{-\mathbf{y}^H\mathbf{C}^{-1}\mathbf{y}} \tag{9}$$

where

$$\mathbf{C} = \sigma^2\mathbf{I} + \mathbf{X}\mathbf{A}^{-1}\mathbf{X}^H \tag{10}$$

refers to the covariance matrix of the measurement vector.

For convenience, we maximize the logarithm of the marginal likelihood, which gives the objective function

$$\mathcal{L}(\boldsymbol{\alpha}) = -\ln|\sigma^2\mathbf{I} + \mathbf{X}\mathbf{A}^{-1}\mathbf{X}^H| - \mathbf{y}^H(\sigma^2\mathbf{I} + \mathbf{X}\mathbf{A}^{-1}\mathbf{X}^H)^{-1}\mathbf{y}. \tag{11}$$

The strategy of sparse-Bayesian learning is the maximization of $\mathcal{L}(\boldsymbol{\alpha})$ focusing on the contribution of a single hyperparameter α_i. This splitting is possible if the $M \times M$ matrix $\mathbf{C}$ is decomposed as

$$\begin{aligned}\mathbf{C} &= \sigma^2\mathbf{I} + \sum_{m\neq i}\alpha_m^{-1}\boldsymbol{\phi}_m\boldsymbol{\phi}_m^H + \alpha_i^{-1}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^H \\ &= \mathbf{C}_{-i} + \alpha_i^{-1}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^H\end{aligned} \tag{12}$$

where $\mathbf{C}_{-i}$ is $\mathbf{C}$ without the contribution of the regressor $\boldsymbol{\phi}_i$.

For its inversion, one can use the following identity:

$$\mathbf{C}^{-1} = \mathbf{C}_{-i}^{-1} - \frac{\mathbf{C}_{-i}^{-1}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^H\mathbf{C}_{-i}^{-1}}{\alpha_i + \boldsymbol{\phi}_i^H\mathbf{C}_{-i}^{-1}\boldsymbol{\phi}_i} \tag{13}$$

and, through the substitution of (12) and (13) into (11), the objective function $\mathcal{L}(\boldsymbol{\alpha})$ additively decouples as

$$\mathcal{L}(\boldsymbol{\alpha}) = \mathcal{L}(\boldsymbol{\alpha}_{-i}) + \ell(\alpha_i) \tag{14}$$

where $\mathcal{L}(\boldsymbol{\alpha}_{-i})$ denotes the value of the objective function when it is evaluated without the contribution of regressor $\boldsymbol{\phi}_i$ and $\ell(\alpha_i)$ refers to the increment obtained in the objective function due to the incorporation of this regressor. This increment is given by

$$\ell(\alpha_i) = \frac{|q_i|^2}{s_i - \alpha_i} - \ln\left(1 - \frac{s_i}{\alpha_i}\right) \tag{15}$$

where we have used the definitions

$$\begin{aligned}s_i &= \boldsymbol{\phi}_i^H\mathbf{C}_{-i}^{-1}\boldsymbol{\phi}_i \quad \text{and} \\ q_i &= \boldsymbol{\phi}_i^H\mathbf{C}_{-i}^{-1}\mathbf{y}.\end{aligned} \tag{16}$$

The "sparsity factor" s_i is a measure of the extent that regressor $\boldsymbol{\phi}_i$ "overlaps" those already present in the model. The "quality factor" q_i is complex-valued in this approach and its magnitude gives a measure of how well $\boldsymbol{\phi}_i$ increases the marginal likelihood by helping to explain the data. The objective function $\mathcal{L}(\boldsymbol{\alpha})$ has a unique maximum with respect to α_i at

$$\alpha_i = \frac{s_i^2}{|q_i|^2 - s_i}, \quad \text{if } |q_i|^2 > s_i \tag{17}$$

or at $\alpha_i = \infty$ otherwise. During the regressor pursuit, many α_i's tend to infinity, meaning that these coefficients are peaked at zero, i.e., $h_i = 0$ and the corresponding regressors are not included in the active set.

B. Sparse-Bayesian Pursuit

The search procedure of this treatment is a complex-valued reformulation of the learning method in [14] and is an alternative to [15]. The pursuit is initiated with an empty active set of regressors, and therefore, $\mathbf{C}_{-i} = \sigma^2$ in (12). Initially, the potential set is the full stock of N potential

regressors and σ^2 is set to some sensible value, for example, $\sigma^2 = (1/M)\|\mathbf{y}\|^2 \times 10^{-6}$.

The values of s_i and q_i are computed and the potential regressor $\boldsymbol{\phi}_i$ that maximizes $\mathcal{L}(\boldsymbol{\alpha})$ or, equivalently, maximizes

$$\ell(\alpha_i) = \frac{|q_i|^2 - s_i}{s_i} + \ln \frac{s_i}{|q_i|^2} \tag{18}$$

is incorporated to the set of active regressors.

Notice that this step is equivalent to selecting the regressor with the greatest projection $|\boldsymbol{\phi}_i^H \mathbf{y}|$, as in the greedy pursuits, but unlike OMP and DOMP, this procedure includes an additional condition so that the potential regressors not satisfying $|\boldsymbol{\phi}_i^H \mathbf{y}|^2 > \sigma^2$ in (17), i.e. with projection below the noise level, are deleted with $\alpha_i = \infty$.

After the first pursuit iteration, the active set is increased by one regressor with a new computed α_i and the number of potential regressors is reduced. Likewise, the posterior covariance $\boldsymbol{\Sigma}$ and mean $\boldsymbol{\mu}$ of the coefficients, which are scalars in this first iteration, are computed along with the updated values of s_i and q_i for all potential regressors.

Repeating these operations, in each successive iteration, the SBP retrieves the regressor that maximizes the marginal likelihood, and the posterior covariance and mean of the coefficients are updated using the following formulas:

$$\tilde{\boldsymbol{\Sigma}} = \begin{bmatrix} \boldsymbol{\Sigma} + \beta^2 \boldsymbol{\Sigma}_{ii} \boldsymbol{\Sigma} \mathbf{X}^H \boldsymbol{\phi}_i \boldsymbol{\phi}_i^H \mathbf{X} \boldsymbol{\Sigma} & -\beta \boldsymbol{\Sigma}_{ii} \boldsymbol{\Sigma} \mathbf{X}^H \boldsymbol{\phi}_i \\ -\beta \boldsymbol{\Sigma}_{ii} (\boldsymbol{\Sigma} \mathbf{X}^H \boldsymbol{\phi}_i)^H & \boldsymbol{\Sigma}_{ii} \end{bmatrix} \tag{19}$$

$$\tilde{\boldsymbol{\mu}} = \begin{bmatrix} \boldsymbol{\mu} - \mu_i \beta \boldsymbol{\Sigma} \mathbf{X}^H \boldsymbol{\phi}_i \\ \mu_i \end{bmatrix} \tag{20}$$

where $\boldsymbol{\Sigma}_{ii} = (\alpha_i + s_i)^{-1}$, $\mu_i = \boldsymbol{\Sigma}_{ii} q_i$, and $\mathbf{e}_i = \boldsymbol{\phi}_i - \beta \mathbf{X} \boldsymbol{\Sigma} \mathbf{X}^H \boldsymbol{\phi}_i$. The active set is increased until the candidate regressor set is exhausted. The pseudocode of this SBP technique is shown in Algorithm 1.

Enunciated in terms of the posterior coefficient covariance $\boldsymbol{\Sigma} = (\beta \mathbf{X}^H \mathbf{X} + \mathbf{A})^{-1}$, the first benefit of the method is a formulation understood as a regularization procedure. Given that $\boldsymbol{\Sigma}$ is an $N \times N$ matrix, a second benefit is observed in terms of memory requirements compared to the direct use of the $M \times N$ regressors matrix $\mathbf{X}$ in an overdetermined equation system ($M > N$). Finally, the coefficients are estimated with a sequential algorithm avoiding the expensive computation of the regressor matrix pseudoinverse.

To illustrate the SBP performance, the learning curves in the identification process of a class AB PA are plotted in Fig. 1 for two GMP models of fifth and ninth orders. The experimental test bench is the same used in [16] and is also described in Section III. The resulting NMSE performance is -52.2 dB with 77 active regressors, identified from the unreduced set of 231 candidate regressors of the fifth-order model, and -53.6 dB with 106 active regressors, identified from the set of 451 regressors of the ninth-order model (red circles). In the case of NMSE performance better than the objective, it is possible a further reduction of the active set. For example, if a target NMSE of -51 dB is allowed, the SBP algorithm can reduce further the number of active regressors to 19 and 16, as the red asterisks shown in Fig. 1.

The regressors kept in the reduced ninth-order model are ordered by decreasing likelihood in Table I. The type of regressor and the associated NMSE improvement are also displayed. The results of the fifth-order model are related, except that regressors with higher nonlinear order, such as those numbered 13 and 16, are missing. The NMSE worsens slightly and the algorithm tries to compensate for the missing regressors by adding less effective ones of lower nonlinear order, thus demanding a total of 19 regressors. Notice that the ninth-order GMP model with a richer initial set yields a better NMSE with only 16 coefficients compared to the fifth-order GMP model, which needs 19 coefficients.

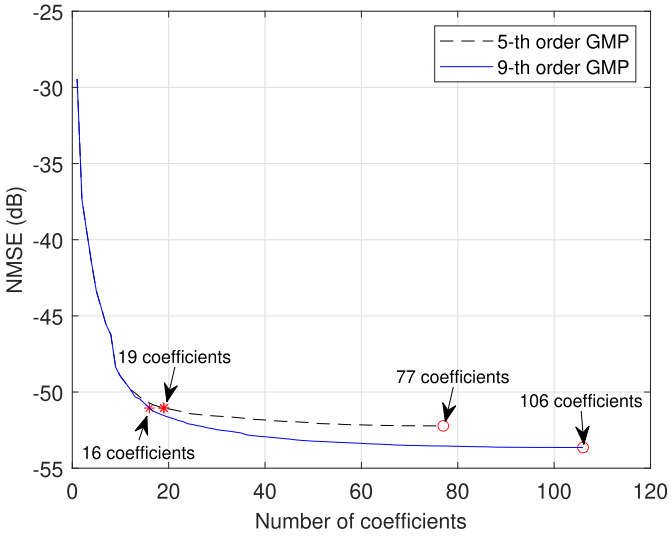


Fig. 1. Performance of the SBP identification procedure using two GMP models with fifth- and ninth-nonlinear orders. NMSE versus number of coefficients for $P_o = 27.4$ dBm. The full active sets (circles) and the active sets that guarantee an NMSE $= -51$ dB (asterisks) are highlighted.

Algorithm 1 SBP

Input: $\mathbf{X} \in \mathbb{C}^{M \times N}$, $\mathbf{y} \in \mathbb{C}^M$

Output: $\mathbf{h} \in \mathbb{C}^{N_a}$, $N_a \ll N$

1: *Initialization*: Initialize σ^2 to some sensible value (for example, $\sigma^2 = \frac{1}{M}\|\mathbf{y}\|^2 \times 10^{-6}$), $\mathbf{C}_{-i}^{(0)} \leftarrow \sigma^2$

2: Compute with (16) the values of s_i and q_i for all potential regressors $\boldsymbol{\phi}_i$, and remove from the potential set the regressors that do not fulfil the requirement $|q_i|^2 > s_i$.

3: Compute α_i and $\ell(\alpha_i)$ using (17) and (18). Move the potential regressor that maximizes $\ell(\alpha_i)$ to the active set.

4: Compute with (19) and (20) the updated values of $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ (which are scalars initially). Update $\mathbf{C}_{-i}$ along with s_i and q_i for all potential regressors.

5: Go to 2 until the *stopping criterion* is met or the potential set is empty.

C. Deselecting Regressors

Once the active set has been identified, it is reasonable to question if all regressors remain actually active or if some regressors can be deselected otherwise, after new regressors have been incorporated to the active set. The procedure to test this active set can be initiated with the N_a active regressors

TABLE I

GMP(9, 10, 2) MODEL INITIAL SET: 451 REGRESSORS. REDUCED SET: 16 REGRESSORS. TARGET NMSE: −51 dB

Likelihood order	Regressor	NMSE (dB)
1	$x(n)$	-29.5
2	$x(n)\|x(n)\|$	-37.3
3	$x(n-1)\|x(n)\|^4$	-39.5
4	$x(n-4)$	-41.6
5	$x(n-1)$	-43.6
6	$x(n-2)$	-44.5
7	$x(n-7)$	-45.4
8	$x(n)\|x(n)\|^3$	-46.1
9	$x(n-1)\|x(n)\|^2$	-48.3
10	$x(n)\|x(n)\|^2$	-49.2
11	$x(n)\|x(n)\|^4$	-49.5
12	$x(n)\|x(n-1)\|$	-49.9
13	$x(n)\|x(n)\|^8$	-50.2
14	$x(n)\|x(n+1)\|$	-50.8
15	$x(n)\|x(n)\|^6$	-50.9
16	$x(n)\|x(n+2)\|$	-51.1

and the corresponding matrices $\mathbf{X}_0 \in \mathbb{C}^{M\times N_a}$, $\mathbf{A}_0 \in \mathbb{C}^{N_a\times N_a}$, observing that

$$\begin{aligned}\mathbf{C}_0 &= \sigma^2\mathbf{I} + \mathbf{X}_0\mathbf{A}_0^{-1}\mathbf{X}_0^H \\ &= \sigma^2\mathbf{I} + \sum_{n=1}^{N_a}\alpha_n^{-1}\boldsymbol{\phi}_n\boldsymbol{\phi}_n^H. \end{aligned} \tag{21}$$

Deselection of the regressor $\boldsymbol{\phi}_i$ yields

$$\mathbf{C} = \sigma^2\mathbf{I} + \sum_{n\neq i}\alpha_n^{-1}\boldsymbol{\phi}_n\boldsymbol{\phi}_n^H \tag{22}$$

or

$$\mathbf{C} = \mathbf{C}_0 - \alpha_i^{-1}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^H \tag{23}$$

where $\mathbf{C}$ is $\mathbf{C}_0$ with the contribution of the regressor i removed (and also the dependence on α_i). Proceeding in a similar way as in the SBP, it is possible to write the objective function after deselection $\mathcal{L}(\boldsymbol{\alpha})$ as

$$\mathcal{L}(\boldsymbol{\alpha}) = \mathcal{L}(\boldsymbol{\alpha}_0) + \ell(\alpha_i) \tag{24}$$

where $\mathcal{L}(\boldsymbol{\alpha}_0)$ is the objective function before deselection of regressor ϕ_i, and the increase of the objective function after deselection is

$$\ell(\alpha_i) = \frac{|q_i|^2}{s_i - \alpha_i} - \ln\left(1 - \frac{s_i}{\alpha_i}\right). \tag{25}$$

The regressor that maximizes the objective function (or the marginal likelihood) after deselection is discarded from the active set, and the posterior covariance and mean of the coefficients are updated using the formulas

$$\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} - \frac{1}{\boldsymbol{\Sigma}_{jj}}\boldsymbol{\Sigma}_j\boldsymbol{\Sigma}_j^H \tag{26}$$

$$\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu} - \frac{\mu_j}{\boldsymbol{\Sigma}_{jj}}\boldsymbol{\Sigma}_j \tag{27}$$

and the appropriate row and/or column j is removed from $\tilde{\boldsymbol{\Sigma}}$ and $\tilde{\boldsymbol{\mu}}$. $\boldsymbol{\Sigma}_j$ is the jth column of $\boldsymbol{\Sigma}$.

Note that deselection is not contemplated in standard greedy pursuit techniques. However, it is a beneficial algorithm of this Bayesian treatment that enables to check whether any regressor of the set selected by the SBP algorithm is actually active. Referring to Fig. 1, the deselection procedure was carried out for the active sets of the two GMP models, 77 and 106 coefficients, with the result that no regressors had to be removed if the respective accuracy must be preserved. This is an indication that each regressor selected by the SBP algorithm is consistently active even if new candidate regressors are included in the set. However, some circumstances can be foreseen for which deselection of regressors would be convenient to reconfigure the pruned model. One possible example is a PA suffering a bias point modification leading to a less nonlinear operation and a consequent application of the deselection subroutine would be positive to remove higher order regressors. In this article, we focus on a different situation in which the power level of the input signal varies, but not the system, i.e., the PA. Therefore, the more likely model learned with the input signal at a given power level is also valid for other points with lower level and it is only necessary for the coefficients' reestimation. The details of model reconfiguration in this context are discussed in the following.

D. Reconfiguring the Model in a Power-Varying Scenario

The PA was tested with the input signal power covering an output dynamic range from 20.9 to 33.8 dBm. Taking into account the benefits of a richer initial model as the power level increases, previously illustrated for a given power level, the SBP algorithm was repeated with the more complete set of candidate regressors of the bivariate circuit-knowledge-based Volterra (bi-CKV) model [19]. By way of explanation, the pruned fifth-order bi-CKV model, optimized also with 19 regressors for 27.4 dBm, was applied directly at other power levels without readjustment of the coefficient values, and the NMSE results (not shown) degraded dramatically, as it could be expected.

A straightforward solution to this issue is to perform again the complete model identification in a point-by-point basis, i.e., to repeat the search of the active regressors and the estimation of the coefficients. Considering that for a scenario with changing power levels, the number of identified regressors would be different, the SBP algorithm has been modified to halt when an objective NMSE is reached. When a target NMSE of −51 dB is selected as the stopping criterion in the SBP algorithm, a large active set of 171 regressors is necessary for the highest level, as the dotted line with circles shown in Fig. 2. These results also demonstrate a viable further reduction of the number of regressors for lower power levels, e.g., only five regressors are necessary for the lowest level.

Since 19 regressors have been already determined at 27.4 dBm, a computationally more efficient procedure is the reuse of these regressors, saving the searching steps, and repeating only the coefficients' reestimation subroutine for the corresponding 19 coefficients. The NMSE performance with the reestimated coefficients is indicated in the same figure (blue line with dots).

The efficiency improvement offered by the reestimation subroutine with respect to the full SBP algorithm can be

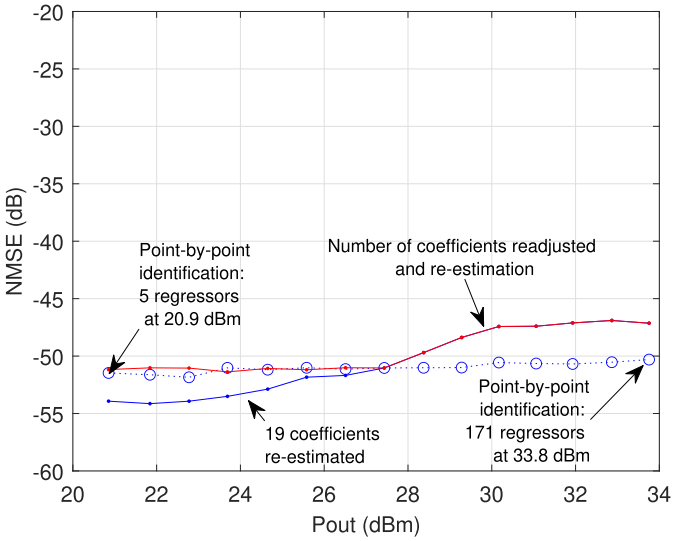


Fig. 2. NMSE without and with the reestimation procedure applied to the 19 regressors model identified at $P_o = 27.4$ dBm. SBP applied point-by-point (circles), reestimation of the 19 coefficients (blue line), and readjustment and reestimation for NMSE $= -51$ dB (red line).

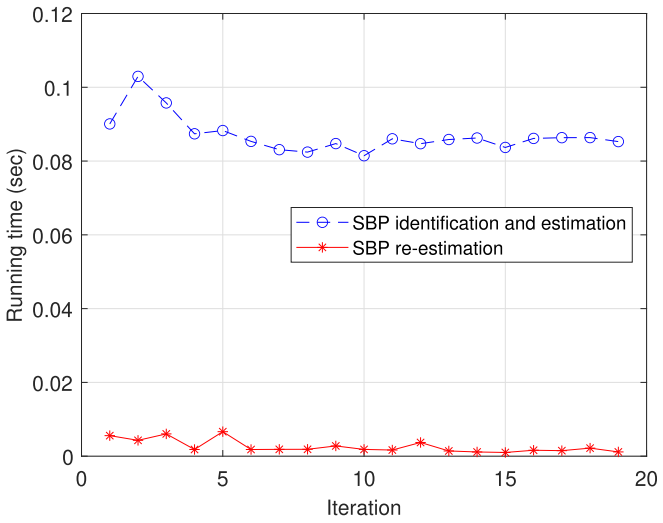


Fig. 3. Comparison of the execution time for the complete identification and reestimation procedures.

evaluated by comparing the respective execution times of each iteration shown in Fig. 3. The reestimation (line with asterisks) is nearly 80 times faster than the complete SBP with active regressor selection and coefficient estimation (dashed line with circles).

Considering that the 19 regressors' set is oversized at lower levels, a further reduction of the active set is achieved if the reestimation subroutine of the SBP algorithm is performed only for the regressors necessary to comply with the -51-dB NMSE objective (red line). A possible procedure can be summarized as follows. Initially, use the SBP to identify the active regressors and estimate the coefficients at the maximum power level. Since the regressors are arranged according to its likelihood order, they can be orderly deselected if the power decreases and the coefficients reestimated.

Finally, above 27.4 dBm, the objective NMSE of -51 dB is not reached for the highest power, even if the number of

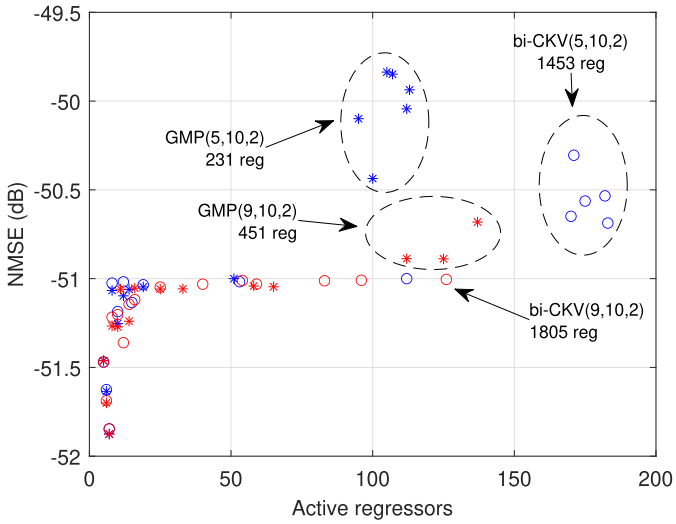


Fig. 4. Graph of NMSE versus number of active regressors for the reconfigured models. The encircled points corresponding to the higher power levels do not comply with the target NMSE.

regressors is increased to 171. This leads to reconsidering the option of a more favorable initial model.

E. Initial Model Selection

The above discussion has illustrated the advantage of a more complete initial model, mainly as the power level increases. For that reason, a comparison of four models is presented: the fifth- and ninth-order GMP models and two bi-CKV models of fifth and ninth orders. In all cases, the primary memory length is 10 and the secondary length is 2. Instead of a traditional representation of the NMSE for a power sweep, the results are plotted in an NMSE-N_a plane to facilitate comparison. The pruned model showing the best NMSE with the minimum number of active regressors is finally selected. The point-by-point results in the output power range from 20.9 to 33.8 dBm are shown in Fig. 4 for each model. In the four cases, the identified active sets are small for the lower powers (the left-hand side in the figure) and become larger as the level increases.

Although the most simple fifth-order GMP model seems advantageous because of its initial set of 231 candidate regressors, it is unable to comply with the objective NMSE for the points of maximum power (encircled blue asterisks). The use of a more complex model, fifth-order bi-CVK with 1453 candidate regressors, improves slightly the NMSE to a value of about -50.5 dB, but it is not sufficient to comply with the objective at the highest powers. This is a consequence of the need for a higher nonlinear order model, as demonstrated in the results of the ninth-order GMP model. In this case, the initial set contains 451 candidate regressors and the accuracy is further improved to about -50.8 dB due to the new higher order regressors. Finally, the ninth-order bi-CKV model with 1805 candidate regressors is able to fulfill the objective NMSE in the whole dynamic range with 126 active regressors.

The results of the figure indicate that a good selection is the ninth-order bi-CKV model to comply with the target NMSE at all power levels and suggests an adaptive procedure to gain further model reduction under low-level PA operation. In this

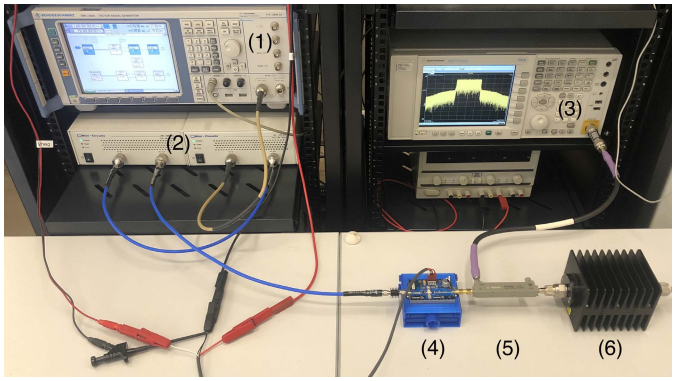


Fig. 5. Photograph of the experimental setup composed of (1) a VSG, (2) two cascaded preamplifiers, (3) a VSA, (4) the PA under test, (5) a directional coupler, and (6) an attenuator terminated with a 50-Ω load.

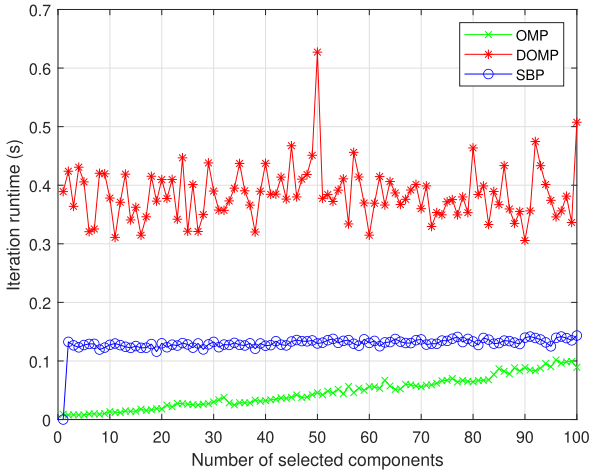


Fig. 6. Iteration runtime for the analysis of a model with 800 components. The implementation of the DOMP algorithm is the RC-DOMP.

way, once the 126 active regressors have been identified at the maximum power, it is possible to reshape the PA model after a variation of the average output power. Since this is a model with sufficient richness and recalling that each regressor is ordered according to its likelihood, the coefficients can be reestimated orderly until the NMSE objective is achieved. In that form, the number of regressors is optimized for each power level.

For similar reasons, this method can also help in the reconfiguration of DPDs in the case of power-varying operation. In a DPD based on indirect learning architecture (ILA), the application is immediately following the procedure presented. In a DPD based on direct learning architecture (DLA), the application is not as direct because the coefficients are readjusted by the DLA procedure. Notwithstanding, the more likely regressors selected at the highest operation level can be reused at lower power points and the active set can be reduced by orderly removing the last regressors, that is, the less likely regressors. This method is implemented in Section III.

III. Linearization Procedures

To illustrate the features of the proposed Bayesian treatment in comparison with other regressor selection approaches, the linearization of two different types of PAs is shown in this section: a commercial class AB PA and a highly efficient class J PA that exhibit a complicated nonlinear characteristic with gain expansion and compression.

A. Experimental Setup

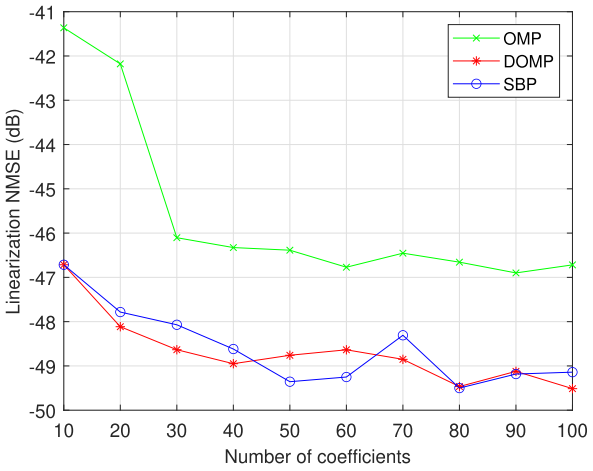
The experimental setup employed in this work is shown in Fig. 5. The probing signal was created by an SMU200A vector signal generator (VSG) from Rohde & Schwarz with built-in arbitrary waveform generator. It allowed a maximum sampling rate of 100 MS/s, thus imposing a limitation to the maximum signal bandwidth that could be linearized with an oversampling factor of 3. Considering that limitation, the probing signal was designed according to the 5G-NR standard with 30-MHz bandwidth, 30-kHz subcarrier spacing, 16-QAM symbols over all the subcarriers, and a peak-to-average power ratio (PAPR) of 10.5 dB. Although this bandwidth is in the low part of the channel values of the 5G-NR standard, this probing signal is valid to illustrate the features of the Bayesian pursuit procedure.

The VSG was followed by two cascaded Mini-Circuits TVA-4W-422A+ preamplifiers to drive the PA under test in a mildly nonlinear operation while keeping the behavior of the modulator sufficiently linear. The first PA under test was the evaluation board of a class AB amplifier based on the CGH40010 GaN HEMT from Cree Inc., operated with a drain-to-source current of about 200 mA at 3.6 GHz. The second PA under test was a continuous-mode class J amplifier based on the CGH35030F GaN HEMT from Cree, which was designed for an operation frequency of 850 MHz. The output of the PA under test for each case was fed to a PXA-N9030A vector signal analyzer (VSA) from Keysight Technologies through a directional coupler and an attenuator to avoid introducing undesired distortion from the equipment. The RF output signal was then downconverted to baseband and acquired in the VSA with a sampling frequency of 92.16 MS/s, where the measurement dynamic range was optimized by averaging 300 repetitions of the measured signal. Finally, the signals were time-aligned to synchronize the input and output datasets.

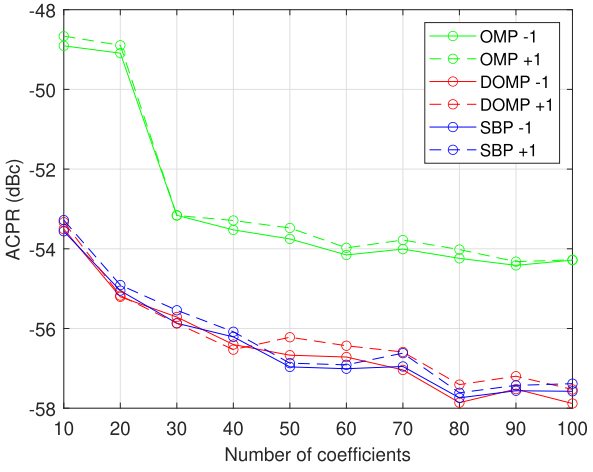
B. Comparative Assessment of the SBP Versus Greedy Algorithms for a Class AB PA

As a first experiment, the pursuit algorithm of the Bayesian treatment proposed in this article will be contrasted with the OMP and DOMP techniques for the linearization of the commercial class AB PA under test with an average output power of 30.1 dBm and a gain compression of 2.34 dB. Comparison with these techniques is justified because the OMP is a reference in [15] and the DOMP has demonstrated superior performance compared to other algorithms [20].

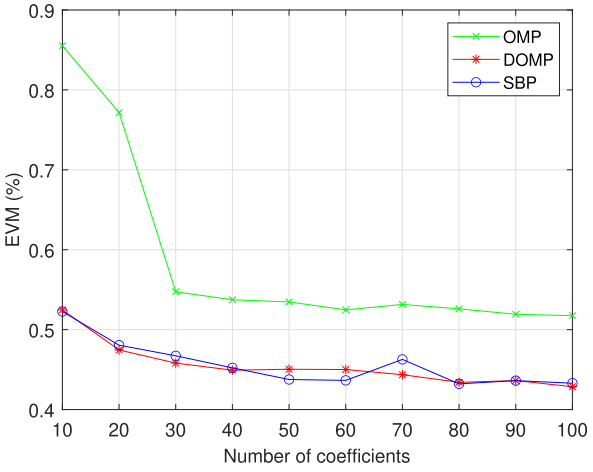
A bi-CKV model was selected to implement the DPD. The structure, with seventh-order and maximum memory depth of seven taps, provides an initial full stock of 812 regressors. The three algorithms under assessment were executed over this model to reduce its complexity by keeping only the most relevant regressors in the active set. A segment of $M = 1843$



(a)



(b)



(c)

Fig. 7. (a) Linearization NMSE; (b) upper and lower ACPR, denoted as $+1$ and -1, respectively; and (c) EVM achieved by the techniques under comparison in a sweep of number of active coefficients for the class AB PA under test.

samples of the input–output dataset was employed for the identification of the reduced-order model structure.

The computational cost of the three algorithms was compared by evaluating the runtime of each iteration, as shown in Fig. 6 for a typical experiment with 10 000 samples and 800 components. In agreement with the runtime figures

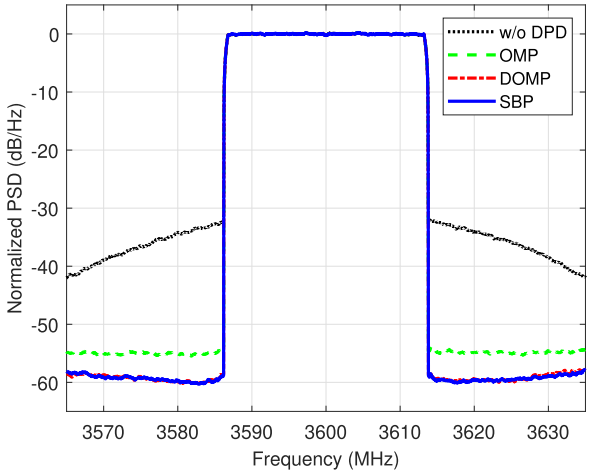


Fig. 8. PSDs of the output signal without DPD and the linearized class AB PA output with 100 coefficients attained by the OMP, DOMP, and SBP techniques. $P_o = 30.1$ dBm.

presented in [16], these results demonstrate the total execution time for the greedy pursuit algorithms, OMP, and the reduced complexity implementation of the DOMP, as well as for the pursuit algorithm of the present Bayesian treatment. Although this issue may vary for other possible scenarios, the OMP algorithm provides a lower running time than the present proposal with a poorer accuracy, and the SBP algorithm outperforms the DOMP in running time with the same accuracy.

Once the SBP has been completed, the result is a set with the most likely active regressors. Then, the DPD is implemented through DLA [21], where 30 iterations with a step size of $1/4$ were executed for training. Afterward, the DPD coefficients were fixed and a different validation signal was applied, randomly generated with different binary data.

The outcomes of Fig. 7 show the evolution of the linearization figures of merit as the number of selected coefficients increases following the order determined by each algorithm. These figures of merit are the linearization NMSE, the lower and upper adjacent channel power ratio (ACPR), and the error vector magnitude (EVM). For comparison, the characteristics of the OMP and DOMP algorithms are also plotted. As it can be observed, the SBP and DOMP algorithms achieve a fine performance with modeling capabilities for a reduced number of coefficients. Based on the presented results for the runtime and linearization performance, the SBP approach offers a reliable procedure for DPD design. The superior reduction of spectral regrowth produced by the SBP and DOMP algorithms can also be corroborated in Fig. 8, where the normalized power spectral densities (PSDs) at the output of the PA are presented without DPD and linearized with the pruned model structures. Results for the OMP-based DPD are also shown.

Observe that if an ACPR goal of -55 dBc is set, Fig. 7(b) shows the feasibility of a DPD with 20 coefficients to deliver the required ACPR at a power level of 30.1 dBm. This pruned set means a significant reduction in model complexity compared with the full 100 coefficients model. According to the discussion in Section II, an oversized DPD with regressors identified by the SBP algorithm at a given output power can be reconfigured by applying the reestimating algorithm to shorten

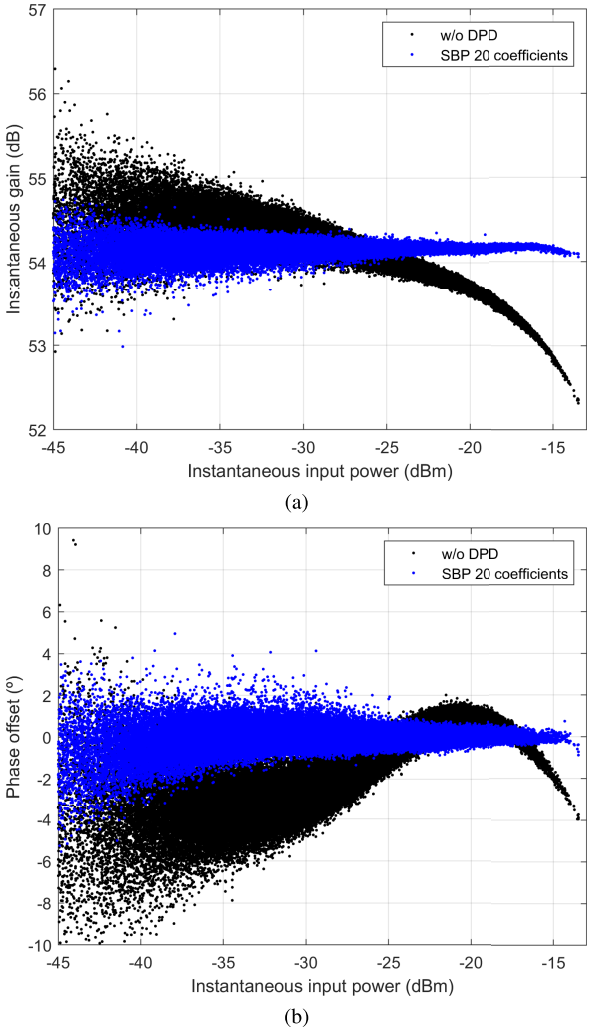


Fig. 9. (a) AM/AM and (b) AM/PM characteristics of the unlinearized class AB PA and with a DPD of 20 coefficients attained through SBP. $P_o = 30.1$ dBm.

the active set. The AM/AM and AM/PM characteristics shown in blue in Fig. 9 also show that, even reducing the number of coefficients of the model to 20, the algorithm is able to linearize the PA output.

C. Comparative Assessment of the SBP Approach Versus Greedy Algorithms for a Class J PA

A second experiment with a different PA under test was performed to show that the feature selection capabilities of the SBP approach are independent of the type of the PA to be linearized. In this case, the class J amplifier under test was employed with an average output power of 31.1 dBm, exhibiting a complicated nonlinear characteristic with gain expansion followed by compression. This behavior produces a remarkable nonlinear distortion for the operation point, in which it exhibits an ACPR of −27.9 and −28.9 dBc, an NMSE of −16.5 dB, and an EVM of 14.7%. Again, the linearization performance of the SBP approach will be compared to that of the OMP and DOMP algorithms.

The selected model structure to implement the DPD, in this case, was the same bi-CKV model with the seventh order and

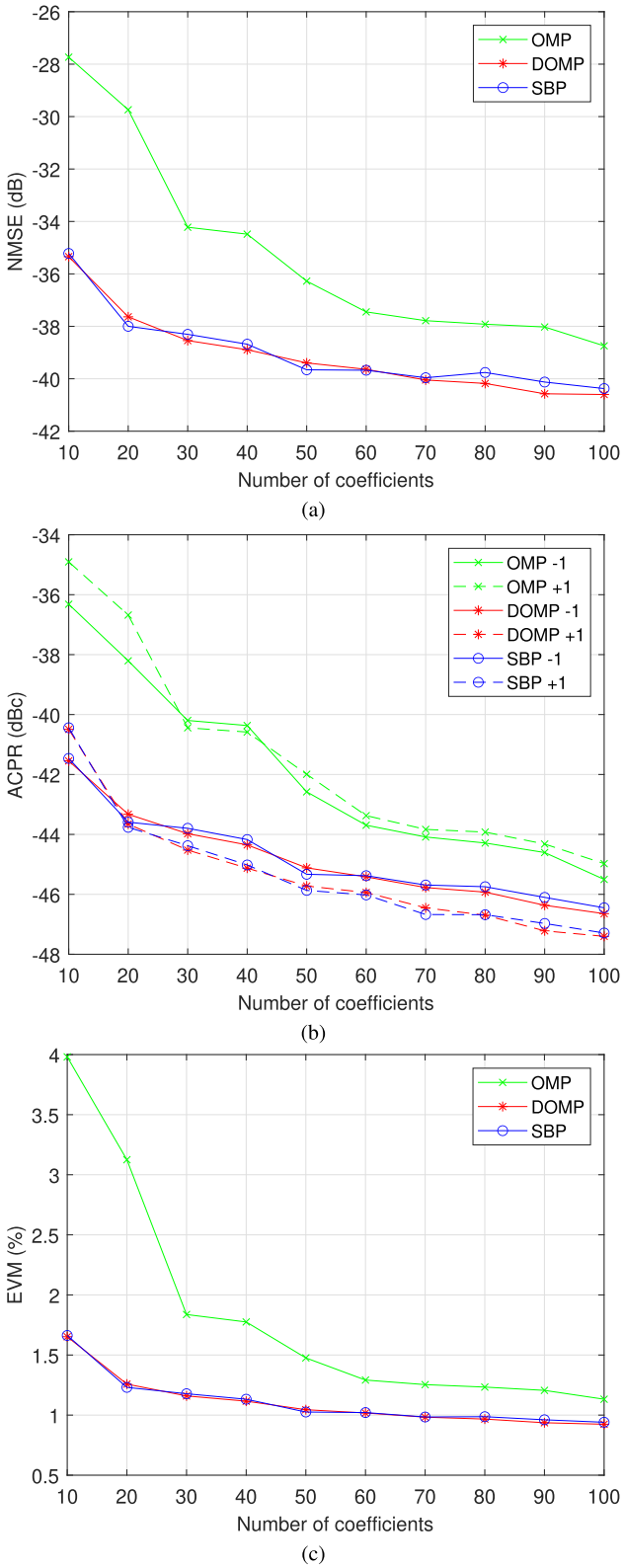


Fig. 10. (a) Linearization NMSE; (b) upper and lower ACPR, denoted as +1 and −1, respectively; and (c) EVM achieved by the techniques under comparison in a sweep of number of active coefficients for the class J PA under test.

memory depth of seven taps as in the previous experiment, providing an initial full stock of 812 regressors. The segment of input–output data employed for the identification of the

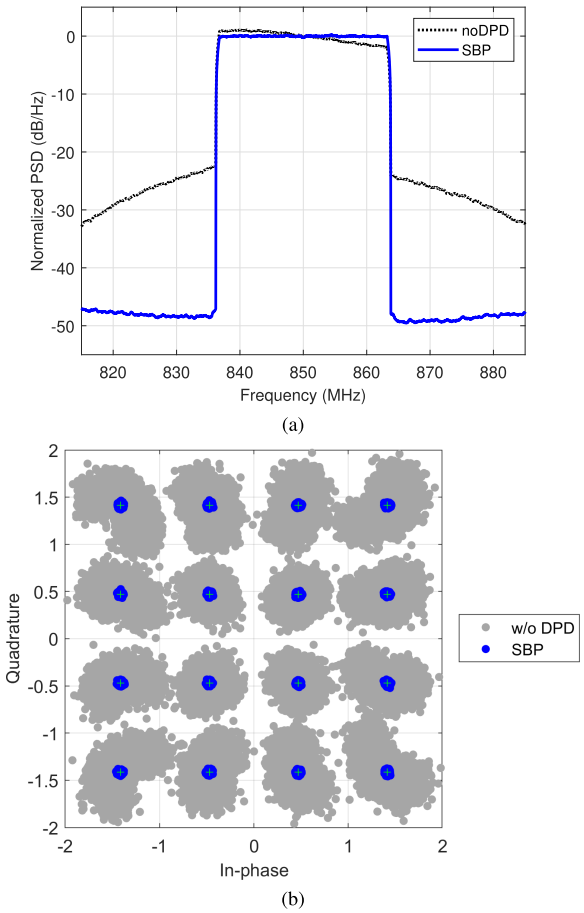


Fig. 11. (a) PSDs and (b) constellations of the output signal without DPD and the linearized class J PA output with 100 coefficients attained by the SBP technique. $P_o = 31.1$ dBm.

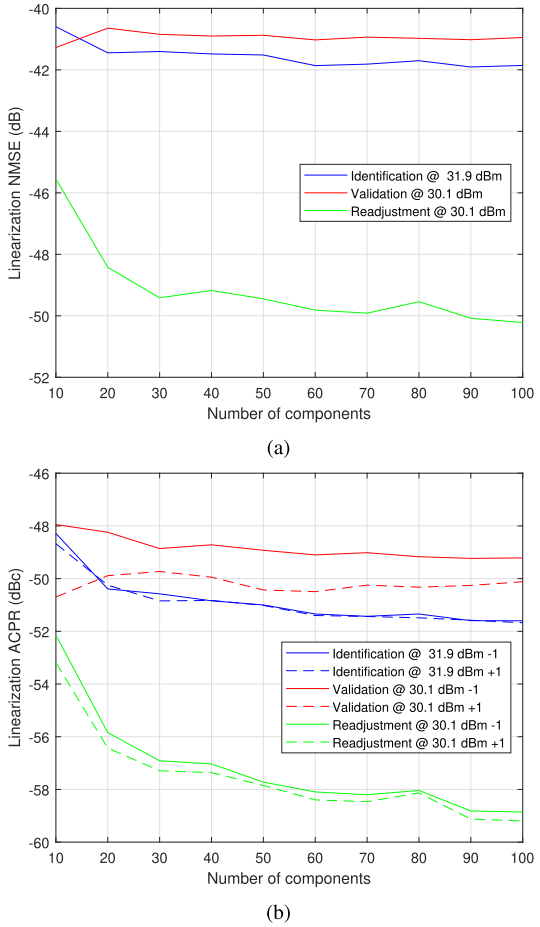


Fig. 12. (a) Linearization NMSE and (b) upper and lower ACPR, denoted as $+1$ and -1, respectively, in a sweep of number of coefficients identified with the SBP at 31.9 dBm of output power (blue line), validating at 30.1 dBm (red line) and readjusting the coefficients at 30.1 dBm (green line).

most relevant regressors by means of the three algorithms under assessment contained $M = 92\,160$ samples. Again, the DPD was implemented through a DLA scheme with 30 iterations and a step size of $1/4$. Then, the DPD coefficients were fixed and the linearization performance was validated with a different signal.

As it is reasonable to expect, the evolution of the linearization figures of merit NMSE, ACPR, and EVM for the class J PA provided in Fig. 10 is similar to the previously explained one for the class AB PA. Due to the stronger nonlinear behavior of the class J amplifier, the values achieved by the three approaches for a large enough number of coefficients show a lower degree of linearization than for the class AB PA. The SBP approach still achieves almost coincident values of NMSE, ACPR, and EVM than the DOMP algorithm, with superior performance than the OMP approach.

The more complicated nonlinear behavior in the case of the class J PA without DPD can also be appreciated in the normalized PSD and constellation at the output of the PA shown in Fig. 11. As it can be observed, in addition to a notable spectral regrowth, the PA output exhibits a nonflat in-band spectrum and the corresponding distortion of the constellation diagram of the received 16-QAM symbols. These impairments are successfully compensated when the DPD is applied with the pruned model structure provided by the SBP approach with 100 coefficients.

D. Robustness Against Power-Level Changes

With the aim of highlighting the robustness of the attained SBP structures with respect to power-level changes, the following experiment was performed with the class AB PA under test employed in Section III-B. First, the 100 most likely regressors of the same model as the one used in the previous experiment were selected at 31.9 dBm of output power. These regressors were used to implement a DPD, and through ten DLA iterations with a step size of $1/2$, the resulting ACPR is shown in Fig. 12(b) (blue line).

Thereafter, the input level was reduced 3 dB and, due to the gain compression, the average output power decreases to 30.1 dBm. The linearization performance was measured without changing the parameters of the DPD, and the results were plotted in the same figure (red line). With the unchanged DPD, the new power level worsens the linearization performance. Finally, ten DLA iterations of the DPD with the same regressor structure were performed at 30.1 dBm of output power obtaining a significant NMSE and ACPR improvement, as it is shown in Fig. 12 (green lines). Observe that, once the DPD has been readjusted, the linearization performance

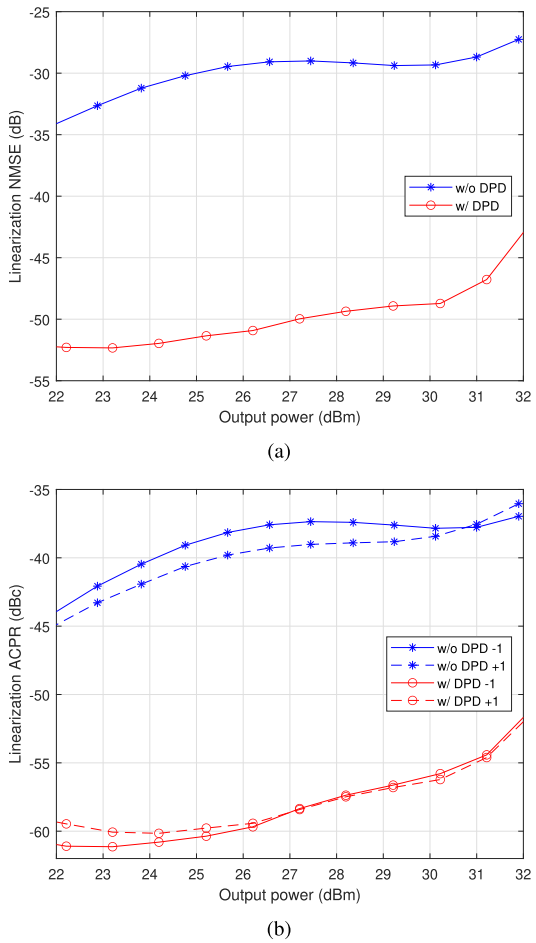


Fig. 13. (a) Linearization NMSE and (b) upper and lower ACPR, denoted as $+1$ and -1, respectively, in a power sweep without DPD and with the DPD that resulted in the analysis of the highest power level (30 most likely regressors).

of the 100-regressor structure identified at 31.9 dBm keeps up after the power-level change. Since this is a decrease of the power level, the ACPR improvement is not surprising, but this outcome also proves the feasibility of reducing orderly the same identified set of most likely regressors to a number that guarantees a given objective ACPR. For example, a model with the first 30 most likely regressors delivers an ACPR of about -57 dBc, keeping up the linearization performance after this power-level change.

As a second test for showcasing the power-level change characteristics of the proposal, the first 30 coefficients of the most likely structure identified at the highest power were trained over a power sweep of 10 dB with a step size of 1 dB, corresponding to an output power-level range from 22 to 32 dBm. Fig. 13 shows the linearization NMSE and ACPR in this power sweep. No new identification of the regressors was necessary after each power change.

The steps of the DPD reshape procedure would be as follows. Identify first the active set of regressors with the SBP algorithm at the highest operation level and adjust the coefficients with the DLA scheme. Under power reduction, the total set of active regressors is suitable to perform with an ACPR much better than necessary. Therefore, a computationally more efficient approach is to orderly remove the least likely regressors until the ACPR target is obtained and then perform the iterations of the DLA scheme. Notice that a second identification search is not necessary.

IV. Conclusion

The DPD approach presented in this work is based on a complex-valued Bayesian treatment, which includes a fast pursuit algorithm to select active regressors. Previously reported techniques are adapted to complex-valued linear regression and a detailed exposition of the statistical background of the proposal is demonstrated. The present approach involves procedures not only for the search of active regressors but also for removing those not relevant. Based on the demonstrated formal deduction, the proposed SBP algorithm joins a pursuit technique with a Bayesian stopping criterion to select the active regressor set following a marginal likelihood maximization order. In this sense, we can say that, given an initial set of candidate regressors, the selected set is the most likely reduced model. It is remarkable that the sparse-Bayesian procedure contains in its nature the valuable benefits of regularization avoiding the regressor matrix pseudoinverse calculation. The evolution of the linearization figures of merit versus the number of active coefficients is examined for the SBP algorithm in comparison to other accepted greedy algorithms. In addition to presenting a reduced computational cost with respect to the reduced complexity DOMP and outperforming the OMP algorithms in modeling and linearization capabilities, this Bayesian treatment includes a stopping criterion and a regressor deselection procedure. An appreciated feature of this Bayesian approach is the possibility that it offers to optimize the DPD performance by reestimating the coefficients after a power-level increase in a DPD with a fixed number of coefficients and by removing regressors of the active set for the case of power-level lowering. It makes possible to realize design techniques for DPDs that are reconfigurable under power-varying operation. Experimental results demonstrate the DPD performance in the linearization of PAs driven by a 5G-NR signal and its robustness to changes in the power level over a 10-dB range.

Appendix I
Model Hypotheses and Bayesian Estimators

In this appendix, we will resort to the maximum entropy principle of Jaynes [22] for the justification of the model hypotheses adopted in this work. These are summarized by the next lemma.

Assumption 1 (A1): We model the joint density of the coefficients and measurements as a proper complex Gaussian density that factorizes as

$$p(\mathbf{h}, \mathbf{y}|\boldsymbol{\alpha}, \sigma^2) = p(\mathbf{y}|\mathbf{h}, \sigma^2)\, p(\mathbf{h}|\boldsymbol{\alpha}) \tag{28}$$

where the likelihood of the measurements

$$\begin{aligned} p(\mathbf{y}|\mathbf{h}, \sigma^2) &= \frac{1}{(\pi\sigma^2)^M} e^{-\frac{1}{\sigma^2}||\mathbf{y}-\mathbf{X}\mathbf{h}||^2} \\ &\equiv \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I}) \end{aligned} \tag{29}$$

and the prior distribution

$$p(\mathbf{h}|\boldsymbol{\alpha}) = \frac{1}{\pi^N |\mathbf{A}|^{-1}} e^{-\mathbf{h}^H \mathbf{A} \mathbf{h}} \equiv \mathcal{CN}(\mathbf{0}, \mathbf{A}^{-1}) \tag{30}$$

are both complex Gaussian densities and proper (with circularly symmetric elements), that is, they have null pseudocovariance matrices, where, by definition, the pseudocovariance matrix of a complex random vector $\mathbf{z}$ of mean $\bar{\mathbf{z}}$ is $E[(\mathbf{z}-\bar{\mathbf{z}})(\mathbf{z}-\bar{\mathbf{z}})^T]$.

Moreover, we will also present in this appendix the minimum mean square error (MMSE), Bayesian MSE, and maximum *a posteriori* (MAP) estimators of the complex vector $\mathbf{h}$, which, as we will show, share the next common expression

$$\hat{\mathbf{h}}_* = (\beta \mathbf{X}^H \mathbf{X} + \mathbf{A})^{-1} \beta \mathbf{X}^H \mathbf{y}. \tag{31}$$

For the proof of the previous results, let us start by denoting the integral of a real function $f(\mathbf{y}) \in \mathbf{R}$, with complex-valued arguments $\mathbf{y} \in \mathbf{C}^M$, as

$$\int_{\mathbf{C}^M} f(\mathbf{y})\, d\mathbf{y} \equiv \int_{\mathbf{R}^M} \int_{\mathbf{R}^M} f(\mathbf{y})\, d\Re e\{\mathbf{y}\}\, d\Im m\{\mathbf{y}\}. \tag{32}$$

With this notation, the MSE for the estimation of $\mathbf{h}$ refers to the following quadratic risk function:

$$\mathcal{R}(\hat{\mathbf{h}}) = \int \int \|\mathbf{h} - \hat{\mathbf{h}}\|^2\, p(\mathbf{h}, \mathbf{y})\, d\mathbf{h}\, d\mathbf{y} \tag{33}$$

where $\hat{\mathbf{h}}$ is the considered estimate and $p(\mathbf{h}, \mathbf{y})$ denotes the joint density of $\mathbf{h}$ and $\mathbf{y}$. Similarly, the Bayesian mse is given by

$$\mathcal{R}(\hat{\mathbf{h}}|\mathbf{y}) = \int \|\mathbf{h} - \hat{\mathbf{h}}\|^2\, p(\mathbf{h}|\mathbf{y})\, d\mathbf{h} \tag{34}$$

where $p(\mathbf{h}|\mathbf{y})$ denotes the conditional density of $\mathbf{h}$ given $\mathbf{y}$.

The MMSE and Bayesian mse are related by

$$\mathcal{R}(\hat{\mathbf{h}}) = \int \mathcal{R}(\hat{\mathbf{h}}|\mathbf{y})\, p(\mathbf{y})\, d\mathbf{y} \tag{35}$$

where $p(\mathbf{y})$ is the marginal density of the measurements.

The minimization of the mse (i.e., the MMSE criterion) leads to the optimization of the Bayesian mse risk function, whose solution is attained at the center of the conditional distribution $p(\mathbf{h}|\mathbf{y})$, so it is given by the conditional mean of the regression coefficients given the measurements [17]

$$\begin{aligned} \hat{\mathbf{h}}_* &= \arg\min_{\hat{\mathbf{h}}}\ \mathcal{R}(\hat{\mathbf{h}}) \\ &= \arg\min_{\hat{\mathbf{h}}}\ \mathcal{R}(\hat{\mathbf{h}}|\mathbf{y}) \\ &= \int \mathbf{h}\, p(\mathbf{h}|\mathbf{y})\, d\mathbf{h}. \end{aligned}$$

Unfortunately, this later integral is usually intractable for arbitrary distributions. Although there are notable exceptions, such as the Gaussian density, for which the integral can be solved analytically, the question is whether one can theoretically justify the use of such favorable distributions. For this purpose, one can resort to the maximum entropy principle of Jaynes [22], which essentially states that in a choice between alternative modeling densities, the one with a maximum degree of randomness or entropy has the desirable property of "being maximally noncommittal with regard to missing information."

It is shown in Appendix II how the distribution of maximum differential entropy that is compatible with our relevant statistics ($\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)^T$ and σ^2) leads to a robust model $p(\mathbf{h}, \mathbf{y}|\boldsymbol{\alpha}, \sigma^2)$ for the joint density of the regression coefficients and measurements. This justifies our choice of a complex proper and Gaussian joint density model whose details have been summarized in Assumption A1.

The joint Gaussianity of the distribution implies Gaussianity of the posterior density $p(\mathbf{h}|\mathbf{y})$, simplifying the evaluation of the Bayesian mse estimator in (36). In this case, the optimal estimator reduces to the linear MMSE estimator

$$\hat{\mathbf{h}}_* = \mathbf{C}_{\mathbf{hy}} \mathbf{C}_{\mathbf{y}}^{-1}\, \mathbf{y} \tag{36}$$

where $\mathbf{C}_{\mathbf{hy}} = E[\mathbf{h}\, \mathbf{y}^H]$ and $\mathbf{C}_{\mathbf{y}} = E[\mathbf{y}\, \mathbf{y}^H]$. By evaluating these covariance matrices and applying the matrix inversion lemma, one obtains the estimator presented in (31), which we reproduce here for the readers' convenience

$$\hat{\mathbf{h}}_* = (\beta \mathbf{X}^H \mathbf{X} + \mathbf{A})^{-1} \beta \mathbf{X}^H \mathbf{y}. \tag{37}$$

This alternative expression to (36) emphasizes the dependence of the estimator on the prior precision matrix of the parameters $\mathbf{A} \equiv \mathrm{diag}(\boldsymbol{\alpha}) = \mathbf{C}_{\mathbf{h}}^{-1}$ and the precision of the measurement noise $\beta \equiv \sigma^{-2}$.

Having seen in (37) the expression of the MMSE estimator, we would like to review its connection with the MAP estimator. By definition, the MAP estimator of $\mathbf{h}$ is the mode of the posterior density of the regression coefficients given the measurements vector. Due to Assumption A1, it can be shown that the posterior distribution $p(\mathbf{h}|\mathbf{y})$ is also Gaussian and proper, with a precision matrix

$$\boldsymbol{\Sigma}_{\mathbf{h}|\mathbf{y}}^{-1} = \beta \mathbf{X}^H \mathbf{X} + \mathbf{A} \tag{38}$$

and mean

$$\boldsymbol{\mu}_{\mathbf{h}|\mathbf{y}} = \boldsymbol{\Sigma}_{\mathbf{h}|\mathbf{y}}\, \beta\, \mathbf{X}^H\, \mathbf{y}. \tag{39}$$

Since the mode of a Gaussian distribution is trivially attained at its mean, the MAP estimate is also given by $\hat{\mathbf{h}}_* \equiv \boldsymbol{\mu}_{\mathbf{h}|\mathbf{y}}$. Therefore, under Assumption A1, the MMSE, Bayesian mse, and MAP estimators of $\mathbf{h}$ coincide with (31).

Besides, in general, the regression coefficients are sparse, i.e., only a few of them are active. Thus, we still need a suitable way to select and estimate those coefficients that are more relevant for the considered problem while keeping the remaining ones inactive. As we show in Section II-A, the relevance vector learning approach considered by Tipping [12] provides a computationally convenient way to automatically infer the most relevant $\alpha_1, \ldots, \alpha_N$, through the maximization of the evidence (or marginal likelihood) of the measurements. This approach further regularizes the estimator in (31), which conveniently contributes to avoiding the potential overfitting to the measurements.

Appendix II
Distribution of Maximum Differential Entropy

In this section, we obtain the distribution of maximum differential entropy that is compatible with the relevant statistics

of our model, which are given by the available estimates of $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)^T$ and σ^2.

The joint density of $\mathbf{h}$ and $\mathbf{y}$ trivially decomposes as the product of a likelihood and a prior distribution

$$p(\mathbf{h}, \mathbf{y}|\boldsymbol{\alpha}, \sigma^2) = p(\mathbf{y}|\mathbf{h}, \boldsymbol{\alpha}, \sigma^2)\, p(\mathbf{h}|\boldsymbol{\alpha}, \sigma^2). \tag{40}$$

Note also that $p(\mathbf{h}, \mathbf{e}|\boldsymbol{\alpha}, \sigma^2) \equiv p(\mathbf{h}, \mathbf{y}|\boldsymbol{\alpha}, \sigma^2)$, and the joint differential entropy of $\mathbf{h}$ and $\mathbf{e}$ given $\boldsymbol{\alpha}$ and σ^2 is bounded above by the sum of differential entropies of the additive noise $\mathbf{e}|\sigma^2$ and the regression coefficients $\mathbf{h}|\boldsymbol{\alpha}$. This upper bound is attained if and only if the model coefficients and the additive noise are mutually independent. A property looks reasonable in our problem since any information on the coefficients does not alter the probability of the noise, i.e.,

$$\begin{aligned} p(\mathbf{y}|\mathbf{h}, \boldsymbol{\alpha}, \sigma^2) &\equiv p(\mathbf{e}|\mathbf{h}, \boldsymbol{\alpha}, \sigma^2) \\ &= p(\mathbf{e}|\sigma^2). \end{aligned} \tag{41}$$

Due to this independence, the pursuit of maximum entropy decouples in the problem of finding the individual densities $p(\mathbf{e}|\sigma^2)$ and $p(\mathbf{h}|\boldsymbol{\alpha})$ that maximize their respective differential entropies given their diagonal second-order statistics. In both cases, the distribution of maximum entropy enforces them to be complex-valued Gaussian random vectors with mutually independent and circularly symmetric elements. Hence, the covariance matrices of the noise $\sigma^2\mathbf{I}$ and the coefficients $\mathbf{A}^{-1} \equiv \text{diag}(\alpha_1^{-1}, \ldots, \alpha_N^{-1})$ are both diagonal, while their respective pseudocovariance matrices vanish. Therefore, the joint distribution of maximum differential entropy for $\mathbf{h}$ and $\mathbf{y}$ is Gaussian and factorizes as

$$p(\mathbf{h}, \mathbf{y}|\boldsymbol{\alpha}, \sigma^2) = p(\mathbf{y}|\mathbf{h}, \sigma^2)\, p(\mathbf{h}|\boldsymbol{\alpha}) \tag{42}$$

with a prior distribution of the coefficients equal to

$$\begin{aligned} p(\mathbf{h}|\boldsymbol{\alpha}) &= \prod_{n=1}^{N} p(h_n|\alpha_n) \\ &= \mathcal{CN}(\mathbf{0}, \mathbf{A}^{-1}) \end{aligned} \tag{43}$$

and a likelihood distribution of the observations

$$\begin{aligned} p(\mathbf{y}|\mathbf{h}, \sigma^2) &= \prod_{m=1}^{M} p(e_m|\sigma^2) \\ &= \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I}). \end{aligned} \tag{44}$$

This justifies the robust model adopted in Assumption A1.

REFERENCES

[1] C. Fager, T. Eriksson, F. Barradas, K. Hausmair, T. Cunha, and J. C. Pedro, "Linearity and efficiency in 5G transmitters: New techniques for analyzing efficiency, linearity, and linearization in a 5G active antenna transmitter context," *IEEE Microw. Mag.*, vol. 20, no. 5, pp. 35–49, May 2019.

[2] A. Katz, J. Wood, and D. Chokola, "The evolution of PA linearization: From classic feedforward and feedback through analog and digital predistortion," *IEEE Microw. Mag.*, vol. 17, no. 2, pp. 32–40, Feb. 2016.

[3] H. Yin *et al.*, "Data-clustering-assisted digital predistortion for 5G millimeter-wave beamforming transmitters with multiple dynamic configurations," *IEEE Trans. Microw. Theory Techn.*, vol. 69, no. 3, pp. 1805–1816, Mar. 2021.

[4] E. Guillena, W. Li, G. Montoro, R. Quaglia, and P. L. Gilabert, "Reconfigurable DPD based on ANNs for wideband load modulated balanced amplifiers under dynamic operation from 1.8 to 2.4 GHz," *IEEE Trans. Microw. Theory Techn.*, vol. 70, no. 1, pp. 453–465, Jan. 2022, doi: 10.1109/TMTT.2021.3091672.

[5] A. Brihuega, M. Abdelaziz, L. Anttila, Y. Li, A. Zhu, and M. Valkama, "Mixture of experts approach for piecewise modeling and linearization of RF power amplifiers," *IEEE Trans. Microw. Theory Techn.*, vol. 70, no. 1, pp. 380–391, Jan. 2022, doi: 10.1109/TMTT.2021.3098867.

[6] J. Kim and K. Konstantinou, "Digital predistortion of wideband signals based on power amplifier model with memory," *Electron. Lett.*, vol. 37, no. 23, pp. 1417–1418, Nov. 2001.

[7] D. R. Morgan, Z. Ma, J. Kim, M. G. 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.

[8] A. Zhu, J. C. Pedro, and T. J. Brazil, "Dynamic deviation reduction-based Volterra behavioral modeling of RF power amplifiers," *IEEE Trans. Microw. Theory Techn.*, vol. 54, no. 12, pp. 4323–4332, Dec. 2006.

[9] A. Abdelhafiz, A. Kwan, O. Hammi, and F. M. Ghannouchi, "Digital predistortion of LTE-A power amplifiers using compressed-sampling-based unstructured pruning of Volterra series," *IEEE Trans. Microw. Theory Techn.*, vol. 62, no. 11, pp. 2583–2593, Nov. 2014.

[10] 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.

[11] J. A. Becerra, M. J. M. Ayora, J. Reina-Tosina, and C. Crespo-Cadenas, "Sparse identification of Volterra models for power amplifiers without pseudoinverse computation," *IEEE Trans. Microw. Theory Techn.*, vol. 68, no. 11, pp. 4570–4578, Nov. 2020.

[12] M. E. Tipping, "The relevance vector machine," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 12. Cambridge, MA, USA: MIT Press, 2000, pp. 652–658.

[13] A. C. Faul and M. E. Tipping, "Analysis of sparse Bayesian learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 14. Cambridge, MA, USA: MIT Press, 2002, pp. 383–389.

[14] M. E. Tipping and A. C. Faul, "Fast marginal likelihood maximisation for sparse Bayesian models," in *Proc. 9th Int. Workshop Artif. Intell. Statist.*, Key West, FL, USA, Jan. 2003, pp. 276–283.

[15] J. Peng, S. He, B. Wang, Z. Dai, and J. Pang, "Digital predistortion for power amplifier based on sparse Bayesian learning," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 63, no. 9, pp. 828–832, Sep. 2016.

[16] C. Crespo-Cadenas, M. J. Madero-Ayora, J. A. Becerra, and S. Cruces, "A fast sparse Bayesian pursuit approach for power amplifier linearization," in *IEEE MTT-S Int. Microw. Symp. Dig.*, May 2021, pp. 1–3.

[17] S. M. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*. Upper Saddle River, NJ, USA: Prentice-Hall, 1998.

[18] P. J. Schreier and L. L. Scharf, *Statistical Signal Processing of Complex-Valued Data*. New York, NY, USA: Cambridge Univ. Press, 2010.

[19] C. Crespo-Cadenas, M. J. Madero-Ayora, and J. A. Becerra, "A bivariate Volterra series approach to modeling and linearization of power amplifiers," in *Proc. IEEE Topical Conf. RF/Microw. Power Modeling Radio Wireless Appl. (PAWR)*, Jan. 2021, pp. 4–7.

[20] A. Barry, W. Li, J. A. Becerra, and P. L. Gilabert, "Comparison of feature selection techniques for power amplifier behavioral modeling and digital predistortion linearization," *Sensors*, vol. 21, no. 17, p. 5772, Aug. 2021.

[21] D. Zhou and V. E. DeBrunner, "Novel adaptive nonlinear predistorters based on the direct learning algorithm," *IEEE Trans. Signal Process.*, vol. 55, no. 1, pp. 120–133, Jan. 2007.

[22] E. T. Jaynes, "Information theory and statistical mechanics," *Phys. Rev.*, vol. 106, no. 4, pp. 620–630, 1957.