Regularized Parametric Models of Nonstationary Processes

Daniel Rudoy and Tryphon Georgiou

***Abstract*— In this article, we study two classes of nonstationary processes respectively parameterized by time-varying autoregressions and time-varying lattice filters. The processes considered are induced by solutions to certain convex optimization problems with local or global constraints, and are consistent with standard models of their stationary counterparts. We show that an underlying nesting property naturally leads to a family of hypothesis tests for stationarity and provide a geometric interpretation of our results on the manifold of all-pole rational transfer functions.**

I. Introduction

Time-varying autoregressions constitute a popular class of parametric models for nonstationary processes, and have been applied to a wide array of signal processing and control problems including speech analysis [1]–[3], EEG analysis [4] and radar processing [5]. Different classes of time-varying autoregressive (TVAR) models can be obtained depending on how the temporal evolution of the autoregressive coefficients is modeled. Previous approaches include using a functional basis expansion [1], [3], [6] or modeling each trajectory as a sample path of a suitably-chosen stochastic process (see, e.g., [7]).

In contrast, we consider TVAR modeling from a geometric point of view. Time-varying AR coefficients at time instants m and n implicitly define two "frozen-time" AR processes with power spectral densities f_m and f_n, which are functions of $\{a_1[m],\dots,a_p[m],\sigma_m^2\}$ and $\{a_1[n],\dots,a_p[n],\sigma_n^2\}$, respectively. This observation forms the basis of our approach to modeling the temporal evolution of TVAR coefficients—we constrain the coefficient variation from one timestep to the next by bounding a measure of distance between the induced AR models. We show that this formulation admits efficient estimators realized through convex optimization programs. Aspects of our approach are similar to the recent work of [8] in the context of signal segmentation, however, our presentation is more general touching on lattice parameterizations, tests for stationarity and Riemannian metrics on the manifold of AR processes.

To this end, we review the direct- and lattice-form representations of time-varying autoregressive processes in Section II, and discuss unconstrained estimation of their parameters in Section III. In Section IV, we narrow the set of nonstationary processes being under consideration by regularizing the temporal trajectories of the time-varying AR and lattice coefficients using local and global constraints, and derive the corresponding constrained estimators. In Section V, we show how to use these estimators to construct hypothesis tests for stationarity. Finally, in Section VI, we contrast the distance measures employed in our regularization framework with distance measures arising from Riemannian metrics on the space of AR processes. We conclude and briefly discuss future directions in Section VII.

The research is supported, in part, by the Air Force Office of Scientific Research under grant FA9550-09-1-0113, the National Science Foundation under grant 0701248, and a Graduate Research Fellowship.

Daniel Rudoy is with the School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA; rudoy@seas.harvard.edu

Tryphon Georgiou is with the Department of Electrical & Computer Engineering, University of Minnesota, Minneapolis, MN 55455, USA; tryphon@umn.edu

II. Time-Varying Autoregressive Models

Consider the following discrete-time difference equation with time-varying coefficients:

$$\text{TVAR}(p): \quad x[n] = \sum_{i=1}^{p} a_i[n]x[n-i] + \sigma w[n], \qquad (1)$$

which generalizes the classical autoregressive (AR) process and where the sequence $w[n]$ is a zero-mean white Gaussian sequence with unit variance scaled by a gain parameter $\sigma > 0$. The time-dependence of the coefficients $a_i[n]$ implies that the stochastic process specified by (1) is nonstationary.

Next, fix a positive integer j and define $\{a_{i,j}[n] \,|\, 1 \le i \le j\}$ and $\{b_{i,j}[n] \,|\, 1 \le i \le j\}$ to be the time-varying forward and backward linear prediction coefficients that minimize the squared errors of predicting $x[n]$ and $x[n-j]$, respectively. These forward and backward errors are defined according to:

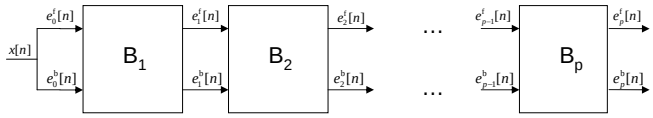
$$e_j^{\mathrm{f}}[n] \triangleq x[n] - \sum_{i=1}^{j} a_{i,j}[n]x[n-i], \qquad (2)$$

$$e_j^{\mathrm{b}}[n] \triangleq x[n-j] - \sum_{i=1}^{j} b_{i,j}[n]x[n-j+i]. \qquad (3)$$

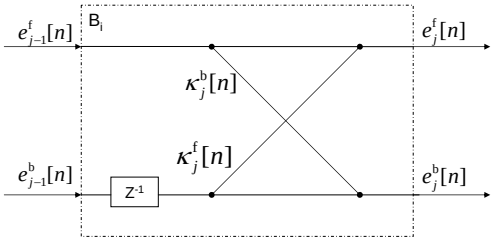
It is clear from (2) that $e_j^{\mathrm{f}}[n]$ is the error of approximating $x[n]$ by its projection onto the space spanned by $\{x[n-1], x[n-2], \dots, x[n-j]\}$. However, since the variables $\{x[n-1], x[n-2], \dots, x[n-j]\}$ are dependent, the expansion $\widehat{x}[n] \triangleq \sum_{i=1}^{j} a_{i,j}[n]x[n-i]$ is not orthogonal and, consequently, neither is the direct-form realization of (1).

An alternative realization of $\widehat{x}[n]$, using an *orthogonal* set of vectors, may be obtained by applying the Gram-Schmidt procedure to the variables $\{x[n-1], x[n-2], \dots, x[n-j]\}$ in order starting from $x[n-1]$ (see e.g., [9]). This yields a recursive way of computing the optimal forward and backward linear prediction coefficients of order j from those

ISBN 978-963-311-370-7



(a) Whitening lattice filter comprised of p sections



(b) The jth section of whitening lattice filter

Fig. 1. Diagram of whitening lattice filter with time-dependent forward and backward lattice coefficients $\kappa_j^{\mathrm{f}}[n]$ and $\kappa_j^{\mathrm{b}}[n]$, respectively

of order $j-1$ by:

$$a_{i,j}[n] = \begin{cases} a_{i,j-1}[n] + \kappa_j^{\mathrm{f}}[n] b_{j-i,j-1}[n-1] \text{ if } 1 \leq i < j \\ -\kappa_j^{\mathrm{f}}[n] \text{ if } i = j \end{cases}, \tag{4}$$

$$b_{i,j}[n] = \begin{cases} b_{i,j-1}[n-1] + \kappa_j^{\mathrm{b}}[n] a_{j-i,j-1}[n] \text{ if } 1 \leq i < j \\ -\kappa_j^{\mathrm{b}}[n] \text{ if } i = j \end{cases}, \tag{5}$$

corresponding to a generalization of the Levinson recursion. Here the forward and backward time-varying lattice (reflection) coefficients $\kappa_j^{\mathrm{f}}[n]$ and $\kappa_j^{\mathrm{b}}[n]$ are defined via:

$$\begin{aligned} \kappa_j^{\mathrm{f}}[n] &\triangleq -\frac{\langle e_{j-1}^{\mathrm{f}}[n], e_{j-1}^{\mathrm{b}}[n-1] \rangle}{\| e_{j-1}^{\mathrm{b}}[n-1] \|^2}, \text{ and} \\ \kappa_j^{\mathrm{b}}[n] &\triangleq -\frac{\langle e_{j-1}^{\mathrm{f}}[n], e_{j-1}^{\mathrm{b}}[n-1] \rangle}{\| e_{j-1}^{\mathrm{f}}[n] \|^2}. \end{aligned} \tag{6}$$

By substituting (4) and (5) into (2) and (3), respectively, we obtain the following familiar recursive lattice structure:

$$\begin{pmatrix} e_j^{\mathrm{f}}[n] \\ e_j^{\mathrm{b}}[n] \end{pmatrix} = \begin{pmatrix} 1 & \kappa_j^{\mathrm{f}}[n] \\ \kappa_j^{\mathrm{b}}[n] & 1 \end{pmatrix} \begin{pmatrix} e_{j-1}^{\mathrm{f}}[n] \\ e_{j-1}^{\mathrm{b}}[n-1] \end{pmatrix}. \tag{7}$$

Indeed, together with the natural initial condition of $e_0^{\mathrm{f}}[n] = e_0^{\mathrm{b}}[n] = x[n]$, the recursion of (7) corresponds to the time-varying whitening lattice filter shown in Fig. 1. This yields an orthogonal realization of a TVAR(p) process in contrast to the direct form realization of (1).

III. Unconstrained Estimation

If M sequences each consisting of N observations were available, with $M \gg N$, then the sample covariance matrix of the process could be easily estimated (see e.g., [5]). On the other hand, when only a single sequence of observations is available the problem is ill-posed without further constraints. Indeed, even short-time analysis presupposes that signal statistics are piecewise-constant which is not suitable in our general nonstationary setting. Consequently, we describe appropriately constrained estimators in Section IV below; in preparation, we first develop the unconstrained estimators in this section.

A. Time-varying Autoregressive Coefficients

Given N observations of the process $x[n]$ partitioned according to:

$$(\boldsymbol{x}_p \,|\, \boldsymbol{x}_{N-p})^T \triangleq \big(x[0] \cdots x[p-1] \;\mid\; x[p] \cdots x[N-1] \big)^T,$$

we would like to estimate the TVAR coefficients grouped into a vector $\boldsymbol{a} \in \mathbb{R}^{(N-p)p \times 1}$ according to:

$$\boldsymbol{a} \triangleq \begin{pmatrix} \boldsymbol{a}_p & \boldsymbol{a}_{p+1} & \cdots & \boldsymbol{a}_{N-1} \end{pmatrix}^T,$$

where $\boldsymbol{a}_m \triangleq \begin{pmatrix} a_1[m] & a_2[m] & \cdots & a_p[m] \end{pmatrix}$.

The unconditional likelihood of the TVAR coefficients and σ^2 can be factored according to:

$$\begin{aligned} p(\boldsymbol{x}_{N-p}, \boldsymbol{x}_p \,;\, \boldsymbol{a}_0, \ldots, \boldsymbol{a}_{N-1}, \sigma^2) &= p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p \,;\, \boldsymbol{a}, \sigma^2) \\ &\quad \times p(\boldsymbol{x}_p; \boldsymbol{a}_0, \ldots, \boldsymbol{a}_{p-1}, \sigma^2). \end{aligned}$$

Here the notation "|" reflects conditioning on random variables, whereas ";" indicates dependence of the density on non-random parameters. As is standard practice, we approximate the above unconditional data likelihood by the conditional likelihood $p(\boldsymbol{x}_{N-p} \,|\, \boldsymbol{x}_p \,;\, \boldsymbol{a}, \sigma^2)$. The "edge" effect of disregarding $p(\boldsymbol{x}_p; \boldsymbol{a}_0, \ldots, \boldsymbol{a}_{p-1}, \sigma^2)$ becomes less noticeable with increasing N.

Due to the Gaussianity of $w[n]$, maximizing the conditional likelihood is equivalent to solving the following least-squares problem:

$$\widehat{\boldsymbol{a}} = \underset{\boldsymbol{a}}{\text{minimize}} \; \| \boldsymbol{x}_{N-p} - \boldsymbol{H}_{\boldsymbol{x}} \boldsymbol{a} \|_2^2, \tag{8}$$

where $\boldsymbol{H}_{\boldsymbol{x}}$ is the appropriate data-dependent block-Hankel matrix. Its solution is readily obtained as $\widehat{\boldsymbol{a}} = \left(\boldsymbol{H}_{\boldsymbol{x}}^T \boldsymbol{H}_{\boldsymbol{x}} \right)^{-1} \boldsymbol{H}_{\boldsymbol{x}}^T \boldsymbol{x}_{N-p}$.

B. Time-Varying Reflection Coefficients

The time-varying reflection coefficients may also be directly estimated from data. Specifically, given N observations, we need to estimate the parameter vectors $\{\boldsymbol{\kappa}_1^{\mathrm{f}}, \boldsymbol{\kappa}_2^{\mathrm{f}}, \ldots, \boldsymbol{\kappa}_p^{\mathrm{f}}\}$ and $\{\boldsymbol{\kappa}_1^{\mathrm{b}}, \boldsymbol{\kappa}_2^{\mathrm{b}}, \ldots, \boldsymbol{\kappa}_p^{\mathrm{b}}\}$ where $\boldsymbol{\kappa}_j^{\mathrm{f}}, \boldsymbol{\kappa}_j^{\mathrm{b}} \in \mathbb{R}^{(N-j) \times 1}$ are defined for all $1 \leq j \leq p$ by

$$\begin{aligned} \boldsymbol{\kappa}_j^{\mathrm{f}} &\triangleq \begin{pmatrix} \kappa_j^{\mathrm{f}}[j] & \kappa_j^{\mathrm{f}}[j+1] & \cdots & \kappa_j^{\mathrm{f}}[N-1] \end{pmatrix}^T, \quad \text{and} \\ \boldsymbol{\kappa}_j^{\mathrm{b}} &\triangleq \begin{pmatrix} \kappa_j^{\mathrm{b}}[j] & \kappa_j^{\mathrm{b}}[j+1] & \cdots & \kappa_j^{\mathrm{b}}[N-1] \end{pmatrix}^T. \end{aligned}$$

We may group these first j sets of coefficients according to $\boldsymbol{\theta}_j^{\mathrm{f}} \triangleq \{\boldsymbol{\kappa}_1^{\mathrm{f}}, \boldsymbol{\kappa}_2^{\mathrm{f}}, \ldots, \boldsymbol{\kappa}_j^{\mathrm{f}}\}$ and $\boldsymbol{\theta}_j^{\mathrm{b}} \triangleq \{\boldsymbol{\kappa}_1^{\mathrm{b}}, \boldsymbol{\kappa}_2^{\mathrm{b}}, \ldots, \boldsymbol{\kappa}_j^{\mathrm{b}}\}$.

The estimation approach, structurally similar to the time-invariant case, is to estimate $\widehat{\boldsymbol{\kappa}}_j^{\mathrm{f}}$ and $\widehat{\boldsymbol{\kappa}}_j^{\mathrm{b}}$ only after $\widehat{\boldsymbol{\theta}}_{j-1}^{f}$ and $\widehat{\boldsymbol{\theta}}_{j-1}^{b}$ are obtained. In particular, suppose that $\boldsymbol{\theta}_{j-1}$ and consequently $e_{j-1}^{\mathrm{f}}[n]$ and $e_{j-1}^{\mathrm{b}}[n]$ are known. Then, in analogy to Burg's method, we estimate $\boldsymbol{\kappa}_j^{\mathrm{f}}$ and $\boldsymbol{\kappa}_j^{\mathrm{b}}$ to

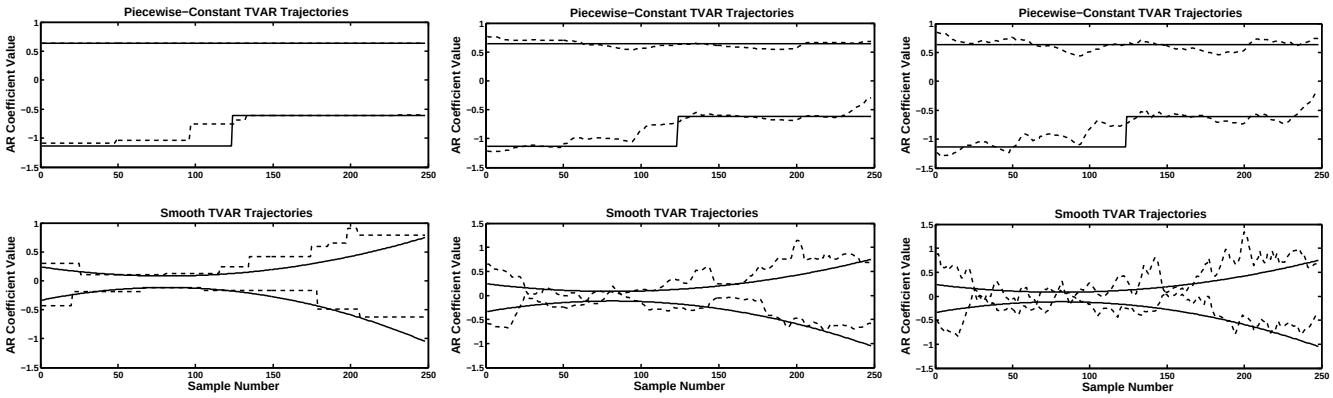


Fig. 2. Fitting TVAR processes with a constraint on the overall "path-length" of coefficient trajectories. The quadratic program of (13) is applied to two TVAR(2) processes with piecewise-constant (top panels) and smooth (bottom panels) coefficient trajectories. The three examples correspond to different norms used: $q = 1$ (left), $q = 1.5$ (middle) and $q = 2$ (right).

minimize, respectively, the sums of squared forward and backward prediction errors:

$$\begin{aligned}\widehat{\boldsymbol{\kappa}}_j^{\mathrm{f}} &= \underset{\boldsymbol{\kappa}_j^{\mathrm{f}}}{\operatorname{argmin}}\; L(\boldsymbol{\kappa}_j^{\mathrm{f}}; \boldsymbol{\theta}_{j-1}^{\mathrm{f}}) \triangleq \sum_{n=j-1}^{N-1} \|e_j^{\mathrm{f}}[n]\|^2, \\ \widehat{\boldsymbol{\kappa}}_j^{\mathrm{b}} &= \underset{\boldsymbol{\kappa}_j^{\mathrm{b}}}{\operatorname{argmin}}\; L(\boldsymbol{\kappa}_j^{\mathrm{b}}; \boldsymbol{\theta}_{j-1}^{\mathrm{b}}) \triangleq \sum_{n=j-1}^{N-1} \|e_j^{\mathrm{b}}[n]\|^2.\end{aligned} \tag{9}$$

The objective function of (9) may be conveniently rewritten in linear regression form. To this end, define the vectors $\boldsymbol{e}_j^{\mathrm{f}}$ and $\boldsymbol{e}_j^{\mathrm{b}} \in \mathbb{R}^{(N-j)\times 1}$ according to:

$$\begin{aligned}\boldsymbol{e}_j^{\mathrm{f}} &\triangleq \begin{pmatrix} e_j^{\mathrm{f}}[j] & e_j^{\mathrm{f}}[j+1] & \cdots & e_j^{\mathrm{f}}[N-1] \end{pmatrix}^T \\ \boldsymbol{e}_j^{\mathrm{b}} &\triangleq \begin{pmatrix} e_j^{\mathrm{b}}[j-1] & e_j^{\mathrm{b}}[j] & \cdots & e_j^{\mathrm{b}}[N-2] \end{pmatrix}^T,\end{aligned}$$

and the matrices $\boldsymbol{E}_j^{\mathrm{f}}, \boldsymbol{E}_j^{\mathrm{b}} \in \mathbb{R}^{(N-j)\times(N-j)}$ specified entrywise by: $\boldsymbol{E}^{\mathrm{f}}(m,n) \triangleq e_j^{\mathrm{f}}[m]\delta[m-n]$ and $\boldsymbol{E}^{\mathrm{b}}(m,n) \triangleq e_j^{\mathrm{b}}[m]\delta[m-n]$ for $1 \le m, n \le N-j$. Then we may write (9) according to:

$$\begin{aligned}\widehat{\boldsymbol{\kappa}}_j^{\mathrm{f}} &= \underset{\boldsymbol{\kappa}_j^{\mathrm{f}}}{\operatorname{argmin}}\; L(\boldsymbol{\kappa}_j^{\mathrm{f}}; \boldsymbol{\theta}_{j-1}^{\mathrm{f}}) \\ &= \left(\boldsymbol{e}_{j-1}^{\mathrm{f}} + \boldsymbol{E}_{j-1}^{\mathrm{b}}\boldsymbol{\kappa}_j^{\mathrm{f}}\right)^T \left(\boldsymbol{e}_{j-1}^{\mathrm{f}} + \boldsymbol{E}_{j-1}^{\mathrm{b}}\boldsymbol{\kappa}_j^{\mathrm{f}}\right), \\ \widehat{\boldsymbol{\kappa}}_j^{\mathrm{b}} &= \underset{\boldsymbol{\kappa}_j^{\mathrm{b}}}{\operatorname{argmin}}\; L(\boldsymbol{\kappa}_j^{\mathrm{b}}; \boldsymbol{\theta}_{j-1}^{\mathrm{b}}) \\ &= \left(\boldsymbol{e}_{j-1}^{\mathrm{b}} + \boldsymbol{E}_{j-1}^{\mathrm{f}}\boldsymbol{\kappa}_j^{\mathrm{b}}\right)^T \left(\boldsymbol{e}_{j-1}^{\mathrm{b}} + \boldsymbol{E}_{j-1}^{\mathrm{f}}\boldsymbol{\kappa}_j^{\mathrm{b}}\right).\end{aligned}$$

Consequently, the estimators are given by:

$$\begin{aligned}\widehat{\boldsymbol{\kappa}}_j^{\mathrm{f}} &= -\left({\boldsymbol{E}_{j-1}^{\mathrm{b}}}^T \boldsymbol{E}_{j-1}^{\mathrm{b}}\right) \boldsymbol{e}_{j-1}^{\mathrm{f}}, \quad \text{and} \\ \widehat{\boldsymbol{\kappa}}_j^{\mathrm{b}} &= -\left({\boldsymbol{E}_{j-1}^{\mathrm{f}}}^T \boldsymbol{E}_{j-1}^{\mathrm{f}}\right) \boldsymbol{e}_{j-1}^{\mathrm{b}},\end{aligned} \tag{10}$$

or pointwise according to:

$$\begin{aligned}\widehat{\kappa}_j^{\mathrm{f}}[n] &= -\frac{e_{j-1}^{\mathrm{f}}[n]e_{j-1}^{\mathrm{b}}[n-1]}{e_{j-1}^{\mathrm{b}}[n]e_{j-1}^{\mathrm{b}}[n]}, \quad \text{and} \\ \widehat{\kappa}_j^{\mathrm{b}}[n] &= -\frac{e_{j-1}^{\mathrm{f}}[n]e_{j-1}^{\mathrm{b}}[n-1]}{e_{j-1}^{\mathrm{f}}[n]e_{j-1}^{\mathrm{f}}[n]},\end{aligned}$$

which may be viewed as a simple plug-in estimator based on (6). This is the best estimator available in the absence of further assumptions.

IV. Constrained Estimation

Estimating $(N-p)p$ parameters from N observations is an ill-posed problem and so the estimators of (8) and (9) need to be constrained. Here we discuss two strategies based, respectively, on a set of *local* and *global* constraints on how fast the time-varying AR or lattice coefficients are allowed to vary.

A. Time-Varying AR Models

Local constraints on the TVAR coefficient trajectories may be imposed by bounding finite-difference approximations to the first d derivatives of each TVAR coefficient trajectory. In the case of $d = 1$, for instance, this yields the following optimization problem:

$$\begin{aligned}\widehat{\boldsymbol{a}} = \min_{\boldsymbol{a}}\; & \|\boldsymbol{x}_{N-p} - \boldsymbol{H}_{\boldsymbol{x}}\boldsymbol{a}\|_2^2 \\ \text{subject to}\quad & |a_i[n] - a_i[n-1]| \le \epsilon \\ & \epsilon \ge 0, \quad \forall\; 1 < n \le N;\; 1 \le i \le p.\end{aligned} \tag{11}$$

Note that, in this case, a time-invariant AR process may be recovered by setting $\epsilon = 0$.

The *local* constraints of (11) preclude large local fluctuations since the amount of coefficient variation per timestep is bounded. An alternative is to consider a *global* regularizer based on the total "length" of all AR trajectories captured via a q-norm-derived distance as follows:

$$L_q(\boldsymbol{a}) \triangleq \left(\sum_{i=1}^{p}\sum_{n=p+1}^{N-1} |a_i[n] - a_i[n-1]|^q\right)^{1/q}. \tag{12}$$

Indeed, one may think of a TVAR process as a path on the manifold of AR processes, with each "frozen-time" AR process as a point along the path, and of (12) as a proxy for measuring the path length—we discuss this interpretation further in Section VI. In this manner (12) gives rise to a class

of time-varying AR processes for some $C > 0$ described by the feasible set of solutions to:

$$\begin{aligned} \widehat{\boldsymbol{a}} = \min_{\boldsymbol{a}} \ & \|\boldsymbol{x}_{N-p} - \boldsymbol{H}_{\boldsymbol{x}}\boldsymbol{a}\|_2^2 \\ \text{subject to} \quad & L_q(\boldsymbol{a}) \leq C. \end{aligned} \tag{13}$$

In the statistics literature, when $q = 1$, the solution to the quadratic program of (13) may be obtained by an algorithm called the Lasso; its properties were originally studied in [10] in the context of shrinkage estimators.

Two alternatives to the quadratic program of (13) include minimizing the path-length while constraining the energy of the residual

$$\begin{aligned} \widehat{\boldsymbol{a}} = \min_{\boldsymbol{a}} \ & L_q(\boldsymbol{a}) \\ \text{subject to} \quad & \|\boldsymbol{x}_{N-p} - \boldsymbol{H}_{\boldsymbol{x}}\boldsymbol{a}\|_2^2 < C, \end{aligned} \tag{14}$$

or minimizing a dual-objective function as in:

$$\widehat{\boldsymbol{a}} = \min_{\boldsymbol{a}} \ \lambda L_q(\boldsymbol{a}) + (1-\lambda)\|\boldsymbol{x}_{N-p} - \boldsymbol{H}_{\boldsymbol{x}}\boldsymbol{a}\|_2^2, \tag{15}$$

for some $\lambda \in [0, 1]$. Setting $C = \sigma^2$ in (14) allows us to effectively measure the length of the process; however, σ^2 is not always known in practice. The sum-of-norms objective function of (15) has also been recently studied in the context of time-series segmentation algorithms [8].

We make the following observations about the problems (13), (14), and (15):

1) All three problems are *convex* when $1 < q \leq \infty$, and solutions may be efficiently found using freely-available software packages such as CVX [11]. Highly-efficient algorithms are available if $q \in \{1, 2, \infty\}$.
2) Using the ℓ_1-norm ($q = 1$) induces a sparse solution by penalizing the number of changes in value in the TVAR coefficient trajectories.
3) In the formulation of (13), a time-invariant AR process may be recovered by setting $C = 0$. But, the quadratic programs in (14) and (15) do not allow such a specialization for any value of C or λ.

As an illustration, the results of estimating two TVAR(2) processes, with piecewise-constant and smooth coefficient trajectories, using (13) with different norms are shown in Fig. 2. The value of the length upper bound C was set based on the true path length of the generated process, i.e., by calculating $L_q(\boldsymbol{a})$ using the true coefficient trajectories.

Our second example illustrates the application of (13) to a temporarily-unstable TVAR(2) process generated by filtering white Gaussian noise through a second-order digital resonator whose minimum-phase poles ($z, z^* \,|\, z = re^{-j\theta}\, r = .8\, \theta = \pi/4$) are briefly moved to their conjugate-reciprocal locations outside the unit circle resulting in an amplitude spike seen in the top panel of Fig. 3. As shown in the middle panel of Fig. 3, the power spectral density remains unchanged, since reflecting poles about the boundary of the unit circle only changes the phase of the filter and leaves the magnitude spectrum unchanged. Nonetheless, the ℓ_1-constrained estimator of (13) yields accurate estimates of the TVAR trajectories at every time instant.

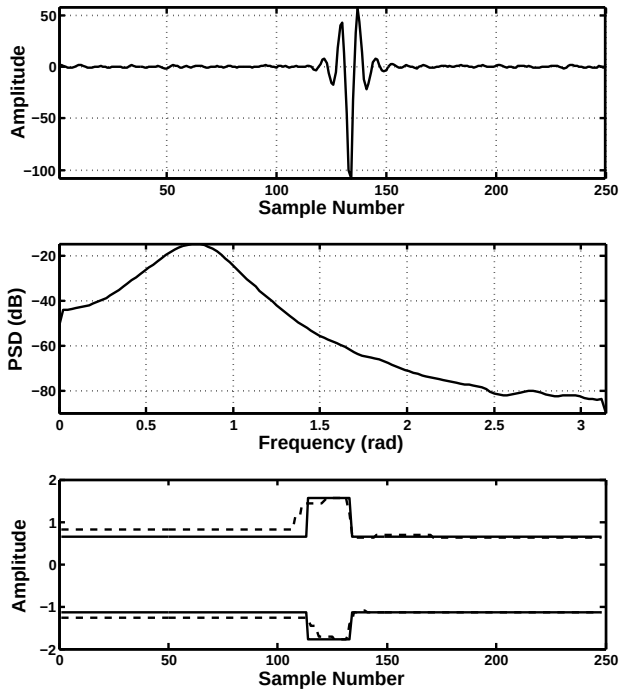


Fig. 3. Fitting a TVAR(2) process with an ℓ_1 constraint on the overall "path-length" of coefficient trajectories. The time-domain signal showing the temporary instability (top) is shown together with its estimated power spectral density (middle), and with the true (solid) and estimated (dashed) estimates of TVAR coefficient trajectories (bottom).

B. Time-varying Lattice Filters

In order to constrain the estimator of (10), we first make a standard assumption (see, e.g., [1]) that:

$$\|e_j^{\mathrm{f}}[n]\|^2 = \|e_j^{\mathrm{b}}[n-1]\|^2 \quad \text{for all} \quad 0 \leq n \leq N-1,$$

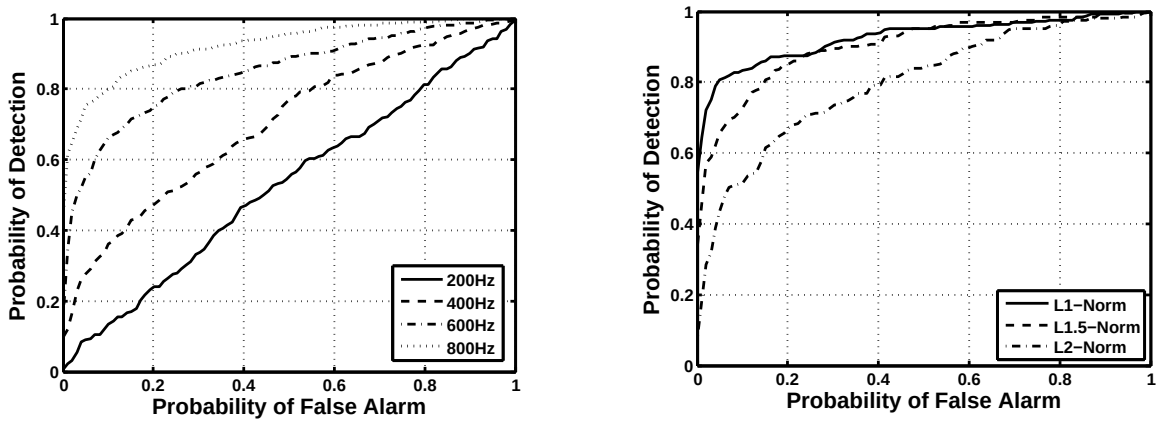
which together with (6) implies that $\kappa_j^{\mathrm{f}}[n] = \kappa_j^{\mathrm{b}}[n]$. This reduces the number of coefficients that need to be estimated by a factor of two. In this case, $\boldsymbol{\kappa}_j \triangleq \boldsymbol{\kappa}_j^{\mathrm{f}} = \boldsymbol{\kappa}_j^{\mathrm{b}}$ and the estimator of (10) becomes:

$$\widehat{\boldsymbol{\kappa}}_j = \operatorname*{argmin}_{\boldsymbol{\kappa}_j} L(\boldsymbol{\kappa}_j; \boldsymbol{\theta}_{j-1}) \triangleq \sum_{n=j-1}^{N-1} \|e_j^{\mathrm{f}}[n]\|^2 + \|e_j^{\mathrm{b}}[n]\|^2. \tag{16}$$

The estimator of (16) may be further constrained in the same vein as (11), (13), (14), or (15). For instance, in the case of a path-length constraint we obtain the following quadratic program (convex if $1 < q \leq \infty$) for some $C > 0$ for the jth time-varying reflection coefficient:

$$\begin{aligned} \widehat{\boldsymbol{\kappa}}_j = \operatorname*{argmin}_{\boldsymbol{\kappa}_j} & \sum_{n=j-1}^{N-1} \|e_j^{\mathrm{f}}[n]\|^2 + \|e_j^{\mathrm{b}}[n]\|^2 \\ \text{subject to} & \left(\sum_{i=1}^{p} \sum_{n=p+1}^{N-1} |\kappa_i[n] - \kappa_i[n-1]|^q \right)^{1/q} \leq C, \end{aligned} \tag{17}$$

where it is assumed that the first $j-1$ lattice coefficients have already been estimated. Since time-varying lattice coefficients are estimated iteratively, the quadratic program of (17)



(a) (b)

Fig. 4. Example of GLRT detection performance for a 100-sample synthetic TVAR(2) signal: (a) GLRT operating characteristics for various frequency jumps $\delta \in \{\pi/40, \pi/20, 3\pi/40, \pi/10\}$ radians; (b) GLRT operating characteristics for different norms $q \in \{1, 1.5, 2\}$.

has a factor of p (assuming p coefficients are desired) less constraints than the TVAR formulation of (13), which leads to overall computational savings.

V. Testing for Stationarity

Setting $\epsilon = 0$ in (11) or setting $C = 0$ in (13) and (17) constrains the feasible solution set to include only time-invariant AR processes. Assuming BIBO stability of the underlying process, this allows us to formulate statistical hypothesis tests for the *stationarity* of $x[n]$ as follows:

$$\begin{array}{lllll} \text{Local:} & \mathcal{H}_0: & \epsilon = 0 & \text{Global:} & \mathcal{H}_0: \quad C = 0 \\ & \mathcal{H}_1: & \epsilon = \epsilon_0 > 0 & & \mathcal{H}_1: \quad C = C_0 > 0 \end{array} \tag{18}$$

Given N observations $\boldsymbol{x}$, both of these hypothesis tests may be realized using a generalized likelihood-ratio test (GLRT) statistic according to:

$$T(\boldsymbol{x}) \triangleq 2 \ln \frac{\sup_{\boldsymbol{a},\sigma^2} p_{\mathcal{H}_1}(\boldsymbol{x}; \boldsymbol{a}, \sigma^2)}{\sup_{\boldsymbol{a},\sigma^2} p_{\mathcal{H}_0}(\boldsymbol{x}; \boldsymbol{a}, \sigma^2)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma, \tag{19}$$

where the subscript $p_{\mathcal{H}_i}(\boldsymbol{x}; \boldsymbol{\theta})$ denotes the likelihood of the parameters $\boldsymbol{\theta}$ given the data $\boldsymbol{x}$ under hypothesis $i \in \{0, 1\}$.

In the case of the direct form parametrization, the maximum likelihood estimates of $\boldsymbol{a}$ and σ^2 using (11) or (13) under $\mathcal{H}_1$ with $\epsilon = \epsilon_0$ or $C = C_0$, and under $\mathcal{H}_0$ using the same estimators with $\epsilon = 0$ or $C = 0$, respectively. In the case of an orthogonal parametrization, first an estimate of $\boldsymbol{\kappa}_1, \ldots, \boldsymbol{\kappa}_p$ is obtained under $\mathcal{H}_1$ using (17) and under $\mathcal{H}_0$ using (17) with $C = 0$. Then an estimate of $\boldsymbol{a}$ may be obtained from the estimated time-varying lattice parameters via the generalized Levinson recursion of (4) and (5). Finally note that in the case of a global regularizer on the coefficient variation, when (13) or (17) are employed, the test statistic depends on the choice of q in (12).

To illustrate typical behavior of the likelihood-ratio test statistic of (19), we consider a synthetic 100-sample TVAR(2) signal obtained by filtering white Gaussian noise through a second-order digital resonator. The resonator's center frequency is increased by δ radians halfway through the duration of the signal, while its bandwidth is kept constant. The detection performance of the GLRT statistic of (19), computed for $q = 1$ to encourage sparsity in the resultant estimates, is illustrated in the left panel of Fig. 4, which shows receiver operating characteristic (ROC) curves computed for different frequency jump sizes $\delta \in \{\pi/40, \pi/20, 3\pi/40, \pi/10\}$ radians (200 Hz increments). To generate data under $\mathcal{H}_0$, δ was set to zero and 500 trial simulations were performed for each combination. The value of C_0 was set based on the true path length of the generated process, i.e., by calculating $L_q(\boldsymbol{a})$ using the true coefficient trajectories.

In agreement with our intuition, detection performance improves when δ is increased—larger changes are easier to detect. We also considered the effect of the regularization norm q on the power of the likelihood ratio test—it is illustrated in the right panel of Fig. 4. As expected, the detection performance improves with decreasing q since the underlying signal is sparse.

Note that the formulation of the hypothesis tests in (18) requires the specification of the constants ϵ_0 and C_0, which is natural since these constants define nested classes of nonstationary processes. Estimating these constants from data to arrive at a hypothesis test of the form

$$\begin{array}{lllll} \text{Local:} & \mathcal{H}_0: & \epsilon = 0 & \text{Global:} & \mathcal{H}_0: \quad C = 0 \\ & \mathcal{H}_1: & \epsilon > 0 & & \mathcal{H}_1: \quad C > 0 \end{array} \tag{20}$$

is not a straightforward hypothesis testing question in our framework, since changing ϵ or C changes the underlying class of stochastic processes. It may be possible to address (20) through a multiple-testing approach whereby ϵ_0 and C_0 are systematically increased in the test of (18) until the null hypothesis is rejected.

VI. Relationship to Metrics on Manifold of Power Spectral Densities

Thus far we have considered classes of nonstationary processes defined by limiting variation of the time-varying AR or lattice coefficients using a coefficient-domain distance

measure such as (12). Below, we explore the relationship of these coefficient-domain distances to intrinsic (Riemannian) metrics on the manifold of AR processes. We touch upon these concepts from a point of view consistent with the spirit of the paper—in order to highlight viable alternatives for regularizing geodesics across "frozen-time" models of nonstationary processes.

A wide variety of metrics and so-called "distortion measures" have been used to quantify distance between stochastic processes and track drift of their characteristics [12], [13]. Early on, Itakura and Saito [14] presented the "error-matching measure"

$$d_{\mathrm{IS}}(f_0, f_1) \triangleq \int_{-\pi}^{\pi} \left| \frac{f_0}{f_1} - \log(\frac{f_0}{f_1}) - 1 \right| \frac{d\theta}{2\pi},$$

between power spectral densities f_0 and f_1, now commonly referred to as the Itakura-Saito distance. This indexing is consistent with the earlier development and suggests that $i \in \{0, 1\}$ represents two different points in time where signal statistics (e.g., power spectral densities) have been estimated from time-series data. Itakura also introduced the so-called "gain-optimized" distortion defined by:

$$d_{\mathrm{I}}(f_0, f_1) \triangleq \min_{\lambda \geq 0} d_{\mathrm{IS}}(f_0, \lambda f_1) = \log \int_{-\pi}^{\pi} \frac{f_0/\sigma_0^2}{f_1/\sigma_1^2} \frac{d\theta}{2\pi}, \quad (21)$$

where $\sigma_i^2 \triangleq \exp(\int_{-\pi}^{\pi} \log(f_i) \frac{d\theta}{2\pi})$ is the variance of the optimal one-step-ahead prediction error. The gain-optimized distortion $d_{\mathrm{I}}(f_0, f_1)$ quantifies the differences in "shape" not total power between f_0 and f_1. Note that in the case of the TVAR process of (1), we have that $\sigma_i^2 = \sigma^2$ for all $i \in \mathbb{Z}$ since we have assumed that the noise variance is time-invariant; we relax this assumption to obtain a more general treatment below.

An expression similar to (21) was obtained in [15, Proposition 3] based on considering the so called "degradation of the prediction error variance." This measure is the ratio of two prediction error variances $\sigma_{01}^2/\sigma_{00}^2$ where σ_{ij}^2 is the variance of the prediction error obtained when a random process with power spectrum f_i is predicted using the optimal predictor designed based on the power spectrum f_j. This measures evaluates how well the optimal predictor designed for one process works when applied to predicting the other; it is given by:

$$\rho(f_0, f_1) \triangleq \frac{\int_{-\pi}^{\pi} \frac{f_0(\theta)}{f_1(\theta)} \frac{d\theta}{2\pi}}{\exp\left(\int_{-\pi}^{\pi} \log\left(\frac{f_0(\theta)}{f_1(\theta)} \right) \frac{d\theta}{2\pi} \right)} = \log(d_{\mathrm{I}}(f_0, f_1)).$$

Clearly, $\rho(f_0, f_1)$ is the ratio of the *arithmetic* and *geometric* means of the quantity f_0/f_1. More interesting, however, is the fact that $\rho(\cdot, \cdot) - 1$ can be used to induce a Riemannian metric [15] on the topological space of power spectral densities. Indeed, for small perturbations Δ—thought of as an element of the tangent space of the power spectral densities—$\rho(f, f + \Delta)$ can be locally approximated by the metric

$$g_f(\Delta) \triangleq \sqrt{\int_{-\pi}^{\pi} \left(\frac{\Delta}{f} \right)^2 \frac{d\theta}{2\pi} - \left(\int_{-\pi}^{\pi} \frac{\Delta}{f} \frac{d\theta}{2\pi} \right)^2},$$

which defines the Riemannian structure. It turns out that geodesics are easy to compute in this setting, and that geodesic distances between power spectra take the form of (normalized) L_2-distances between their respective logarithms:

$$d(f_0, f_1) \triangleq \sqrt{\int_{-\pi}^{\pi} \left(\log \frac{f_0}{f_1} \right)^2 \frac{d\theta}{2\pi} - \left(\int_{-\pi}^{\pi} \log \frac{f_0}{f_1} \frac{d\theta}{2\pi} \right)^2},$$

which is similar to "log-spectral deviations" used in speech processing (see, e.g., [12, page 370]).

Specializing to the case of autoregressive spectra, let

$$f_i(\theta) \triangleq \frac{\sigma_i^2}{|1 + a_1[i] e^{-j\theta} + \ldots + a_p[i] e^{-jp\theta}|^2},$$

for $i \in \{0, 1\}$. Then $\rho(f_0, f_1)$ takes the following form:

$$\begin{aligned} \rho(f_0, f_1) &= \int_{-\pi}^{\pi} \frac{\left|1 + a_1[1] e^{-j\theta} + \ldots + a_p[1] e^{-jp\theta}\right|^2}{|1 + a_1[0] e^{-j\theta} + \ldots + a_p[0] e^{-jp\theta}|^2} \frac{d\theta}{2\pi} \\ &= \int_{-\pi}^{\pi} \left|1 + a_1[1] e^{-j\theta} + \ldots + a_p[1] e^{-jp\theta}\right|^2 \frac{f_0}{\sigma_0^2} \frac{d\theta}{2\pi} \\ &= \frac{\boldsymbol{a}_1^T \boldsymbol{R}_{p+1} \boldsymbol{a}_1}{\sigma_0^2}, \end{aligned}$$

where $\boldsymbol{a}_i \triangleq (1, a_1[i], \ldots, a_p[i])^T$, while $\boldsymbol{R}_{p+1}$ is the Toeplitz covariance matrix of the process with power spectral density f_0. Thus, if

$$\boldsymbol{\delta} \triangleq \begin{pmatrix} \delta_1 & \delta_2 & \cdots & \delta_p \end{pmatrix}^T$$

denotes a tangent direction in the space of autoregressive coefficients and quantifies a "small" perturbation or direction of change in the sense $\boldsymbol{a}_0 \to \boldsymbol{a}_1 = \boldsymbol{a}_0 + \begin{pmatrix} 0 & \boldsymbol{\delta}^T \end{pmatrix}^T$, then the degradation of predictive-error variance $\rho(f_0, f_1) - 1$ between the two nearby spectra is given by:

$$\rho(f_0, f_1) - 1 = \frac{\boldsymbol{\delta}^T \boldsymbol{R}_p \boldsymbol{\delta}}{\sigma^2}.$$

Therefore the metric

$$g_{AR}(\boldsymbol{a}_0, \delta) \triangleq \sqrt{\frac{\boldsymbol{\delta}^T \boldsymbol{R}_p \boldsymbol{\delta}}{\sigma^2}} \quad (22)$$

induces a natural Riemannian structure on the topological space of AR processes from the point of view of prediction theory.

It is crucial to observe the differences between the metric of (22) and the q-norm-based distance of (12). Both forms depend on the perturbation $\boldsymbol{\delta}$, but (22) also takes the location of the poles into account and is, therefore, more sensitive to small perturbations of poles near the unit circle than to perturbations of the poles near the origin—a desirable feature from a systems perspective. On the other hand, it is not clear how to use (22) in order to bound the rate of variation of time-varying AR or lattice coefficients in the manner of (13) because the resultant optimization problem is no longer

convex. Thus, it is desirable to obtain a convenient expression for geodesic distances between AR models based on (22) and an efficient approach to computing or approximating it.

VII. Discussion

We have presented a general framework for regularizing time-varying autoregressive processes by constraining the underlying coefficient trajectories so that the resultant estimation problems reduce to convex quadratic programs. We considered both direct- and orthogonal-form realizations of TVAR processes—the resultant estimation algorithms can be viewed as generalizations of the covariance and Burg methods popular in the linear prediction literature and the speech processing community. Nonstationary processes defined in this manner admit a nesting structure—stationary counterparts are a special case—that leads to natural hypothesis tests for stationarity. In addition, we contrasted the distance measures employed in our regularization framework with distance measures arising from Riemannian metrics on the space of AR processes.

This framework is readily extensible to time-varying vector AR processes, periodic AR processes and the case when multiple sequences of observations are available. The lattice-domain formulation of Section IV-B can be further constrained so that $|\kappa_i[n]| \leq 1$ for all $1 \leq i \leq p$ and $0 \leq n \leq N-1$ so that the resultant estimator would yield "frozen-time" stable models (in contrast to the example in Fig. 3) useful in implementing the shaping filter—that is the inverse system to the whitening filter of Fig. 1.

The most interesting direction, however, is to employ the Riemannian metric of (22) in lieu of (12) in order to define classes of nonstationary processes that are amenable to efficient estimation procedures. This will be the subject of our future work.

References

[1] Y. Grenier, "Time-dependent ARMA modeling of nonstationary signals," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 31, pp. 899–911, 1983.

[2] K. S. Nathan and H. F. Silverman, "Time-varying feature selection and classification of unvoiced stop consonants," *IEEE Trans. Speech Audio Process.*, vol. 2, pp. 395–405, 1994.

[3] D. Rudoy, T. F. Quatieri, and P. J. Wolfe, "Time-varying autoregressive tests for multiscale speech analysis," in *Proc. 10th Ann. Conf. Intl. Speech Commun. Ass.*, 2009. [Online]. Available: http://sisl.seas.harvard.edu

[4] M. Juntunen, I. J. Zervo, and J. P. Kaipio, "Root modulus constraints in autoregressive model estimation," *Circuit System Signal Process.*, vol. 17, pp. 709–718, 1998.

[5] Y. I. Abramovich, N. K. Spencer, and M. D. E. Turley, "Time-varying autoregressive (TVAR) models for multiple radar observations," *IEEE Trans. on Signal Process.*, vol. 55, pp. 1298–1311, 2007.

[6] T. S. Rao, "The fitting of non-stationary time series models with time dependent parameters," *J. Roy. Stat. Soc. B*, vol. 32, pp. 312–322, 1970.

[7] J. Vermaak, C. Andrieu, A. Doucet, and S. J. Godsill, "Particle methods for Bayesian modeling and enhancement of speech signals," *IEEE Trans. Speech Audio Process.*, vol. 10, pp. 173–185, 2002.

[8] H. Ohlsson, L. Ljung, and S. Boyd, "Segmentation of ARX-models using sum-of-norms regularization," *Automat., to appear*, 2010.

[9] B. Friedlander, "Lattice filters for adaptive processing," *Proc. IEEE*, vol. 8, pp. 829–866, 1982.

[10] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc. B*, vol. 58, pp. 267–288, 1996.

[11] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming (web page and software)," June 2009. [Online]. Available: http://stanford.edu/ boyd/cvx.

[12] R. M. Gray, A. Buzo, A. H. Gray, and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 28, pp. 3993–4003, 1980.

[13] M. Basseville and A. Benveniste, "Sequential detection of abrupt changes in spectral characteristics of digitial signals," *IEEE Trans. Inf. Theory*, vol. 29, pp. 709–724, 1983.

[14] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method," *Proc. 6th Intl. Congr. Acoust. Tokyo, Japan*, pp. C17–C20, 1968.

[15] T. T. Georgiou, "Distances between power spectral densities," *IEEE Trans. Signal Process.*, vol. 55, pp. 3993–4003, 2007.