Compressive Modeling of Stationary Autoregressive Processes

Georg Kail and Geert Leus
Department of Microelectronics, Delft University of Technology (TU Delft), The Netherlands
email: {g.r.kail, g.j.t.leus}@tudelft.nl

***Abstract*—Compressive covariance sampling (CCS) methods that estimate the correlation function from compressive measurements have achieved great compression rates lately. In stationary autoregressive (AR) processes, the power spectrum is fully determined by the AR parameters, and vice versa. Therefore, compressive estimation of AR parameters amounts to CCS for such signals. However, previous CCS methods typically do not fully exploit the structure of AR power spectra. On the other hand, traditional AR parameter estimation methods cannot be used when only a compressed version of the AR signal is observed. We propose a Bayesian algorithm for estimating AR parameters from compressed observations, using a Metropolis-Hastings sampler. Simulation results confirm the promising performance of the proposed method.**

***Index Terms*—Autoregressive process, power spectrum estimation, compressive sampling, Bayesian sampling, Metropolis-Hastings sampler**

I. Introduction

We consider the problem of estimating the parameters of a stationary autoregressive (AR) process when the observed signal is not the AR signal itself but a compressed version of it. Due to compression, the covariance matrix of the observed signal loses its Toeplitz structure. Classical methods of AR coefficient estimation, e.g. [1], are not designed for compressive observations. The same is true for most methods in the closely related field of AR model fitting [2]–[4], where a non-AR signal is approximated by an AR process.

Since the power spectrum of stationary AR processes is fully determined by the AR parameters, the present problem is equivalent to compressive power spectrum estimation or compressive covariance sensing (CCS) [5]–[7] for such signals. It amounts to *structured CCS* accounting for the particular parametric structure of AR processes. For other classes of signals, CCS methods maintain good performance even at low compression rates. However, they typically do not exploit the particular structure of AR power spectra.

In nonlinear parametric estimation problems, Markov chain Monte Carlo (MCMC) methods [8] are often used (e.g., [9], [10]). The method proposed here belongs to this family, employing *Metropolis-Hastings (MH) within Gibbs sampling* [11]. The performance of this iterative algorithm (for a limited number of iterations) critically depends on the appropriate design of proposal distributions that govern the innovation in each iteration.

Previous Work. Generalizations of the AR covariance and inverse covariance matrices are considered in [12], [13], where different block structures corresponding to 2D time-varying AR models are discussed. In contrast to 2D time-varying AR models, however, the compressive AR model leads to blocks that are not banded. In AR model fitting, compressive observations are considered in [3], where the vector of AR coefficients is forced to be sparse. The present paper, on the other hand, makes no sparsity assumptions. In [4], AR model fitting is performed using irregularly sampled data, which could in principle be reinterpreted as compressive observations. However, the method discards large parts of the irregularly sampled data in a way that makes it unsuited for compressive AR parameter estimation. In signal segmentation, MCMC methods have been used for AR model fitting [9], [14]. These methods rely on uncompressed observations. In CCS, linear methods have been used successfully both for nonparametric covariance sensing [5], [6] and for particular linear parametric models [7]. The specific nonlinear parametric model of AR processes has not been considered in this context.

Contributions. The main contributions of this paper are the following. First, the gap between AR parameter estimation and CCS is closed by formulating compressive AR parameter estimation as a problem of structured CCS. Second, for solving this problem, we propose a method that is substantially different from previous CCS methods due to the nonlinearity of the problem. As simulation results confirm, the proposal distributions we design for this MH within Gibbs sampling method achieve high performance within moderate processing time even for low compression rates.

This paper is organized as follows. Section II describes the compressed AR signal model. The proposed estimation method is presented in Section III. Numerical results assessing the performance of the proposed method are discussed in Section IV.

II. Signal Model

Observation Model. We consider an unobserved complex signal of interest x_n that is modeled as a stationary AR process of order p:

$$x_n = \sum_{i=1}^{p} a_i \, x_{n-i} + e_n \, .$$

Here, $a_i \in \mathbb{R}$ for $i = 1, \ldots, p$ denotes the AR coefficients, and e_n is zero-mean white circularly symmetric complex Gaussian noise with variance σ^2. The AR coefficients $\mathbf{a} = (a_1 \cdots a_p)^{\mathrm{T}}$,

where $^{\mathrm{T}}$ denotes transposition, and the noise variance σ^2 are the parameters which we will aim to estimate. We assume that the model order p is fixed and known. The observed signal y_n is obtained from x_n through periodic compressive sampling, i.e.,

$$\mathbf{y}[k] = \mathbf{\Phi}\,\mathbf{x}[k], \tag{1}$$

with the length-M vector $\mathbf{y}[k] = (y_{(k-1)M+1} \cdots y_{kM})^{\mathrm{T}}$ and the length-N vector $\mathbf{x}[k] = (x_{(k-1)N+1} \cdots x_{kN})^{\mathrm{T}}$. The dimensions of the $M \times N$ compression matrix $\mathbf{\Phi}$ determine the compression rate, which is M/N. Suitable choices for the elements of $\mathbf{\Phi}$ have been studied, e.g., in [15]. In our experiments, we use complex Gaussian elements. Let K denote the number of blocks $\mathbf{y}[k]$ that we observe. Then (1) for $k = 1, \ldots, K$ can be expressed as

$$\mathbf{y} = \widetilde{\mathbf{\Phi}}\,\mathbf{x}, \tag{2}$$

with $\mathbf{y} = (\mathbf{y}^{\mathrm{T}}[1] \cdots \mathbf{y}^{\mathrm{T}}[K])^{\mathrm{T}}$ and $\mathbf{x} = (\mathbf{x}^{\mathrm{T}}[1] \cdots \mathbf{x}^{\mathrm{T}}[K])^{\mathrm{T}}$. The $KM \times KN$ matrix $\widetilde{\mathbf{\Phi}}$ is defined as $\widetilde{\mathbf{\Phi}} = \mathbf{I}_K \otimes \mathbf{\Phi}$, where $\mathbf{I}_K$ is the $K \times K$ identity matrix and $\otimes$ denotes the Kronecker product.

Our goal is to estimate $\mathbf{a}$ and σ^2 based on $\mathbf{y}$. However, rather than estimating $\mathbf{a}$ directly, it will be convenient to first estimate the reflection coefficients $\boldsymbol{\rho} = (\rho_1 \cdots \rho_p)^{\mathrm{T}}$ of the corresponding lattice filter [16, p. 223] (cf. pp. 226, 233, and 236 for the following statements). For a given $\boldsymbol{\rho}$, the corresponding $\mathbf{a}$ is obtained by repeating the following calculations (the Levinson-Durbin recursion) for $i = 1, \ldots, p$:

$$\begin{aligned} a_i^{(i)} &= \rho_i\,, \\ a_j^{(i)} &= a_j^{(i-1)} - \rho_i\, a_{i-j}^{(i-1)} \qquad \text{for } j = 1, \ldots, i-1\,. \end{aligned} \tag{3}$$

The vector $\mathbf{a} = \left(a_1^{(p)} \cdots a_p^{(p)}\right)^{\mathrm{T}}$ is then the $\mathbf{a}$ corresponding to $\boldsymbol{\rho}$. The advantage of estimating $\boldsymbol{\rho}$ is its relation to stability [1], [4], [17]: an AR process with $\mathbf{a} \in \mathbb{R}^p$ is stable if and only if $\boldsymbol{\rho} \in \mathbb{R}^p$ and $|\rho_i| < 1$ for $i = 1, \ldots, p$. Therefore, we can guarantee stability of the estimated AR process by estimating $\boldsymbol{\rho}$ within the domain $(-1, 1)^p$.

Likelihood Function. Our estimates of $\boldsymbol{\rho}$ and σ^2 from $\mathbf{y}$ will be based on the likelihood function $p(\mathbf{y}|\boldsymbol{\rho}, \sigma^2)$. We start our derivation of $p(\mathbf{y}|\boldsymbol{\rho}, \sigma^2)$ by assessing the statistical properties of x_n. For zero-mean circularly symmetric complex Gaussian noise e_n, stationary AR processes x_n are also zero-mean circularly symmetric complex Gaussian distributed. Therefore, $p(\mathbf{x}|\boldsymbol{\rho}, \sigma^2)$ is fully specified by the autocorrelation matrix $\mathbf{R}_x = \mathrm{E}\{\mathbf{x}\mathbf{x}^{\mathrm{H}}|\boldsymbol{\rho}, \sigma^2\}$, where $^{\mathrm{H}}$ denotes conjugate transposition. Due to stationarity, $\mathbf{R}_x$ is a Toeplitz matrix with first column $(r_0 \cdots r_{KN})^{\mathrm{T}}$ and first row $(r_0 \cdots r_{-KN})$, where $r_n = \mathrm{E}\{x_{n'} x^*_{n'+n}\}$. The autocorrelation function r_n is related to $\mathbf{a}$ and σ^2 through the system of $p+1$ equations known as the Yule-Walker equations [16, p. 194]:

$$r_n = \sum_{i=1}^{p} a_i\, r_{n-i} + \sigma^2 \delta_n\,, \tag{4}$$

for $n = 0, \ldots, p$, where δ_n denotes the unit sample. Using the symmetry $r_{-n} = r_n^*$ and solving the above equations with respect to $\mathbf{r} = (r_0 \cdots r_p)^{\mathrm{T}}$ yields

$$\mathbf{r} = \sigma^2 (\mathbf{A}\mathbf{\Lambda})^{-1} \boldsymbol{\delta}_1\,. \tag{5}$$

Here, $\mathbf{A}$ is a Hankel matrix of size $(p{+}1) \times (2p{+}1)$ with the first column $(0 \cdots 0\ 1)^{\mathrm{T}}$ and the last row $(1\ -a_1 \cdots -a_p\ 0 \cdots 0)$. The matrix $\mathbf{\Lambda}$ is composed of the last $p{+}1$ columns of the $(2p{+}1) \times (2p{+}1)$ matrix whose diagonal and antidiagonal elements are 1 and whose other elements are 0. Finally, $\boldsymbol{\delta}_1$ is the unit vector of $p{+}1$ elements whose first element is 1. Note that due to (5) $\mathbf{r}$ is real. With $\mathbf{r}$ given, r_n can be calculated for arbitrary n using (4) and $r_{-n} = r_n$. This allows us to construct $\mathbf{R}_x$. Since according to (4) and (5) the autocorrelation r_n is fully determined by $\mathbf{a}$ and σ^2 (or, equivalently, $\boldsymbol{\rho}$ and σ^2), estimation of these parameters is equivalent to parametric power spectrum estimation for AR signals.

It follows from (5) that r_n, and therefore also $\mathbf{R}_x$, is proportional to σ^2. For later computations, it will be useful to define $\tilde{r}_n$, $\tilde{\mathbf{r}}$, and $\widetilde{\mathbf{R}}_x$ by dividing r_n, $\mathbf{r}$, and $\mathbf{R}_x$ by σ^2. From (5) and (4), we then obtain

$$\tilde{\mathbf{r}} = (\mathbf{A}\mathbf{\Lambda})^{-1}\boldsymbol{\delta}_1 \quad \text{and} \quad \tilde{r}_n = \sum_{i=1}^{p} a_i\, \tilde{r}_{n-i} + \delta_n\,. \tag{6}$$

To emphasize that $\widetilde{\mathbf{R}}_x$ (for given K) is fully determined by $\boldsymbol{\rho}$ through (3) and (6), we will denote it as $\widetilde{\mathbf{R}}_x(\boldsymbol{\rho})$ in the following.

Since $\mathbf{x}$ for given $\boldsymbol{\rho}$ and σ^2 is zero-mean circularly symmetric complex Gaussian distributed, it follows from (2) that $\mathbf{y}$ for given $\boldsymbol{\rho}$ and σ^2 is also zero-mean circularly symmetric complex Gaussian distributed with the autocorrelation matrix $\mathbf{R}_y = \mathrm{E}\{\mathbf{y}\mathbf{y}^{\mathrm{H}}\} = \widetilde{\mathbf{\Phi}}\mathbf{R}_x\widetilde{\mathbf{\Phi}}^{\mathrm{H}}$. In analogy to $\widetilde{\mathbf{R}}_x(\boldsymbol{\rho})$, we define $\widetilde{\mathbf{R}}_y(\boldsymbol{\rho}) = \widetilde{\mathbf{\Phi}}\widetilde{\mathbf{R}}_x(\boldsymbol{\rho})\widetilde{\mathbf{\Phi}}^{\mathrm{H}}$, which (for given K and $\mathbf{\Phi}$, which we assume) is fully determined by $\boldsymbol{\rho}$ through (3) and (6). Finally, we obtain the likelihood function

$$\begin{aligned} p(\mathbf{y}|\boldsymbol{\rho}, \sigma^2) &= \frac{1}{\pi^{KM}\,|\mathbf{R}_y|}\,\exp\left(-\,\mathbf{y}^{\mathrm{H}}\,\mathbf{R}_y^{-1}\mathbf{y}\right) \\ &= \frac{1}{(\pi\sigma^2)^{KM}\,|\widetilde{\mathbf{R}}_y(\boldsymbol{\rho})|}\,\exp\left(-\,\frac{\mathbf{y}^{\mathrm{H}}\left(\widetilde{\mathbf{R}}_y(\boldsymbol{\rho})\right)^{-1}\mathbf{y}}{\sigma^2}\right). \end{aligned} \tag{7}$$

Parameter Priors and Joint Posterior. The Bayesian estimation methodology, which was chosen here, involves assigning prior probability distributions to the parameters of interest, i.e., $\boldsymbol{\rho}$ and σ^2. The absence of prior assumptions about the parameters is reflected by noninformative priors. We adopt this approach and choose uniform priors for the reflection coefficients as well as the noise variance:

$$p(\boldsymbol{\rho}) = 1/2^p\,, \qquad p(\sigma^2) = 1\,, \tag{8}$$

for $\boldsymbol{\rho} \in (-1, 1)^p$ and $\sigma^2 > 0$. In the case of $\boldsymbol{\rho}$, the uniform distribution is in fact the maximum entropy distribution, since the support of $\boldsymbol{\rho}$ is finite and no further prior assumptions were made. Recall that the domain of $\boldsymbol{\rho}$ was chosen to be $(-1, 1)^p$ because this corresponds to the set of all stable AR processes with $\mathbf{a} \in \mathbb{R}^p$. In the case of σ^2, the uniform prior is an improper prior, since the support of σ^2 is infinite. However,

we will see in (11) that the resulting posterior distribution is a proper distribution. Using (7) and (8), we obtain the joint posterior distribution of $\boldsymbol{\rho}$ and σ^2:

$$p(\boldsymbol{\rho}, \sigma^2|\mathbf{y}) \;\propto\; p(\mathbf{y}|\boldsymbol{\rho}, \sigma^2)\; p(\boldsymbol{\rho})\; p(\sigma^2) \;\propto\; p(\mathbf{y}|\boldsymbol{\rho}, \sigma^2)\,, \quad (9)$$

for $(\boldsymbol{\rho},\ \sigma^2) \in (-1, 1)^p \times (0, \infty)$. Here, we dropped factors that are constant with respect to $\boldsymbol{\rho}$ and σ^2. The normalization constant of $p(\boldsymbol{\rho}, \sigma^2|\mathbf{y})$ will not be relevant for our estimation method.

III. Estimation method

Sample-Based Bayesian Estimation. Our goal is to estimate $\boldsymbol{\rho}$ and σ^2 from $\mathbf{y}$ by maximizing their joint posterior distribution $p(\boldsymbol{\rho}, \sigma^2|\mathbf{y})$, i.e., maximum a-posteriori (MAP) estimation. Due to the uniform priors chosen in (8), the MAP estimator coincides with the maximum likelihood (ML) estimator. Since the maximization of $p(\boldsymbol{\rho}, \sigma^2|\mathbf{y})$ is too complex to calculate directly, we resort to an approximate solution. Following the concept of Bayesian sampling [8], we generate a large population of realizations $\big(\boldsymbol{\rho}^{(j)}, (\sigma^2)^{(j)}\big)$ from $p(\boldsymbol{\rho}, \sigma^2|\mathbf{y})$. As the number of realizations increases, the population becomes denser, thus reducing the minimum distance between the true MAP estimate and its nearest neighboring realizations. Since $p(\boldsymbol{\rho}, \sigma^2|\mathbf{y})$ is continuous within the domain of $\boldsymbol{\rho}$ and σ^2, it follows that $\max_j p\big(\boldsymbol{\rho}^{(j)}, (\sigma^2)^{(j)}\big|\mathbf{y}\big)$ approaches $\max_{\boldsymbol{\rho},\sigma^2} p(\boldsymbol{\rho}, \sigma^2|\mathbf{y})$ as the population grows. Therefore, by calculating $p\big(\boldsymbol{\rho}^{(j)}, (\sigma^2)^{(j)}\big|\mathbf{y}\big)$ for all j and picking the maximum, i.e.,

$$\left.\begin{aligned} \widehat{\boldsymbol{\rho}} &= \boldsymbol{\rho}^{(j_{\max})} \\ \widehat{\sigma^2} &= (\sigma^2)^{(j_{\max})} \end{aligned}\right\} \quad \text{with} \quad j_{\max} = \operatorname*{argmax}_j\, p\big(\boldsymbol{\rho}^{(j)}, (\sigma^2)^{(j)}\big|\mathbf{y}\big)\,, \quad (10)$$

we obtain approximate MAP estimates. The efficiency of this procedure stems from the fact that, since the realizations $\big(\boldsymbol{\rho}^{(j)}, (\sigma^2)^{(j)}\big)$ are generated from $p(\boldsymbol{\rho}, \sigma^2|\mathbf{y})$, the density of realizations in the population is especially high near the maximum of $p(\boldsymbol{\rho}, \sigma^2|\mathbf{y})$. Note also that it is not necessary to store the entire population to obtain the global maximum; at each iteration, we can simply compare the new realization $\big(\boldsymbol{\rho}^{(j)}, (\sigma^2)^{(j)}\big)$ to the realization that previously maximized $p\big(\boldsymbol{\rho}, \sigma^2\big|\mathbf{y}\big)$. If the new realization achieves a larger $p\big(\boldsymbol{\rho}, \sigma^2\big|\mathbf{y}\big)$, it replaces the previous maximum, otherwise is can be discarded. Thus, throughout the entire process, only one realization needs to be stored.

Metropolis-Hastings Sampling. For generating the population of realizations $\big(\boldsymbol{\rho}^{(j)}, (\sigma^2)^{(j)}\big)$, we adopt the MCMC approach [8], more specifically the MH algorithm. Formulated in brief for a generic parameter vector $\boldsymbol{\theta}$ and a distribution of interest $p(\boldsymbol{\theta})$, each iteration of this algorithm, indexed by j, produces a new realization $\boldsymbol{\theta}^{(j)}$ drawn from $p(\boldsymbol{\theta})$. This is done in two steps. In the first step, a proposal $\widetilde{\boldsymbol{\theta}}$ is generated from a proposal distribution $q_j\big(\widetilde{\boldsymbol{\theta}}\big|\boldsymbol{\theta}^{(j-1)}\big)$, which is some distribution that is convenient for sampling and introduces some variation based on the previous realization $\boldsymbol{\theta}^{(j-1)}$. In the second step, the new realization $\boldsymbol{\theta}^{(j)}$ is chosen as

$$\boldsymbol{\theta}^{(j)} = \begin{cases} \widetilde{\boldsymbol{\theta}} & \text{with probability } \alpha_j \\ \boldsymbol{\theta}^{(j-1)} & \text{with probability } 1-\alpha_j\,, \end{cases}$$

where the acceptance probability α_j is given as

$$\alpha_j = \min\left\{ \frac{p\big(\widetilde{\boldsymbol{\theta}}\big)\; q_j\big(\boldsymbol{\theta}^{(j-1)}\big|\widetilde{\boldsymbol{\theta}}\big)}{p\big(\boldsymbol{\theta}^{(j-1)}\big)\; q_j\big(\widetilde{\boldsymbol{\theta}}\big|\boldsymbol{\theta}^{(j-1)}\big)}, 1 \right\}.$$

After a certain number of iterations, the subsequent realizations $\boldsymbol{\theta}^{(j)}$ are distributed according to $p(\boldsymbol{\theta})$, independently of the initial realization $\boldsymbol{\theta}^{(0)}$. In contrast to estimation methods that involve averaging or counting some of the realizations $\boldsymbol{\theta}^{(j)}$, the estimator in (10) can use all realizations including those from the first iterations. For terminating the iterative process, different strategies have been proposed, such as assessing whether the distribution of the realizations has converged to a stationary distribution [18]. In our experiments, we chose a simple solution with a fixed number of iterations J, which can be determined based on previous (training) results.

The MH algorithm was chosen here because its general formulation makes it possible to generate a population from almost any distribution. However, the number of iterations that are needed critically depends on the design of the proposal distributions $q_j\big(\widetilde{\boldsymbol{\theta}}\big|\boldsymbol{\theta}^{(j-1)}\big)$. If the dependence of $\widetilde{\boldsymbol{\theta}}$ on $\boldsymbol{\theta}^{(j-1)}$ is either too strong or too weak, or if $q_j\big(\widetilde{\boldsymbol{\theta}}\big|\boldsymbol{\theta}^{(j-1)}\big)$ produces many proposals for which α_j is low, the algorithm may not yield useful results within a tolerable number of iterations. In our estimation problem, the distribution of interest $p(\boldsymbol{\theta})$ corresponds to $p(\boldsymbol{\rho}, \sigma^2|\mathbf{y})$. The proposal distributions $q_j\big(\widetilde{\boldsymbol{\theta}}\big|\boldsymbol{\theta}^{(j-1)}\big)$ are chosen such that they alternately produce a new $\widetilde{\boldsymbol{\rho}}$ or $\widetilde{\sigma}^2$, while the respective other stays the same as its previous realization:

$$q_j\big(\widetilde{\boldsymbol{\rho}}, \widetilde{\sigma}^2\big|\boldsymbol{\rho}^{(j-1)}, (\sigma^2)^{(j-1)}\big) = \begin{cases} q_{\boldsymbol{\rho}}\big(\widetilde{\boldsymbol{\rho}}\big|\boldsymbol{\rho}^{(j-1)}, (\sigma^2)^{(j-1)}\big)\; \delta\big(\widetilde{\sigma}^2 - (\sigma^2)^{(j-1)}\big) & \text{for even } j \\ q_{\sigma^2}\big(\widetilde{\sigma}^2\big|\boldsymbol{\rho}^{(j-1)}, (\sigma^2)^{(j-1)}\big)\; \delta\big(\widetilde{\boldsymbol{\rho}} - \boldsymbol{\rho}^{(j-1)}\big) & \text{for odd } j\,. \end{cases}$$

Here, $\delta(\cdot)$ denotes the Dirac delta function. This alternating pattern in the proposal distributions is well-known from an important special case of MH sampling, the Gibbs sampler [8]. Gibbs sampling corresponds to a particular choice of proposal distributions, which in our case would be $q_{\boldsymbol{\rho}}\big(\widetilde{\boldsymbol{\rho}}\big|\boldsymbol{\rho}^{(j-1)}, (\sigma^2)^{(j-1)}\big) = p\big(\widetilde{\boldsymbol{\rho}}\big|(\sigma^2)^{(j-1)}, \mathbf{y}\big)$ and $q_{\sigma^2}\big(\widetilde{\sigma}^2\big|\boldsymbol{\rho}^{(j-1)}, (\sigma^2)^{(j-1)}\big) = p\big(\widetilde{\sigma}^2\big|\boldsymbol{\rho}^{(j-1)}, \mathbf{y}\big)$. The convenient consequence is that $\alpha_j = 1$ for all j, as is easily shown. This is particularly useful for fast convergence of the estimator in (10), because the estimate does not improve when $\boldsymbol{\theta}^{(j)} = \boldsymbol{\theta}^{(j-1)}$. However, this solution requires that realizations of $\boldsymbol{\rho}$ and σ^2 can be generated directly from the posterior distributions $p(\boldsymbol{\rho}|\sigma^2, \mathbf{y})$ and $p(\sigma^2|\boldsymbol{\rho}, \mathbf{y})$, which are proportional to $p(\boldsymbol{\rho}, \sigma^2|\mathbf{y})$ and normalized with regard to $\boldsymbol{\rho}$ and σ^2, respectively. Inspection of (7) shows that $p(\sigma^2|\boldsymbol{\rho}, \mathbf{y})$ is in fact a distribution from which we can easily generate realizations, as will be discussed shortly. However, the same is not true for $p(\boldsymbol{\rho}|\sigma^2, \mathbf{y})$, which precludes Gibbs sampling for

the given problem. By using $p(\sigma^2|\boldsymbol{\rho},\mathbf{y})$ to propose $\widetilde{\sigma}^2$ in odd-numbered iterations and defining a different type of proposal distribution for $\widetilde{\boldsymbol{\rho}}$ in even-numbered distributions, we employ *MH within Gibbs sampling* [11].

Proposal and Decision Steps for σ^2. As mentioned above, $\widetilde{\sigma}^2$ is generated from $p(\widetilde{\sigma}^2|\boldsymbol{\rho}^{(j-1)},\mathbf{y})$ (in odd-numbered iterations). This can be done because—as inspection of (7) shows—for fixed $\boldsymbol{\rho}$ and $\mathbf{y}$, $p(\boldsymbol{\rho},\sigma^2|\mathbf{y})$ is an inverse gamma distribution of σ^2, up to a normalization constant. We thus obtain

$$q_{\sigma^2}\big(\widetilde{\sigma}^2\big|\boldsymbol{\rho}^{(j-1)},(\sigma^2)^{(j-1)}\big) = p(\widetilde{\sigma}^2|\boldsymbol{\rho}^{(j-1)},\mathbf{y}) = \frac{\Big(\gamma\big(\boldsymbol{\rho}^{(j-1)}\big)\Big)^{\psi}}{\Gamma(\psi)}\,\frac{1}{(\widetilde{\sigma}^2)^{\psi+1}}\,\exp\left(-\frac{\gamma\big(\boldsymbol{\rho}^{(j-1)}\big)}{\widetilde{\sigma}^2}\right), \quad (11)$$

for $\widetilde{\sigma}^2>0$, where $\Gamma(\cdot)$ denotes the gamma function and

$$\psi = KM-1\,, \quad \gamma(\boldsymbol{\rho}) = \mathbf{y}^{\mathrm{H}}\left(\widetilde{\mathbf{R}}_y(\boldsymbol{\rho})\right)^{-1}\mathbf{y}\,. \quad (12)$$

We recall that this choice of $q_{\sigma^2}\big(\widetilde{\sigma}^2\big|\boldsymbol{\rho}^{(j-1)},(\sigma^2)^{(j-1)}\big)$ is advantageous because it ensures that $\alpha_j=1$ for all odd j.

Proposal and Decision Steps for $\boldsymbol{\rho}$. For the proposal distribution $q_{\boldsymbol{\rho}}\big(\widetilde{\boldsymbol{\rho}}\big|\boldsymbol{\rho}^{(j-1)},(\sigma^2)^{(j-1)}\big)$, we choose a beta distribution on the interval $(-1,1)$, independently for each element $\widetilde{\rho}_i$:

$$q_{\boldsymbol{\rho}}\big(\widetilde{\boldsymbol{\rho}}\big|\boldsymbol{\rho}^{(j-1)},(\sigma^2)^{(j-1)}\big) = \prod_{i=1}^{p} q_{\rho,i}\big(\widetilde{\rho}_i\big|\rho_i^{(j-1)}\big)\,,$$

with

$$q_{\rho,i}\big(\widetilde{\rho}_i\big|\rho_i^{(j-1)}\big) = \frac{1}{2\,\mathrm{B}\Big(\eta\big(\rho_i^{(j-1)}\big),\xi\big(\rho_i^{(j-1)}\big)\Big)} \times\left(\frac{1+\widetilde{\rho}_i}{2}\right)^{\eta(\rho_i^{(j-1)})-1}\left(\frac{1-\widetilde{\rho}_i}{2}\right)^{\xi(\rho_i^{(j-1)})-1}. \quad (13)$$

Here, $\mathrm{B}(\cdot,\cdot)$ denotes the beta function and

$$\eta(\rho_i)=\frac{2\,(1+\rho_i)}{1-|\rho_i|}, \quad \xi(\rho_i)=\frac{2\,(1-\rho_i)}{1-|\rho_i|}.$$

This particular distribution is chosen for the following reasons. First, it is nonzero on the entire domain $(-1,1)$ but concentrated around its mean, which is $\rho_i^{(j-1)}$. As experiments confirm, this ensures a good balance between large enough variation to quickly become independent from the initialization and small enough variation for a fairly steady improvement of the estimate. Second, this distribution converges towards zero as $|\widetilde{\rho}_i|$ approaches 1, which is important because proposals that are extremely close to 1 may lead to numerical problems in the calculation of $p\big(\widetilde{\boldsymbol{\rho}},(\sigma^2)^{(j-1)}\big|\mathbf{y}\big)$. For the acceptance probability α_j in iterations with even j, we thus obtain

$$\alpha_j = \min\left\{\frac{p\big(\widetilde{\boldsymbol{\rho}},(\sigma^2)^{(j-1)}\big|\mathbf{y}\big)\,q_{\boldsymbol{\rho}}\big(\boldsymbol{\rho}^{(j-1)}\big|\widetilde{\boldsymbol{\rho}},(\sigma^2)^{(j-1)}\big)}{p\big(\boldsymbol{\rho}^{(j-1)},(\sigma^2)^{(j-1)}\big|\mathbf{y}\big)\,q_{\boldsymbol{\rho}}\big(\widetilde{\boldsymbol{\rho}}\big|\boldsymbol{\rho}^{(j-1)},(\sigma^2)^{(j-1)}\big)},1\right\} = \min\left\{\frac{p\big(\mathbf{y}\big|\widetilde{\boldsymbol{\rho}},(\sigma^2)^{(j-1)}\big)}{p\big(\mathbf{y}\big|\boldsymbol{\rho}^{(j-1)},(\sigma^2)^{(j-1)}\big)}\,\prod_{i=1}^{p}\frac{q_{\rho,i}\big(\rho_i^{(j-1)}\big|\widetilde{\rho}_i\big)}{q_{\rho,i}\big(\widetilde{\rho}_i\big|\rho_i^{(j-1)}\big)},1\right\}. \quad (14)$$

In the last step, we used (9). Note that even though the proposal distribution $q_{\boldsymbol{\rho}}\big(\widetilde{\boldsymbol{\rho}}\big|\boldsymbol{\rho}^{(j-1)},(\sigma^2)^{(j-1)}\big)$ does not depend on $\mathbf{y}$ directly, $\boldsymbol{\rho}^{(j)}$ still depends on $\mathbf{y}$ because of (14).

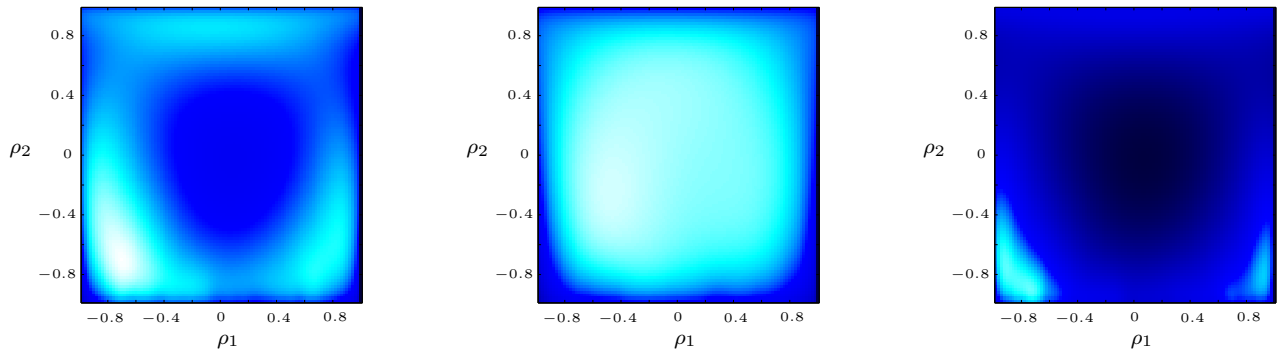
Approximation of the Likelihood. Every other iteration of the algorithm described above requires evaluating (7) for the respective $\widetilde{\boldsymbol{\rho}}$, which involves calculating the corresponding autocorrelation function $\tilde{r}_n$ for $n=0,\ldots,KN$ and inverting the $KM\times KM$ matrix $\widetilde{\mathbf{R}}_y(\widetilde{\boldsymbol{\rho}})$. For large enough values of KN and M/N to ensure reliable estimation of the AR parameters, this leads to an unacceptable computational complexity. We resolve this by using an approximation of (7) that was proposed in [7] for arbitrary autocorrelation functions r_n. We start by observing that $\mathbf{y}^{\mathrm{H}}\left(\widetilde{\mathbf{R}}_y(\boldsymbol{\rho})\right)^{-1}\mathbf{y}$ in the exponent of (7) is equal to $\mathrm{Tr}\left(\left(\widetilde{\mathbf{R}}_y(\boldsymbol{\rho})\right)^{-1}\mathbf{y}\mathbf{y}^{\mathrm{H}}\right)$ with the rank-1 sample covariance matrix $\mathbf{y}\mathbf{y}^{\mathrm{H}}$. Second, we observe that $\widetilde{\mathbf{R}}_y(\boldsymbol{\rho})$ has a block Toeplitz structure, where each $M\times M$ block on the kth block diagonal below the main block diagonal is given by $\boldsymbol{\Phi}\,\mathrm{E}\{\mathbf{x}[k'+k]\,\mathbf{x}^{\mathrm{H}}[k']\}\,\boldsymbol{\Phi}^{\mathrm{H}}$, for $k=-K+1,\ldots,K-1$. The likelihood function (7) can be closely approximated by replacing $\mathbf{y}\mathbf{y}^{\mathrm{H}}$ with a modified sample covariance matrix $\mathbf{S}$ that has the same block Toeplitz structure as $\widetilde{\mathbf{R}}_y(\boldsymbol{\rho})$. The blocks on the kth block diagonal below the main block diagonal of $\mathbf{S}$ are given by

$$\mathbf{S}[k] = \frac{1}{(K-k)}\sum_{k'=1}^{K-k}\mathbf{y}[k'+k]\,\mathbf{y}^{\mathrm{H}}[k']\,.$$

Due to their block Toeplitz structure, the matrices $\widetilde{\mathbf{R}}_y(\boldsymbol{\rho})$ and $\mathbf{S}$ contain a large amount of redundant information. The useful information is mainly concentrated around their main block diagonal. Motivated by these two observations, [7] proposes to crop the two matrices and use only their first $L\times L$ blocks of size $M\times M$, where L may be much smaller than K. Denoting the cropped $LM\times LM$ matrices by $\widetilde{\mathbf{R}}_{y,L}(\boldsymbol{\rho})$ and $\mathbf{S}_L$, we obtain the approximate likelihood

$$p(\mathbf{y}|\boldsymbol{\rho},\sigma^2) = \frac{1}{(\pi\sigma^2)^{LM}\,|\widetilde{\mathbf{R}}_{y,L}(\boldsymbol{\rho})|} \times\exp\left(-\frac{\mathrm{Tr}\left(\left(\widetilde{\mathbf{R}}_{y,L}(\boldsymbol{\rho})\right)^{-1}\mathbf{S}_L\right)}{\sigma^2}\right). \quad (15)$$

Note that reducing the dimensions of $\widetilde{\mathbf{R}}_y(\boldsymbol{\rho})$ also implies that $\tilde{r}_n$ needs to be calculated only for $n=0,\ldots,LN$ whenever $p(\mathbf{y}|\widetilde{\boldsymbol{\rho}},\sigma^2)$ is calculated (i.e., in every other iteration of the algorithm). Further note that, since $q_{\sigma^2}\big(\widetilde{\sigma}^2\big|\boldsymbol{\rho}^{(j-1)},(\sigma^2)^{(j-1)}\big)$ according to (11) is derived from $p(\mathbf{y}|\boldsymbol{\rho},\sigma^2)$, we must also



(a) (b) (c)

Fig. 1. Evaluation of $p(\mathbf{y}|\boldsymbol{\rho},\sigma^2)$ according to (15), where $\mathbf{y}$ was generated using $\rho_{1,\text{true}} = -0.7$, $\rho_{2,\text{true}} = -0.7$, and $\sigma^2_{\text{true}} = 0.26$. Light (dark) colors represent large (small) values of $p(\mathbf{y}|\boldsymbol{\rho},\sigma^2)$, which is shown as a function of $\boldsymbol{\rho} \in (-1,1)^2$ for (a) $\sigma^2 = \sigma^2_{\text{true}}$, (b) $\sigma^2 = 3\sigma^2_{\text{true}}$, (c) $\sigma^2 = \sigma^2_{\text{true}}/3$.

Algorithm 1 MH Sampler for Compressive AR Modeling

Initialize with any $\boldsymbol{\rho}^{(0)}$, $(\sigma^2)^{(0)}$ from $(-1,1)^p \times (0,\infty)$
$\widehat{\boldsymbol{\rho}} \leftarrow \boldsymbol{\rho}^{(0)}$, $\widehat{\sigma^2} \leftarrow (\sigma^2)^{(0)}$
Iterate for $j = 1,\ldots,J$:
  If j is even:
    $(\sigma^2)^{(j)} \leftarrow (\sigma^2)^{(j-1)}$
    Generate $\widetilde{\rho}_i$ from (13) for $i = 1,\ldots,p$
    Calculate α_j from (14) using (15) and (13)
    With probability α_j: $\boldsymbol{\rho}^{(j)} \leftarrow \widetilde{\boldsymbol{\rho}}$
    In the converse case: $\boldsymbol{\rho}^{(j)} \leftarrow \boldsymbol{\rho}^{(j-1)}$
  If j is odd:
    $\boldsymbol{\rho}^{(j)} \leftarrow \boldsymbol{\rho}^{(j-1)}$
    Generate $(\sigma^2)^{(j)}$ from (11) using (16)
  If $p\big(\boldsymbol{\rho}^{(j)},(\sigma^2)^{(j)}\big|\mathbf{y}\big) > p\big(\widehat{\boldsymbol{\rho}},\widehat{\sigma^2}\big|\mathbf{y}\big)$
    $\widehat{\boldsymbol{\rho}} \leftarrow \boldsymbol{\rho}^{(j)}$, $\widehat{\sigma^2} \leftarrow (\sigma^2)^{(j)}$

replace (12) by

$$\psi = LM - 1\,, \quad \gamma(\boldsymbol{\rho}) = \mathrm{Tr}\left(\big(\widetilde{\mathbf{R}}_{y,L}(\boldsymbol{\rho})\big)^{-1}\mathbf{S}_L\right). \qquad (16)$$

The resulting algorithm is summarized in Algorithm 1.

IV. Numerical results

To assess the performance of the proposed method, we generated 2500 AR signals of length $KN = 240000$ and order $p = 2$, with different AR parameters. For $\boldsymbol{\rho}$, we used the 25 elements of $\{-0.7,-0.4,0,0.2,0.9\}^2$, each for 100 AR signals. For each $\boldsymbol{\rho}$, we chose σ^2 such that $r_0 = 1$. Each AR signal was compressed using an individual compression matrix $\boldsymbol{\Phi}$, whose elements were randomly generated from a zero-mean circularly symmetric complex Gaussian distribution with variance 1. Different compression rates between 0.1 and 0.4 were achieved by using $M = 10$ and $N \in \{25,30,40,60,100\}$, where each combination of the different values of N and $\boldsymbol{\rho}$ was used for 20 signals. For one such signal $\mathbf{y}$ with $M/N = 0.1$, Fig. 1 shows the distribution $p(\mathbf{y}|\boldsymbol{\rho},\sigma^2)$ according to (15), using $L = 1$. The figure illustrates the potentially multimodal shape of $p(\mathbf{y}|\boldsymbol{\rho},\sigma^2)$. For higher model order p or for a smaller amount of data (i.e., a smaller number of observed blocks K), the potential number of local maxima increases.

For each AR signal, $\boldsymbol{\rho}$ and σ^2 were estimated according to Algorithm 1, using $L = 1$. One such estimate with $J = 20000$ iterations took about 11s in an unoptimized MATLAB R2011b 64-bit implementation on a 2.8-GHz Intel Core i7 processor. From each estimate $\widehat{\boldsymbol{\rho}}$, the corresponding AR coefficients $\widehat{\mathbf{a}}$ were calculated according to (3). As a performance benchmark, we compared the proposed method (abbreviated PM in the following) to the non-structured CSS method proposed in [5] (referred to as RM in the following). This method estimates r_n from $\mathbf{y}$ using a least-squares approach, without assuming a parametric model for r_n. From this estimate $\widehat{r}_{n,\text{LS}}$, we calculated $\widehat{\mathbf{a}}_{\text{RM}}$ and $\widehat{\sigma^2}_{\text{RM}}$ using the Yule-Walker equations (4) for $n = 0,\ldots,p$. As mentioned in [5], the RM requires full column rank of some matrices of size $M^2 \times N$. For a given M or N, this implies that the compression rate M/N cannot be reduced below a certain minimum. Conversely, for a given compression rate, this amounts to a minimum block size (M,N). Such strong restrictions do not apply to the PM (although the restrictions on (M,N) for the PM need to be studied further). In particular, in our simulations $(M,N) = (10,100)$ often led to failure of the RM because of the full rank condition. Therefore, we used $(M,N) = (12,120)$ instead of $(M,N) = (10,100)$ for the RM.

Fig. 2 shows the empirical normalized mean squared error (NMSE) of $\widehat{\mathbf{a}}$ and $\widehat{\sigma^2}$ of both the PM and the RM for different compression rates M/N. Each NMSE value was obtained from averaging over 500 AR signals. It can be seen that the PM performs consistently better than the RM for the given compression rates. In particular, the results show that the PM is significantly more robust to low compression rates than the RM.

Similar conclusions can be drawn from Fig. 3, which considers the parametric estimate of the autocorrelation function $\widehat{r}_n$ based on $\widehat{\mathbf{a}}$ and $\widehat{\sigma^2}$. From the AR parameter estimates of both the PM and the RM method, we calculated the respective $\widehat{\mathbf{r}}_{36} = (\widehat{r}_{-36},\ldots,\widehat{r}_{36})^{\mathrm{T}}$. Fig. 3 compares the NMSE of the two estimates for different compression rates M/N. Again, the

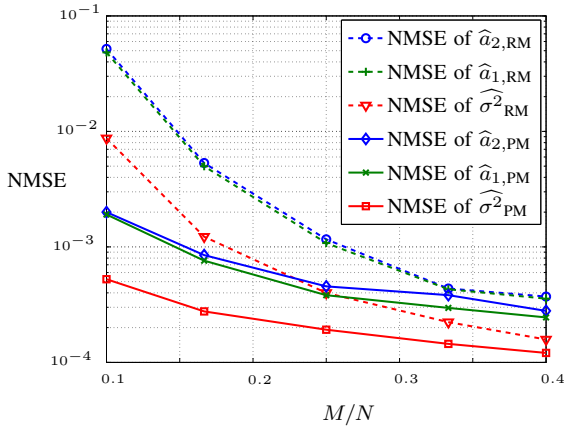


Fig. 2. Empirical NMSE of $\widehat{\mathbf{a}}$ and $\widehat{\sigma^2}$ for various compression rates M/N. Solid lines correspond to results of the proposed method, dashed lines correspond to results of the reference method. Each NMSE was obtained using 500 different AR signals.

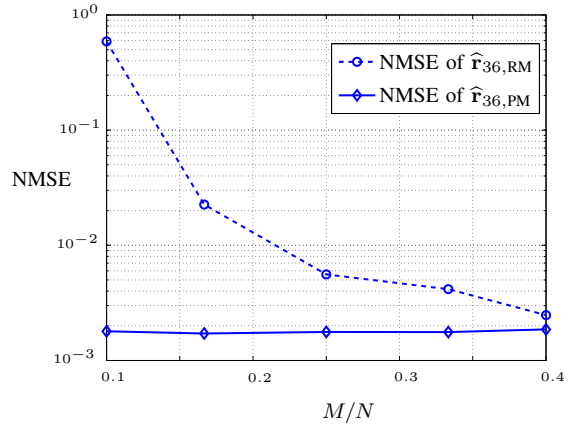


Fig. 3. Empirical NMSE of $\widehat{\mathbf{r}}_{36}$ for various compression rates M/N, comparing the proposed method (solid line) and the reference method (dashed line). Each NMSE was obtained using 500 different AR signals.

results confirm good performance of the PM and particularly show its high robustness with respect to strong compression.

As an illustrative example, Fig. 4 shows the Fourier transforms of $\mathbf{r}_{36}$ and $\widehat{\mathbf{r}}_{36}$, i.e., the power spectrum of one of the AR signals, along with its estimates. Note that, since the NMSE is invariant to multiplication of the parameter vector with a Fourier matrix $\mathbf{F}$, the NMSE shown in Fig. 3 is also the NMSE of the estimated power spectrum $\widehat{\mathbf{s}}_{36} = \mathbf{F}\,\widehat{\mathbf{r}}_{36}$.

V. Conclusion

We proposed a Bayesian algorithm for estimating AR parameters from compressed observations, which can be seen as a problem of structured CCS. Due to its nonlinearity, this problem cannot be solved efficiently by previous CCS methods. We presented an algorithm employing MH within Gibbs sampling, which is a powerful methodology but requires careful design of the iterative steps. Simulation results confirm the promising performance of the chosen design even for low compression rates.

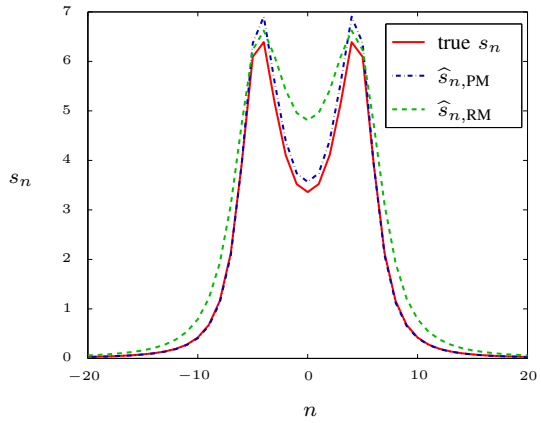


Fig. 4. Power spectrum s_n of an AR signal and estimates $\widehat{s}_n$ obtained with the proposed method (dash-dotted line) and the reference method (dashed line) at a compression rate $M/N = 0.17$.

VI. Acknowledgement

This work was supported by the Austrian Science Fund (FWF) under Grant J3495.

References

[1] J. P. Burg, "Maximum entropy spectral analysis," in *Proc. 37th Meeting Soc. Exploration Geophys.*, Oklahoma City, OK, USA, Oct. 1967.

[2] H. Akaike, "Power spectrum estimation through autoregressive model fitting," *Ann. Inst. Statist. Math.*, vol. 21, no. 1, pp. 407–419, 1969.

[3] X. Wu and X. Zhang, "Adaptive structured recovery of compressive sensing via piecewise autoregressive modeling," in *Proc. IEEE ICASSP-2010*, Dallas, TX, USA, March 2010, pp. 3906–3909.

[4] R. Bos, S. de Waele, and P. M. T. Broersen, "Autoregressive spectral estimation by application of the Burg algorithm to irregularly sampled data," *IEEE Trans. Instrum. Meas.*, vol. 51, no. 6, pp. 1289–1294, Dec. 2002.

[5] D. D. Ariananda and G. Leus, "Compressive wideband power spectrum estimation," *IEEE Trans. Signal Process.*, vol. 60, no. 9, pp. 4775–4789, Sept. 2012.

[6] D. D. Ariananda, D. Romero, and G. Leus, "Cooperative compressive power spectrum estimation," in *Proc. IEEE Sensor Array Multichannel Signal Process. Workshop (SAM)*, A Coruña, Spain, June 2014, pp. 97–100.

[7] D. Romero and G. Leus, "Wideband spectrum sensing from compressed measurements using spectral prior information," *IEEE Trans. Signal Process.*, vol. 61, no. 24, pp. 6232–6246, Dec. 2013.

[8] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer, New York, NY, USA, 2004.

[9] N. Dobigeon, J.-Y. Tourneret, and M. Davy, "Joint segmentation of piecewise constant autoregressive processes by using a hierarchical model and a Bayesian sampling approach," *IEEE Trans. Signal Process.*, vol. 55, no. 4, pp. 1251–1263, Apr. 2007.

[10] G. Kail, J.-Y. Tourneret, F. Hlawatsch, and N. Dobigeon, "Blind deconvolution of sparse pulse sequences under a minimum distance constraint: A partially collapsed Gibbs sampler method," *IEEE Trans. Signal Process.*, vol. 60, no. 6, pp. 2727–2743, June 2012.

[11] S. Chib and E. Greenberg, "Understanding the Metropolis-Hastings algorithm," *The American Statistitian*, vol. 49, pp. 327–335, 1995.

[12] Y. I. Abramovich, B. A. Johnson, and N. K. Spencer, "Two-dimensional multivariate parametric models for radar applications – Part I: Maximum-entropy extensions for Toeplitz-block matrices," *IEEE Trans. Signal Process.*, vol. 56, no. 11, pp. 5509–5526, Nov. 2008.

[13] Y. I. Abramovich, B. A. Johnson, and N. K. Spencer, "Two-dimensional multivariate parametric models for radar applications – Part II: Maximum-entropy extensions for Hermitian-block matrices," *IEEE Trans. Signal Process.*, vol. 56, no. 11, pp. 5527–5539, Nov. 2008.

[14] E. Punskaya, C. Andrieu, A. Doucet, and W. J. Fitzgerald, "Bayesian curve fitting using MCMC with applications to signal segmentation," *IEEE Trans. Signal Process.*, vol. 50, pp. 747–758, Mar. 2002.
[15] M. Mishali and Y. C. Eldar, "Sub-Nyquist sampling," *IEEE Signal Process. Magazine*, vol. 28, no. 6, pp. 98–124, Nov. 2011.
[16] M. H. Hayes, *Statistical digital signal processing and modeling*, Wiley, New York, NY, USA, 1996.
[17] M. J. L. de Hoon, T. H. J. J. van der Hagen, H. Schoonewelle, and H. van Dam, "Why Yule-Walker should not be used for autoregressive modelling," *Ann. Nuclear Energy*, vol. 23, no. 15, pp. 1219–1228, Oct. 1996.
[18] A. Gelman and D. B. Rubin, "Inference from iterative simulation using multiple sequences," *Statist. Science*, vol. 7, no. 4, pp. 457–472, Nov. 1992.