Fast and Stable Subspace Tracking

Xenofon G. Doukopoulos, *Member, IEEE*, and George V. Moustakides, *Senior Member, IEEE*

***Abstract*—We consider the problem of adaptive subspace tracking, when the rank of the subspace we seek to estimate is assumed to be known. Starting from the Data Projection Method (DPM), which constitutes a simple and reliable means for adaptively estimating and tracking subspaces, we develop a fast and numerically robust implementation of DPM, which comes at a considerably lower computational cost. Most existing schemes track subspaces corresponding either to the largest or to the smallest singular values, while our DPM version can switch from one subspace type to the other with a simple change of sign of its single parameter. The proposed algorithm provides orthonormal vector estimates of the subspace basis that are *numerically stable* since they do not accumulate roundoff errors. In fact, our scheme constitutes the *first* numerically stable, low complexity, algorithm for tracking subspaces corresponding to the smallest singular values. Regarding convergence towards orthonormality our scheme exhibits the fastest speed among all other subspace tracking algorithms of similar complexity.**

***Index Terms*—Fast adaptive algorithms, numerically stable subspace tracking, orthogonal iteration, subspace tracking.**

I. Introduction

In this paper, we are interested in developing adaptive techniques capable of performing subspace tracking. Common key ingredient in such methodologies constitutes the assumption that the subspace rank is considered known. This is very common in the subspace tracking literature and therefore adopted in the present work as well. The interested reader may refer to [1] for information theoretic criteria, like Akaike Information Criterion (AIC) and Minimum Description Length (MDL) in order to cope with the rank selection problem. Adaptive subspace tracking algorithms are of great importance, since they find numerous applications as: telecommunication systems, adaptive filtering, direction of arrivals, antenna array processing, linear systems, etc.

A. Problem Definition

In a typical application of subspace-based adaptive signal processing we are receiving, sequentially, observation vectors $\mathbf{y}_n \in \mathbb{R}^N$. Assuming stationarity we define the observation covariance matrix as $\mathbf{R} = \mathbb{E}[\mathbf{y}_n \mathbf{y}_n^t]$. Applying a singular value decomposition (SVD) on $\mathbf{R}$, we can write

$$\mathbf{R} = [\mathbf{U}_\mathrm{s} \quad \mathbf{U}_\mathrm{n}] \begin{bmatrix} \mathbf{\Lambda}_\mathrm{s} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Lambda}_\mathrm{n} \end{bmatrix} [\mathbf{U}_\mathrm{s} \quad \mathbf{U}_\mathrm{n}]^t \tag{1}$$

where $\mathbf{\Lambda}_\mathrm{s} = \mathrm{diag}\{\lambda_1, \ldots, \lambda_{L_\mathrm{s}}\}$ and $\mathbf{\Lambda}_\mathrm{n} = \mathrm{diag}\{\lambda_{L_\mathrm{s}+1}, \ldots, \lambda_N\}$. The diagonal elements $\lambda_i, i = 1, \ldots, N$ are the singular values of $\mathbf{R}$ satisfying

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \geq 0. \tag{2}$$

Hence, the diagonal matrices $\mathbf{\Lambda}_\mathrm{s}$, $\mathbf{\Lambda}_\mathrm{n}$ contain the L_s largest and the $L_\mathrm{n} = N - L_\mathrm{s}$ smallest singular values of $\mathbf{R}$, respectively, while $\mathbf{U}_\mathrm{s}$, $\mathbf{U}_\mathrm{n}$ contain the corresponding singular vectors. The matrices $\mathbf{U}_\mathrm{s}$, $\mathbf{U}_\mathrm{n}$ are both orthonormal, constituting *orthonormal bases* for the corresponding subspaces and they are orthogonal to each other since $\mathbf{U}_\mathrm{s}^t \mathbf{U}_\mathrm{n} = 0$. The problem we would like to solve is as follows.

> Assuming that the data sequence $\{\mathbf{y}_n\}$ is available sequentially, we would like to provide adaptive estimates either for $\mathbf{U}_\mathrm{s}$ or for $\mathbf{U}_\mathrm{n}$.

Perhaps the most common data case encountered in practice corresponds to the following signal plus noise model:

$$\mathbf{y}_n = \mathbf{x}_n + \mathbf{w}_n \tag{3}$$

where $\{\mathbf{x}_n\}$ is a sequence of length-N vectors lying on an L_s-dimensional linear subspace and $\{\mathbf{w}_n\}$ are independent and identically distributed (white) noise vectors with independent elements. In this case, the SVD of (1) takes the special form

$$\mathbf{R} = [\mathbf{U}_\mathrm{s} \quad \mathbf{U}_\mathrm{n}] \begin{bmatrix} \mathbf{D}_s + \sigma^2 \mathbf{I}_{L_\mathrm{s}} & \mathbf{0} \\ \mathbf{0} & \sigma^2 \mathbf{I}_{L_\mathrm{n}} \end{bmatrix} [\mathbf{U}_\mathrm{s} \quad \mathbf{U}_\mathrm{n}]^t \tag{4}$$

where $\mathbf{I}_K$ denotes the identity matrix of size K; $\mathbb{E}[\mathbf{x}_n \mathbf{x}_n^t] = \mathbf{U}_\mathrm{s} \mathbf{D}_\mathrm{s} \mathbf{U}_\mathrm{s}^t$ is the SVD of the covariance matrix of $\mathbf{x}_n$ with $\mathbf{D}_\mathrm{s}$ nonsingular and $\mathbb{E}[\mathbf{w}_n \mathbf{w}_n^t] = \sigma^2 \mathbf{I}_N$ the covariance matrix of $\mathbf{w}_n$ with σ^2 the noise power. Matrix $\mathbf{U}_\mathrm{s}$ is then said to span the *signal* subspace, whereas $\mathbf{U}_\mathrm{n}$ spans its complement, i.e., the *noise* subspace.

The algorithm we are going to develop in Section III along with the corresponding analysis of its convergence properties can be applied to the general data case. However, with a slight abuse of notation, we will call the subspace corresponding to the largest singular values "the signal subspace" and the subspace corresponding to the smallest singular values "the noise subspace," keeping of course in mind that these names originate from the signal plus noise data model of (3).

Manuscript received October 12, 2006; revised September 1, 2007. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Vitor Heloiz Nascimento.

X. G. Doukopoulos is with DibCom, Paris, France (e-mail: xenofon.doukopoulos@dibcom.fr).

G. V. Moustakides is with the Department of Electrical and Computer Engineering, University of Patras, 26500 Rion, Greece (e-mail: moustaki@ece.upatras.gr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2007.909335

B. Literature Review

Before proceeding with a description of the existing work, let us first classify the subspace tracking algorithms with respect to their computational complexity. A straightforward way to estimate the subspace of interest is to apply an SVD on the covariance matrix $\mathbf{R}$. This approach, known as *direct SVD*, requires $O(N^3)$ operations. In order to avoid the excessively high computational complexity needed by direct SVD, alternative schemes requiring less operations were developed. If L denotes the rank of the (signal or noise) subspace we would like to estimate, since usually $L \ll N$, we propose the following classification of the existing algorithms according to their computational complexity. Schemes requiring $O(N^2L)$ or $O(N^2)$ operations will be classified as *high complexity*; algorithms with complexity $O(NL^2)$ as *medium complexity* and finally methods with $O(NL)$ operations as *low complexity*. The last category constitutes the most important one from a real time implementation point of view, and schemes belonging to the low complexity class are also known in the literature as *fast subspace tracking algorithms*.

The literature referring to the problem of subspace tracking is extremely rich. The survey article [2] constitutes an excellent review of results up to 1990, treating the first two classes, since the last class was not available at the time. It is clearly the continuous interest in the subject and significant recent developments that gave rise to the third class. In the present work, we mainly emphasize on this low complexity class for both signal and noise subspace tracking, while we briefly address the most important schemes of the other two classes. An exhaustive literature review can be found in [3, pp. 30–43].

The great majority of articles addressing the problem of subspace tracking focus mainly on signal subspace, while the literature intended for the noise subspace is unfortunately very limited. Starting from the former, let us first introduce the most important methods belonging to the high complexity class. Owsley [4] was the first to introduce an adaptive procedure in order to update signal subspace estimates with $O(N^2L)$ operations. His adaptation was basically the application of an Orthogonal Iteration variant (which is going to be introduced in Section II) that uses the exponentially windowed sample autocorrelation matrix as an estimate of the covariance matrix $\mathbf{R}$. According to [2], the version that presents the overall best performance is a Lanczos-type algorithm. The same type of algorithm, is also proposed in [5], [6] adopting the Rayleigh-Ritz approximation of [7]. All these Lanczos-type schemes require $O(N^2L)$ operations. Based on the interesting idea of rank revealing decompositions, first proposed in [8], alternative subspace tracking techniques were developed requiring $O(N^2)$ complexity [9]–[11].

Let us now continue with signal subspace tracking algorithms belonging to the medium complexity $O(NL^2)$ class. The Data Projection Method (DPM) [12] is an important representative of this class. Since this algorithm will serve as the basis for our novel fast algorithmic scheme we defer its detailed presentation for the next section. The most popular algorithm of the medium complexity class was proposed by Karasalo [13]. Karasalo's algorithm offers the best performance to cost ratio [2] and thus serves as a point of reference for all subsequent low complexity $O(NL)$ techniques. Its overall complexity is $O(NL^2 + L^3)$, with the L^3 part coming from the need to perform an SVD on an $(L+1) \times (L+2)$ matrix.

Focusing on the low complexity class, the Projection Approximation Subspace Tracking (PAST) algorithm is a very well known approach for signal subspace tracking proposed in [14]. The main advantage of this $O(NL)$ scheme is its simple structure having a single parameter to be specified and its main handicap its inability to provide orthonormal estimates. The next two algorithms of interest are MALASE [15] and PROTEUS-2 [16]. Both algorithms have a rather complicated structure but they provide orthonormal estimates for the subspace basis. The first has four parameters and the second one that need to be tuned. The next algorithm, known as the Low Rank Adaptive Filter (LORAF), was proposed in [17] and came out by suitably projecting the observed data onto the signal subspace instead of the complete data space. Its fastest version (of complexity $O(NL)$), known under the acronym LORAF-3, has one parameter to be specified and provides orthonormal estimates. Finally, in a recent work [18], we have the algorithm FAPI which has a single parameter and provides orthonormal estimates with a very high overall performance. A very serious drawback of all previous algorithms is the fact that they do not have a noise subspace tracking counterpart.

The noise subspace tracking literature is *very limited* compared to the wide variety of methodologies offered for signal subspace estimation. However, noise subspace tracking can be of equal importance for many applications. For example, in the domain of telecommunications, there exist several blind channel estimation methods that rely on the knowledge of the noise subspace basis as in [19] for CDMA and in [20], [21] for OFDM systems. In [22], the SQRI minor subspace tracking scheme is proposed. This method employs a sequential inverse iteration based on square root QR update of the covariance matrix. It exhibits extremely good performance characteristics however, due to its computational complexity which is $O(N^2L)$, it belongs to the high complexity class. The estimated subspace is orthonormal at each iteration to machine accuracy. In [23], an algorithm for tracking the minor subspace of a correlation matrix is proposed. Due to the fact that the data autocorrelation matrix is explicitly calculated and used in matrix-vector multiplications the algorithm presents an overall complexity of $O(pN^2)$. However, in the special case of time series analysis it can be reduced to $O(pNL)$, where p corresponds to the number of additional dimensions (with respect to L) in which the new estimate of the desired basis is searched and is equal to 1 or 2 according [23]. The resulting algorithm presents a similar performance with [22], however, as it is stated in [23], it deviates very slowly from orthonormality.

As far as low complexity noise subspace tracking is concerned, it is interesting to point out that, in general, it is not possible to obtain such an algorithm by a straightforward modification of a signal subspace scheme. Existing noise subspace algorithms of complexity $O(NL)$, are *all* problematic, either exhibiting instability (divergence) or nonrobustness (numerical instability due to roundoff error accumulation). In the first case, we have [24], which diverges in noise subspace tracking. Efforts to eliminate its instability were attempted in [25] and [26]

by forcing the estimates to be orthonormal at each time step. Although there is a definite improvement in the stability characteristics, the resulting schemes are nonrobust, deviating slowly from orthonormality due to round-off error accumulation.

Another interesting scheme is the FRANS algorithm of [27], which we discuss in detail in the sequel, and which combines the DPM method with a fast orthonormalization procedure. However, FRANS is not stable and diverges fast from onrthonormality as it is also argued in [3], [28]. In [29] the application of Householder transformation in order to cope with FRANS's divergence is proposed. The new algorithm, called HFRANS presents indeed a much more stable behavior. This approach is different from the one we propose in the present manuscript. Unfortunately, even HFRANS continues to diverge from orthonormality (much slower than FRANS). This is proved in [30], where upper limits concerning the progress of the divergence are derived for both FRANS and HFRANS. Furthermore it is also interesting to mention that if orthonormality is lost, HFRANS is incapable of regaining this desirable property.

The fast DPM scheme we present in this work, was first developed for complex arithmetic, and successfully applied to the problem of adaptive OFDM channel estimation [21]. Here we introduce its real arithmetic version and focus on performance comparisons with all other existing low complexity schemes. We mainly insist on convergence and robustness under finite precision arithmetic. Regarding both issues, our algorithm will be equipped with a number of analytical results that sufficiently capture the essence of its somewhat unique behavior (especially in the case of robustness). Thus, a much more complete analysis of FDPM can be found here than its initial presentation in [28].

In [31], the authors apply our orthonormalization procedure (i.e., the one proposed form FDPM in [28]) and finally arrive in correcting the instability exhibited by all existing variants of the OJA algorithm. The FOOJA of [31] does not perform any better than our scheme despite its higher computational complexity.

II. Orthogonal Iteration and Variants

In this section, we present the Orthogonal Iteration and two of its variants that are capable of estimating the subspaces corresponding to both the largest and smallest singular values. The Orthogonal Iteration [32, p. 353] is the simplest iterative procedure, coming from Numerical Analysis, which can be used to compute the singular vectors corresponding to the L dominant singular values of a symmetric, nonnegative definite matrix.

Lemma 1: Let $\mathbf{R}$ be a symmetric, nonnegative definite matrix of size N, denote with $\lambda_1 \geq \ldots \geq \lambda_L > \lambda_{L+1} \geq \ldots \geq \lambda_N \geq 0$ its singular values and with $\mathbf{u}_1, \ldots, \mathbf{u}_N$ the corresponding singular vectors. Consider the sequence of matrices $\{\mathbf{U}_n\}$ of dimensions $N \times L$, defined by the iteration

$$\mathbf{U}_n = \text{orthnorm}\{\mathbf{R}\mathbf{U}_{n-1}\}, \quad n = 1, 2, \ldots \tag{5}$$

where "orthnorm" stands for orthonormalization using QR-decomposition or the modified Gram-Schmidt procedure, then

$$\lim_{n\to\infty} \mathbf{U}_n = [\mathbf{u}_1 \ \cdots \ \mathbf{u}_L] \tag{6}$$

provided that the matrix $\mathbf{U}_0^t[\mathbf{u}_1 \cdots \mathbf{u}_L]$ is not singular.

Proof: The proof can be found in [32, pp. 410, 411] ■.

For the orthogonal iteration to converge, it is imperative that $\lambda_L > \lambda_{L+1}$. Whenever the latter is valid, the convergence is exponential and proportional to $(\lambda_{L+1}/\lambda_L)^n$ or, equivalently, with an *exponential rate* equal to $-\log(\lambda_{L+1}/\lambda_L)$. According to the above definition, it becomes clear that positive rates correspond to convergence, while negative to divergence. At this point, several remarks are necessary.

Remark 1: If some of the L largest singular values coincide, then the corresponding singular vectors are not unique. In such case the Orthogonal Iteration converges to a basis in the corresponding subspace.

Remark 2: If instead of QR or Gram-Schmidt, we use any other orthonormalization procedure, the sequence $\{\mathbf{U}_n\}$ converges to an orthonormal basis in the subspace spanned by the first L singular vectors. The latter is, of course, unimportant whenever the L largest singular values are all equal to each other and, thus, the vectors of any orthonormal basis are also singular vectors.

We proceed with the presentation of two variants of the orthogonal iteration that allow for adaptive implementations. The two proposed variants are the following:

$$\mathbf{U}_n = \text{orthnorm}\{(\mathbf{I}_N + \mu\mathbf{R})\mathbf{U}_{n-1}\}, \quad n = 1, 2, \ldots \tag{7}$$

$$\mathbf{U}_n = \text{orthnorm}\{(\mathbf{I}_N - \mu\mathbf{R})\mathbf{U}_{n-1}\}, \quad n = 1, 2, \ldots \tag{8}$$

where $\mu > 0$ is a "small" scalar parameter known as *step size*. We observe that the two variants differ only in the sign that precedes μ. Notice that $\mathbf{I}_N \pm \mu\mathbf{R}$ has the same singular vectors as $\mathbf{R}$, while the singular values are equal to $1 \pm \mu\lambda_i$. In the "+" case, since $\mu > 0$, the eigenvalues have the same ordering as the original λ_i. This suggests that (7) will converge to the singular vectors corresponding to the L largest singular values, exactly as in (5). In the "−" case on the other hand, for μ sufficiently small (so that the matrix is nonnegative definite), the singular values are ordered in exactly the *opposite* way as compared to the original λ_i. This forces (8) to converge to the subspace of $\mathbf{R}$ corresponding to the L smallest singular values. Regarding convergence, it is again exponential but at a rate equal to $-\log\{(1+\mu\lambda_{L+1})/(1+\mu\lambda_L)\} = \mu(\lambda_L - \lambda_{L+1}) + o(\mu)$ for (7) and $-\log\{(1-\mu\lambda_{N-L})/(1-\mu\lambda_{N-L+1})\} = \mu(\lambda_{N-L} - \lambda_{N-L+1}) + o(\mu)$ for (8). In both cases, the convergence rate is of the form $\mu c + o(\mu)$ where $c > 0$. Since for sufficiently small μ the rate is positive (provided of course that $\lambda_L > \lambda_{L+1}$ or $\lambda_{N-L} > \lambda_{N-L+1}$), the corresponding algorithm is stable.

A. Adaptive Orthogonal Iteration

When matrix $\mathbf{R}$ is unknown and, instead, we acquire the data vector sequence $\{\mathbf{y}_n\}$ sequentially, we can replace $\mathbf{R}$ in (7) and (8) with an *adaptive estimate* $\mathbf{R}_n$ that satisfies $\mathbb{E}[\mathbf{R}_n] = \mathbf{R}$. This leads to the *Adaptive Orthogonal Iteration* algorithm

$$\mathbf{U}_n = \text{orthnorm}\{(\mathbf{I}_N \pm \mu\mathbf{R}_n)\mathbf{U}_{n-1}\} \tag{9}$$

where the "+" sign generates estimates for the signal subspace and the "−" sign for the noise subspace. Depending on the choice of $\mathbf{R}_n$ we can obtain alternative subspace tracking algorithms.

Clearly the simplest selection for $\mathbf{R}_n$ is the *instantaneous* estimate of the covariance matrix, that is, $\mathbf{R}_n = \mathbf{y}_n\mathbf{y}_n^t$ which gives rise to the Data Projection Method (DPM) first introduced in [12]

$$\mathbf{r}_n = \mathbf{U}_{n-1}^t\mathbf{y}_n \tag{10}$$

$$\mathbf{T}_n = \mathbf{U}_{n-1} \pm \mu\mathbf{y}_n\mathbf{r}_n^t \tag{11}$$

$$\mathbf{U}_n = \text{orthnorm}\{\mathbf{T}_n\} \tag{12}$$

with the orthonormalization performed using Gram-Schmidt. Due to this latter requirement, it is clear that the overall computational complexity of DPM is equal to $O(NL^2)$. This simple adaptive scheme will serve as a starting point for developing our subspace tracking algorithm in the next section. In fact the goal is to replace the Gram-Schmidt part with a faster orthonormalization procedure that will reduce the overall complexity to $O(NL)$.

B. Projection Operator and Convergence Analysis

Let us first denote with $\mathbf{U}$ the subspace basis we would like to estimate; i.e., $\mathbf{U} = \mathbf{U}_\text{s}$ if the "+" sign is employed in (11) or $\mathbf{U} = \mathbf{U}_\text{n}$ if instead we use the "−" sign. Notice that in the majority of applications involving subspace tracking we are mainly interested in the *projection* of the data sequence $\{\mathbf{y}_n\}$ onto the subspace of interest and not so much in the subspace basis $\mathbf{U}$ itself. If $\mathbf{U}_n$ is an estimate of $\mathbf{U}$ then $\mathbf{U}_n\mathbf{U}_n^t$ constitutes an estimate of the *projection operator* while the ideal projector is equal to $\mathbf{U}\mathbf{U}^t$.

Regarding convergence of $\mathbf{U}_n\mathbf{U}_n^t$ towards $\mathbf{U}\mathbf{U}^t$ there are two issues that are of interest: transient and steady-state behavior. Transient phase is characterized through the *mean convergence* of the estimation scheme, in other words how $\mathbb{E}[\mathbf{U}_n\mathbf{U}_n^t]$ converges towards $\mathbf{U}\mathbf{U}^t$. Steady-state behavior, on the other hand, is quantified through the steady-state *projection error power*. To define this quantity let us first introduce the projection error power at time n as

$$e_p(n) = \mathbb{E}\left[\left\|\mathbf{U}_n\mathbf{U}_n^t - \mathbf{U}\mathbf{U}^t\right\|_\text{F}^2\right] \tag{13}$$

where $\|\mathbf{A}\|_\text{F} = \sqrt{\text{tr}\{\mathbf{A}^t\mathbf{A}\}}$ denotes the Frobenius norm of the matrix $\mathbf{A}$. Then projection error power at steady state is simply the limit $\lim_{n\to\infty} e_p(n)$. We have the following lemma that treats the two points of interest for the case of the adaptive orthogonal iteration.

Theorem 1: Consider the adaptation defined in (9) where $\mathbf{R}_n$ is any adaptive estimate of $\mathbf{R}$ satisfying $\mathbb{E}[\mathbf{R}_n] = \mathbf{R}$, then the mean projector trajectory $\mathbb{E}[\mathbf{U}_n\mathbf{U}_n^t]$ tends to the ideal projector $\mathbf{U}\mathbf{U}^t$ at an exponential rate of the form $\mu c_1 + o(\mu)$; where $c_1 > 0$ is a positive constant (stability) independent from μ; furthermore the steady-state projection error power is of the form $\mu c_2 + o(\mu)$, where $c_2 > 0$ is also a positive constant independent from μ.

Proof: The proof can be found in Appendix I. ∎

As we can see from Theorem 1, any increase in the step size μ produces an equivalent increase in the convergence rate and in the steady-state projection error power. Due to this fact, the value of μ in the adaptive algorithm becomes a compromise between the two conflicting requirements of having a fast converging algorithm and a small steady-state projection error power (this is a general observation for *all* adaptive estimation algorithms with constant step size).

C. Alternative Implementations

DPM is a special case of the adaptive orthogonal iteration in (9) with $\mathbf{R}_n = \mathbf{y}_n\mathbf{y}_n^t$. One might wonder whether it is possible to improve subspace tracking by using less crude estimates of $\mathbf{R}$. Such possibility constitutes, for example, the sliding window sample covariance matrix

$$\mathbf{R}_n = \frac{1}{K}\sum_{i=0}^{K-1}\mathbf{y}_{n-i}\mathbf{y}_{n-i}^t. \tag{14}$$

With $K > 1$, (14) becomes a better estimate of $\mathbf{R}$ than the simple outer product $\mathbf{y}_n\mathbf{y}_n^t$. It is therefore only natural to expect that this will translate into better subspace estimates when used in (9). As we shall see in the next lemma, this intuitively plausible conjecture is in fact *false*.

Lemma 2: Consider the adaptation in (9) with

$$\mathbf{R}_n = \sum_{i=0}^{K-1}\alpha_i\mathbf{y}_{n-i}\mathbf{y}_{n-i}^t;\ \text{where}\ \sum_{i=0}^{K-1}\alpha_i = 1;\ \alpha_i \geq 0. \tag{15}$$

Then, to a first-order approximation in μ, the mean projector trajectory $\mathbb{E}[\mathbf{U}_n\mathbf{U}_n^t]$ and the steady-state projection error power $\lim_{n\to\infty}\mathbb{E}[\|\mathbf{U}_n\mathbf{U}_n^t - \mathbf{U}\mathbf{U}^t\|_\text{F}^2]$ are independent from the specific sequence $\{\alpha_i\}$ used in the adaptation.

Proof: The proof can be found in Appendix I. ∎

From Lemma 2 we deduce that by using the more efficient estimate of $\mathbf{R}$ given by (14) or more generally by (15), the adaptive orthogonal iteration exhibits very similar performance as the simple DPM algorithm. Indeed this is true since the transient and the steady-state phase, to a first-order approximation in μ, have exactly the same statistics. We are going to corroborate this theoretical finding with the first simulation example in Section IV. In fact, with the simulation we are going to realize that these algorithms, not only match their statistics but they also have very similar trajectories. Let us now continue with the introduction of the fast orthonormalization versions of DPM.

III. Fast Orthonormalizations of DPM

Going from complexity $O(NL^2)$ to $O(NL)$ is possible by *sacrificing exact orthonormality at each step*. In a subspace tracking problem, however, orthonormality is a leading and particularly desirable characteristic. Therefore the relaxation of this property must be made very cautiously. Indeed the majority of existing $O(NL)$ schemes attempt to comply with two very strict requirements regarding orthonormality:

R1) When the initial estimate is orthonormal, then *all* subsequent estimates must satisfy the same property.

R2) If for some reason orthonormality is lost (when for example we initialize with a nonorthonormal matrix), then the scheme must *converge*, as quickly as possible, to an orthonormal matrix.

The first requirement is very susceptible to nonrobust behavior (round-off error accumulation). For the second, it is clear that we would like convergence towards orthonormality to be

significantly faster than the time required by the estimates to converge (in the mean) to the true subspace. As we are going to see the fast DPM version we are going to propose here, avoids robustness problems and exhibits an extremely high convergence speed towards orthonormality.

Besides our own orthonormalization procedure we are also going to present and analyze the FRANS algorithm of [27] which constitutes an alternative means of fast orthonormalization of DPM. As we will demonstrate in the sequel, FRANS exhibits robustness problems and significantly slower convergence towards orthonormality as compared to our adaptive scheme.

We recall that DPM is the result of using the crude estimate for the data autocorrelation matrix $\mathbf{R}_n = \mathbf{y}_n\mathbf{y}_n^t$. According to [3], for the algorithm to converge we need to select a step size μ such that $0 < \mu \ll (1/\lambda_1)$, where λ_1 is the largest singular value of $\mathbf{R}$. To satisfy this requirement we propose the use of a normalized step size $\mu = \bar{\mu}/\mathrm{trace}\{\mathbf{R}\}$. We know that $\mathrm{trace}\{\mathbf{R}\} \geq \lambda_1$, however most of the time we have $\mathrm{trace}\{\mathbf{R}\} \gg \lambda_1$, therefore, selecting $\bar{\mu}$ even close to unity results in $\mu \ll 1/\lambda_1$. Since $\mathrm{trace}\{\mathbf{R}\}$ is not available, we can replace it by an estimate. There are two possibilities, we can either replace $\mathbf{R}$ by its instantaneous estimate or apply a simple exponential window. More precisely we can select

$$\mu = \frac{\bar{\mu}}{\|\mathbf{y}_n\|^2};\ \text{or}\ \mu = \frac{\bar{\mu}}{\sigma_{\mathbf{y}}^2(n)},$$
$$\text{where}\ \sigma_{\mathbf{y}}^2(n) = \nu\sigma_{\mathbf{y}}^2(n-1) + (1-\nu)\|\mathbf{y}_n\|^2 \tag{16}$$

and $0 < \nu < 1$ is an exponential forgetting factor.

Going back to DPM (10)–(12), we observe that we can write

$$\mathbf{U}_n = \mathbf{T}_n\mathbf{H}_n = \left(\mathbf{I}_N \pm \mu\mathbf{y}_n\mathbf{y}_n^t\right)\mathbf{U}_{n-1}\mathbf{H}_n \tag{17}$$

where matrix $\mathbf{H}_n$ is responsible for performing (exact or approximate) orthonormalization. It is the different choices of $\mathbf{H}_n$ that give rise to alternative fast versions. Let us first find a recursion for $\mathbf{U}_n^t\mathbf{U}_n$ using (17), expecting that this will guide us in properly selecting $\mathbf{H}_n$. We have

$$\mathbf{U}_n^t\mathbf{U}_n = \mathbf{H}_n^t\left(\mathbf{U}_{n-1}^t\mathbf{U}_{n-1} + \delta_n\mathbf{r}_n\mathbf{r}_n^t\right)\mathbf{H}_n \tag{18}$$

where $\delta_n = (\pm 2\mu + \mu^2\|\mathbf{y}_n\|^2)$. Clearly the goal is to make the product $\mathbf{U}_n^t\mathbf{U}_n = \mathbf{I}$, thus assuring orthonormality for our estimates. Recursion (18) is the key relation to define FRANS.

A. The FRANS Algorithm

FRANS can be derived from (18) by selecting $\mathbf{H}_n$ to satisfy Requirement R1. In other words, if $\mathbf{U}_{n-1}$ is orthonormal then $\mathbf{H}_n$ must be selected so that $\mathbf{U}_n$ enjoys the same property. Using (18) this means that $\mathbf{H}_n$ must satisfy the equality

$$\mathbf{I}_L = \mathbf{H}_n^t\left(\mathbf{I}_L + \delta_n\mathbf{r}_n\mathbf{r}_n^t\right)\mathbf{H}_n. \tag{19}$$

A straightforward selection that fulfills (19) is

$$\begin{aligned}\mathbf{H}_n &= \left(\mathbf{I}_L + \delta_n\mathbf{r}_n\mathbf{r}_n^t\right)^{-1/2}\\ &= \mathbf{I}_L - \frac{1}{\|\mathbf{r}_n\|^2}\left(1 - \frac{1}{\sqrt{1+\delta_n\|\mathbf{r}_n\|^2}}\right)\mathbf{r}_n\mathbf{r}_n^t\end{aligned} \tag{20}$$

which gives rise to the FRANS algorithm [27]

$$\mathbf{r}_n = \mathbf{U}_{n-1}^t\mathbf{y}_n \tag{21}$$
$$\mathbf{T}_n = \mathbf{U}_{n-1} \pm \mu\mathbf{y}_n\mathbf{r}_n^t \tag{22}$$
$$\rho_n = \frac{1}{\|\mathbf{r}_n\|^2}\left(1 - \frac{1}{\sqrt{1+(\pm 2\mu + \mu^2\|\mathbf{y}_n\|^2)\,\|\mathbf{r}_n\|^2}}\right) \tag{23}$$
$$\mathbf{U}_n = \mathbf{T}_n - \rho_n(\mathbf{T}_n\mathbf{r}_n)\mathbf{r}_n^t \tag{24}$$

and we recall that μ is normalized according to (16). The adaptation depicted in (21)–(24), has complexity $O(NL)$. As we are going to see in the next section, in the case of signal subspace estimation, this scheme is robust (not accumulating round-off errors) and stable (when initialized with a nonorthonormal matrix it converges to an orthonormal one). When however we use the algorithm to estimate the noise subspace, we have exactly the opposite behavior since FRANS becomes nonrobust and unstable. In other words when initialized with an orthonormal matrix, because of round-off error accumulation, after some point orthonormalization is lost and, even worse, when initialized with a nonorthonormal matrix it *diverges*. Even in the case of signal subspace estimation where the FRANS algorithm is robust and stable [33], the speed of convergence towards orthononormality is of the *same order of magnitude* as the speed of convergence of the estimates $\mathbf{U}_n$ towards their limit $\mathbf{U}$. It turns out that this latter fact is common practice to all existing algorithms of complexity $O(NL)$, except the one proposed next.

B. The Fast DPM Algorithm

In order to derive our scheme we are going to follow a slightly different logic than the one applied to derive FRANS. Instead of requiring, as in (17), the matrix $\mathbf{Z}_n = \mathbf{T}_n\mathbf{H}_n$ to be orthonormal, we simply ask it to be *orthogonal*. Normalization of the columns of $\mathbf{Z}_n$ is then a process that requires $O(NL)$ operations, assuring that the overall target computational complexity will not be exceeded.

Now, let us see how we can select $\mathbf{H}_n$. Forming the product

$$\mathbf{Z}_n^t\mathbf{Z}_n = \mathbf{H}_n^t\left(\mathbf{U}_{n-1}^t\mathbf{U}_{n-1} + \delta_n\mathbf{r}_n\mathbf{r}_n^t\right)\mathbf{H}_n \tag{25}$$

we recall that this is not equal to $\mathbf{U}_n^t\mathbf{U}_n$ as in FRANS, since the columns of $\mathbf{Z}_n$ need normalization to produce $\mathbf{U}_n$. The idea is to make this product a *diagonal* (instead of identity) matrix, property that assures orthogonality of the columns of $\mathbf{Z}_n$. We will select $\mathbf{H}_n$ so that requirement R1 is valid. According to R1 if we assume that $\mathbf{U}_{n-1}$ is orthonormal, then (25) takes the form

$$\mathbf{Z}_n^t\mathbf{Z}_n = \mathbf{H}_n^t\left(\mathbf{I}_L + \delta_n\mathbf{r}_n\mathbf{r}_n^t\right)\mathbf{H}_n. \tag{26}$$

Notice now that if $\mathbf{H}_n$ is an *orthonormal* matrix having as first column the vector $\mathbf{r}_n/\|\mathbf{r}_n\|$ (and, thus, the remaining $L-1$ columns containing vectors orthogonal to $\mathbf{r}_n$), then (26) becomes the following diagonal matrix:

$$\mathbf{Z}_n^t\mathbf{Z}_n = \mathbf{I}_L + \delta_n\|\mathbf{r}_n\|^2\mathbf{e}_1\mathbf{e}_1^t \tag{27}$$

where $\mathbf{e}_1 = [1\ 0\cdots 0]^t$. It is fortunate that there exists a very well-known and widely used matrix in Numerical Analysis with

TABLE I
STEPS OF THE FDPM ADAPTIVE ALGORITHM

Fast Data Projection Method (FDPM)
Initialization: Start with an orthonormal matrix $\mathbf{U}_0$.
Available from previous instant: $\mathbf{U}_{n-1}$. New data: $\mathbf{y}_n$.
Apply:
1. Compute μ from (16)
2. $\mathbf{r}_n = \mathbf{U}_{n-1}^t \mathbf{y}_n$
3. $\mathbf{T}_n = \mathbf{U}_{n-1} + \mu \mathbf{y}_n \mathbf{r}_n^t$ for signal subspace
   $\mathbf{T}_n = \mathbf{U}_{n-1} - \mu \mathbf{y}_n \mathbf{r}_n^t$ for noise subspace

If subspace rank $L > 1$
4. $\mathbf{a}_n = \mathbf{r}_n - \|\mathbf{r}_n\|\mathbf{e}_1$
5. $\mathbf{Z}_n = \mathbf{T}_n - \frac{2}{\|\mathbf{a}_n\|^2}[\mathbf{T}_n \mathbf{a}_n]\mathbf{a}_n^t$
6. $\mathbf{U}_n = \text{normalize}\{\mathbf{Z}_n\}$

If subspace rank $L = 1$
4. $\mathbf{U}_n = \frac{\mathbf{T}_n}{\|\mathbf{T}_n\|}$

the desired properties for $\mathbf{H}_n$. We refer to the *Householder Reflector* matrix [32, p. 195] which is defined as follows:

$$\mathbf{a}_n = \mathbf{r}_n - \|\mathbf{r}_n\|\mathbf{e}_1 \tag{28}$$

$$\mathbf{H}_n = \mathbf{I}_L - \frac{2}{\|\mathbf{a}_n\|^2}\mathbf{a}_n\mathbf{a}_n^t \tag{29}$$

and gives rise to our fast DPM (FDPM) algorithm. The complete adaptation is summarized in Table I, where "normalize" stands for normalization of the columns of the matrix $\mathbf{Z}_n$. The overall complexity is[1] $(6NL + 3L)$ multiplications, $6NL$ additions, $L + 1$ divisions, and $L + 1$ square roots. The strong points of our FDPM algorithm are as follows.

- It has a simple structure with a single parameter (the step size) to be specified.
- By changing the sign in front of μ we can switch from signal to noise subspace estimation (this property is inherited from DPM).
- It is robust with respect to finite precision for both signal and noise subspace estimation (there is no round-off error accumulation).
- It converges to an orthonormal matrix when it is initialized with a nonorthonormal one (stability).
- Finally, it has an extremely high convergence rate towards orthonormality, which is the fastest among all competing schemes of the same $O(NL)$ computational complexity.

The weak point of our algorithm is:

- It provides an orthonormal basis for the desired subspace and *not* the corresponding singular vectors. Actually most of the other algorithms have exactly the same problem. This weakness however is not very crucial since in the majority of applications we are interested in *projecting* the data onto the signal or noise subspace and not in finding the actual singular vectors. Projection can be performed with the help of *any* orthonormal basis that spans the subspace since all bases produce the *same* projection operator. There are even very common situations where this problem is unavoidable as in the signal plus noise model described in (3), when we are interested in the noise subspace. Since in the noise subspace there exists only one (multiple) singular value (the noise power), the singular vectors are not unique, therefore, the vectors of any basis spanning the noise subspace can play the role of singular vectors.

[1]This complexity is obtained if for a scalar α and vectors $\mathbf{x}, \mathbf{y}$ we perform $\alpha\mathbf{x}\mathbf{y}^t$ as $(\alpha\mathbf{x})\mathbf{y}^t$ and $\mathbf{x}\mathbf{y}^t/\alpha$ as $((1/\alpha)\mathbf{x})\mathbf{y}^t$.

C. Stability and Robustness Analysis of FRANS and FDPM

The goal is to analyze the divergence of the estimate $\mathbf{U}_n$ from orthonormality or equivalently the divergence of the product $\mathbf{U}_n^t\mathbf{U}_n$ from the identity matrix. Notice that the orthonormality error is always present due to finite precision arithmetic and possible initialization with a nonorthonormal matrix. By writing $\mathbf{U}_n^t\mathbf{U}_n = \mathbf{I}_L + \mathcal{E}_n$, where $\mathcal{E}_n$ expresses the difference from the identity matrix, we are going to study the behavior of $\mathcal{E}_n$ in each algorithmic case.

Let us first consider FRANS which is simpler to analyze. From (18) we can write

$$\mathcal{E}_n = \mathbf{H}_n^t\mathcal{E}_{n-1}\mathbf{H}_n + \mathcal{V}_n \tag{30}$$

where $\mathbf{H}_n$ is given in (20) and $\mathcal{V}_n$ denotes the finite precision errors due to numerical operations at step n. We have the following theorem that describes the error performance of FRANS.

Theorem 2: FRANS is robust and stable in the case of signal subspace estimation and nonrobust and unstable in the case of noise subspace estimation. The exponential rate of convergence towards orthonormality is of the form $\mu c + o(\mu)$, where $c > 0$ in the signal subspace (stability) and $c < 0$ in the noise subspace (instability).

Proof: The proof can be found in Appendix II. ■

Let us now proceed with FDPM. Substituting in (25) $\mathbf{U}_{n-1}^t\mathbf{U}_{n-1} = \mathbf{I}_L + \mathcal{E}_{n-1}$, we obtain

$$\mathbf{Z}_n^t\mathbf{Z}_n = \mathbf{I}_L + \delta_n\|\mathbf{r}_n\|^2\mathbf{e}_1\mathbf{e}_1^t + \mathbf{H}_n^t\mathcal{E}_{n-1}\mathbf{H}_n. \tag{31}$$

If $\mathbf{D}_n = \text{diag}\{\mathbf{Z}_n^t\mathbf{Z}_n\}$ (a diagonal matrix containing the diagonal elements of the product), then $\mathbf{Z}_n$ can be normalized to produce $\mathbf{U}_n$ as $\mathbf{U}_n = \mathbf{Z}_n\mathbf{D}_n^{-1/2}$. We can, thus, write

$$\mathbf{U}_n^t\mathbf{U}_n = \mathbf{D}_n^{-1/2}\left(\mathbf{I}_L + \delta_n\|\mathbf{r}_n\|^2\mathbf{e}_1\mathbf{e}_1^t + \mathbf{H}_n^t\mathcal{E}_{n-1}\mathbf{H}_n\right)\mathbf{D}_n^{-1/2} + \mathcal{V}_n. \tag{32}$$

Replacing the left-hand side (LHS) with $\mathbf{I}_L + \mathcal{E}_n$ we obtain

$$\mathcal{E}_n = \mathbf{D}_n^{-1/2}\left(\mathbf{H}_n^t\mathcal{E}_{n-1}\mathbf{H}_n - \text{diag}\left\{\mathbf{H}_n^t\mathcal{E}_{n-1}\mathbf{H}_n\right\}\right)\mathbf{D}_n^{-1/2} + \mathcal{V}_n \tag{33}$$

with the last equality being true because $\mathbf{I}_L + \delta_n\|\mathbf{R}_n\|^2\mathbf{e}_1\mathbf{e}_1^t$ is diagonal. From the same equality we also observe that the diagonal elements of $\mathcal{E}_n$ contain only round-off errors (no accumulation) due to the normalization of the columns of $\mathbf{Z}_n$ performed at each step. The analysis of (33) is summarized in the following theorem.

Theorem 3: FDPM is stable and robust for both signal and noise subspace estimation. Regarding exponential convergence towards orthonormality the corresponding rate can be written as $c + o(1)$, with $c > 0$ (stability) and independent from μ.

Proof: The proof can be found in Appendix III. ■

In order to understand the significance of Theorems 2 and 3, let us define the orthonormality error power

$$e_o(n) = \mathbb{E}\left[\|\mathcal{E}_n\|_{\text{F}}^2\right] = \mathbb{E}\left[\left\|\mathbf{U}_n^t\mathbf{U}_n - \mathbf{I}_L\right\|_{\text{F}}^2\right]. \tag{34}$$

If $\mathcal{E}_n$ has an exponential rate of convergence equal to $r(\mu)$, this means that $e_o(n) = O(e^{-2r(\mu)n})$. Consequently, if $r(\mu) > 0$ the error power converges to zero (stability), whereas if $r(\mu) < 0$ the error power diverges (instability). There is another interesting property associated with the rate of convergence. Assume that we are in the stable case where $r(\mu) > 0$ and let M_ϵ be the time required by the error power to attain a small value $\epsilon > 0$ which we regard as convergence in the practical sense. Then we have that $-\log\epsilon = 2r(\mu)M_\epsilon + O(1)$. Now consider the same algorithm with two different step sizes μ_1 and μ_2 and let M_1, M_2 be the iterations required by the algorithm to attain the same error power level ϵ under the two different step sizes. It is then clear that we can write $-\log\epsilon = 2r(\mu_1)M_1 + O(1) = 2r(\mu_2)M_2 + O(1)$, from which we obtain

$$\frac{M_1}{M_2} = \frac{r(\mu_2)}{r(\mu_1)} + o(1) \tag{35}$$

since $M_i \gg 1$. In other words, we end up with the fairly evident conclusion that in an exponentially stable adaptation, convergence time is *inversely proportional* to the rate of convergence.

Let us now see how the previous outcome translates in the case of FRANS and FDPM. For FRANS, since $r(\mu) = \mu c + o(\mu)$, this means that we can write (35) as

$$\frac{M_1}{M_2} = \frac{\mu_2}{\mu_1} + o(1). \tag{36}$$

In other words, convergence time towards orthonormality is inversely proportional to the step size μ. This conclusion has the following interesting consequence: if we decide to reduce μ by a factor $\kappa > 1$ (such need arises whenever we desire to lower the steady-state projection error power by a factor κ), then this will induce an increase by the same factor κ in the time required to converge towards orthonormality.

The situation is drastically different in the case of FDPM where $r(\mu) = c + o(1)$. Relation (35) then becomes

$$\frac{M_1}{M_2} = 1 + o(1). \tag{37}$$

Since for small step sizes we have $o(1) \ll 1$, this suggests that even if we make a drastic change in μ, this will produce only a marginal change (because of $o(1)$) in the time required to attain orthonormality. In other words, convergence towards orthonormality is practically insensitive to changes in the step size. This very desirable characteristic, as we shall see in Section IV, is enjoyed only by our algorithmic scheme. The other algorithms have performance similar to FRANS, i.e., satisfy (36), this being true even for Karasalo's method of complexity $O(NL^2)$.

IV. Simulation Comparisons

Let us begin our simulations section by corroborating the conclusion of Lemma 2. Specifically, with our first simulation we are going to verify that by using better adaptive estimates for $\mathbf{R}$, we do not necessarily improve subspace estimation with the adaptive orthogonal iteration. We consider three different sliding window implementations according to (14) with $K = 1, 10$, and 100. Case $K = 1$ corresponds to DPM. We apply (9) with the "$-$" sign in order to estimate the noise subspace. Our data vector $\mathbf{y}_n$ is of length $N = 8$. We adopt the signal plus noise model of (3) where the random signal $\mathbf{x}_n$ lies on an $L = 4$ dimensional linear subspace. We assume that the singular values of the signal subspace are $\mathbf{D}_s = \mathrm{diag}\{250, 180, 120, 60\}$ and the additive noise is white Gaussian of variance $\sigma^2 = 2$. This value induces an SNR level of approximately 15 dB, which is a very common selection when simulating communication systems. The step size that was employed in this example is equal to $\mu = 5 \times 10^{-5}$.

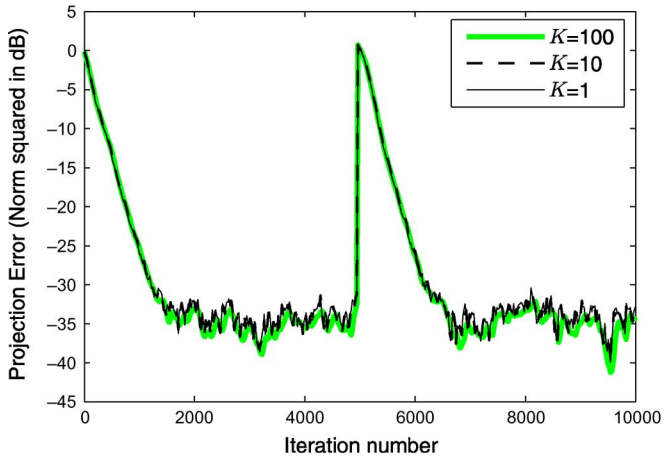


Fig. 1. Performance of (9) for various estimations of the data autocorrelation matrix. Noise subspace tracking.

Fig. 1 depicts the norm squared of the projection error i.e. $\|\mathbf{U}_n\mathbf{U}_n^t - \mathbf{U}\mathbf{U}^t\|_{\mathrm{F}}^2$, of a *single* run. We confirm that the performance is practically insensitive to the window length K, exactly as predicted by our analysis. In fact, the same conclusion holds true even if we estimate the data autocorrelation matrix $\mathbf{R}$ with the popular exponentially windowed sample autocorrelation matrix defined through the recursion $\mathbf{R}_n = \nu\mathbf{R}_{n-1} + (1-\nu)\mathbf{y}_n\mathbf{y}_n^t$, with $0 < \nu < 1$ denoting the exponential forgetting factor.

We continue with comparisons regarding signal and noise subspace estimation and the corresponding convergence towards orthonormality. We use exactly the same signal and noise model as before. For FDPM and FRANS, we apply the simple step size normalization $\mu = \bar{\mu}/\|\mathbf{y}_n\|^2$. In all subsequent graphs, expectation for error power computation, is approximated by averaging 100 independent runs.

A. Signal Subspace Tracking

Let us consider the problem of signal subspace estimation. We compare FDPM against the $O(NL)$ schemes: FAPI, PAST, PROTEUS-2, MALASE, LORAF-3 and also Karasalo's scheme which has computational complexity $O(NL^2)$. The latter, as was aforementioned, is very frequently considered in the literature as point of reference for the $O(NL)$ methods. FRANS is not included in our comparisons since it has exactly the same performance as FDPM. As it will become clear below, the algorithms that compete more directly with FDPM are FAPI and LORAF-3. FAPI's complexity is $3NL + 5L^2 + O(N)$ and LORAF-3 requires $4NL + 3L^2 + O(N)$ operations. However, none of these methods has a noise subspace counterpart, as it is the case for FDPM. We recall that FDPM's complexity is $6NL + O(L)$.

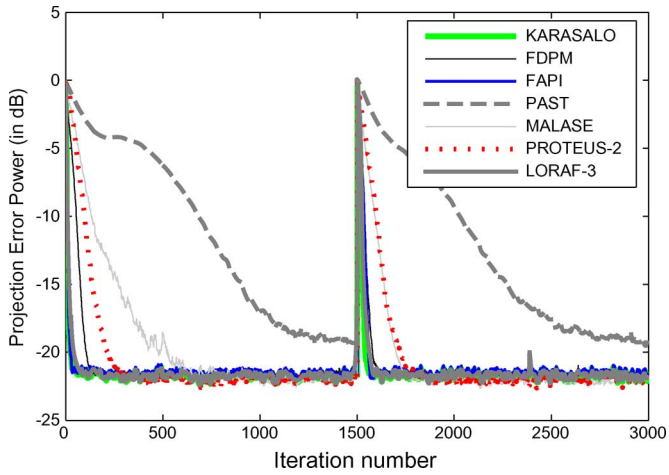


Fig. 2. Performance of the signal subspace tracking schemes, orthonormal change.

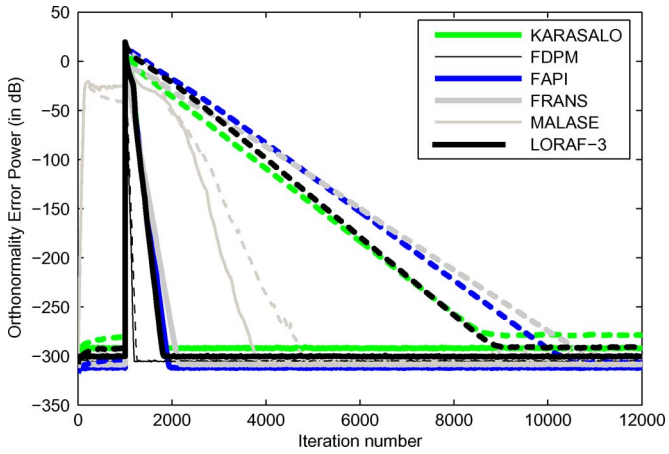


Fig. 3. Deviation from orthonormality of the signal subspace tracking schemes, nonorthonormal change.

Fig. 2 depicts the projection estimation error power $e_p(n)$ defined in (13). At iteration 1500, we apply an abrupt change to the exact basis $\mathbf{U}$ preserving its orthonormal structure. This type of change is imposed to examine the tracking capabilities of all algorithms to follow the new subspace. Step sizes and other parameters are selected so that all algorithms attain *the same steady-state projection error power level*. Specifically, we used for Karasalo $\mu = 0.08413$; for FDPM and FRANS $\mu = 0.0011876$; for FAPI $\beta = 0.95$; for MALASE $\mu = 0.03124$; for LORAF-3 $\alpha = 0.9195$ and for PAST $\mu = 6.98\times10^{-5}$. It is then clear that the algorithm with the fastest rate of convergence can be regarded as the best. FDPM, FAPI, and LORAF-3 exhibit an overall better performance than the other $O(NL)$ schemes, following very closely Karasalo's $O(NL^2)$ method. We observe that FAPI slightly outperforms our FDPM scheme. However, as we are going to see in our next figure, this small advantage disappears in the case of convergence towards orthonormality where FAPI's speed of convergence is significantly inferior to FDPM and, furthermore, FAPI lacks a noise subspace tracking version which is very useful for several important applications. In Fig. 2 we can also see that MALASE and PROTEUS-2 have similar performance requiring twice as many iterations compared to FDPM to attain the same error power level. Finally, PAST has by far the worst performance requiring significantly more time to converge. Regarding PAST we should point out that its performance, relative to the other schemes, improves considerably under lower SNR levels (for more details see [3, pp. 40, 41]).

In Fig. 3 we plot the orthonormality error power, as defined in (34), for FDPM, FAPI, FRANS, MALASE, LORAF-3 and Karasalo's algorithm. PAST and PROTEUS-2 are deliberately excluded from the comparison since the former does not provide orthonormal estimates while the latter has an *extremely* slow convergence. To generate this figure, at iteration $n = 1000$ we replace the estimate $\mathbf{U}_{1000}$ by a nonorthonormal matrix to examine convergence towards orthonormality. Each algorithm is executed with two step sizes μ_1, μ_2 with $\mu_1/\mu_2 = 10$. Since FAPI and LORAF-3 use an exponential forgetting factor α, we know that $1-\alpha$ plays the role of an equivalent step size. Therefore, the two forgetting factors we are using are related through the equation $(1-\alpha_1)/(1-\alpha_2) = 10$. Graphs in solid line correspond to the larger step size. In particular: for Karasalo $\mu_1 = 0.0210$; for FDPM and FRANS $\mu_1 = 2\times10^{-4}$; for FAPI $\beta_1 = 0.98$; for MALASE $\mu_1 = 0.007812$; and for LORAF-3 $\alpha_1 = 0.9977$. Graphs in dashed line correspond to the smaller step size (which is 1/10 of the previous case). We can see that all the tested algorithms practically attain orthonormality within machine accuracy. FDPM has by far the highest convergence speed which changes only marginally with the drastic change in its step size. This very interesting property of our algorithm was accurately predicted by our analysis in Section III-C. From the same figure we also conclude that all other algorithms exhibit a considerable convergence speed reduction when the smaller step size is employed (this is true even for Karasalo's algorithm). In fact, all other algorithms, except MALASE, roughly require 10 times more iterations to converge under the smaller step size, suggesting that the corresponding rate is proportional to the step size of the algorithm.

B. Noise Subspace Tracking

Here we examine the noise subspace tracking characteristics of our algorithm. Apart FDPM, FRANS [27] is also capable of performing noise subspace estimation. We recall that both algorithms constitute different orthonormalization versions of DPM. Our analysis predicted that, in noise subspace estimation, FRANS is nonrobust, that is, when initialized with an orthonormal matrix, it accumulates round-off errors and finally diverges from orthonormality. FDPM on the other hand is robust, retaining orthonormality at all times. Our analysis also predicted that FRANS is unstable, meaning that when initialized with a nonorthonormal matrix it diverges immediately. FDPM on the other hand is stable since it regains orthonormality in (very) few iterations.

To demonstrate the validity of the previous statements, we perform our last set of simulations. We use the same signal and noise model as in the previous cases. We compute the projection error power and the orthonormality error power for DPM (of complexity $O(NL^2)$), FRANS, HFRANS, FDPM, and FOOJA. FOOJA is the most direct competitor of FDPM and comes with a leading computational complexity term of $7NL$ that is higher than FDPM. FOOJA is the first version, after multiple unsuccessful trials that is robust. As we recall from Section I, this

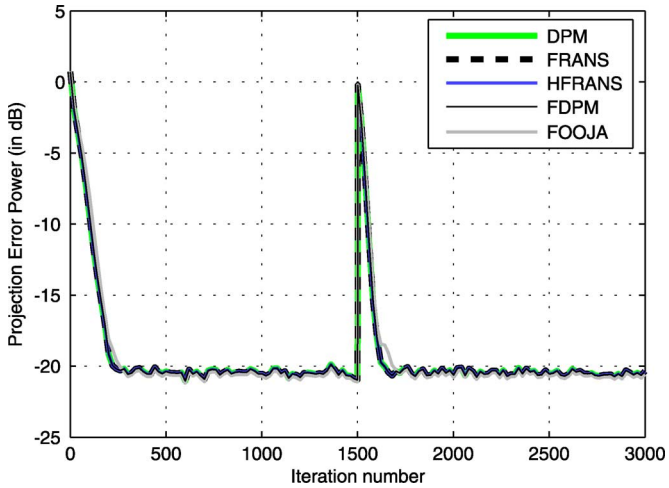


Fig. 4. Performance of the noise subspace tracking schemes, orthonormal change. Deviation from orthonormality of the noise subspace tracking schemes, orthonormal change.

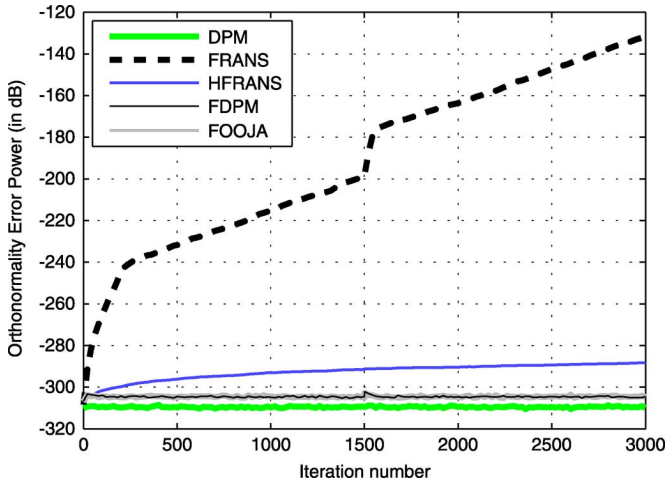


Fig. 5. Deviation from orthonormality of the noise subspace tracking schemes, orthonormal change.

was achieved by employing FDPMS's orthonormalization procedure first presented in [28]. The following step sizes were adopted: DPM, FDPM, FRANS, HFRANS, and FOOJA $\bar{\mu} = 0.2969$. The noise subspace of interest has rank $L = 4$ and a multiple singular value equal to $\sigma^2 = 2$. Again we impose changes to observe the performance of the schemes under comparison. We adopt two different scenarios. In the first, we change the ideal basis $\mathbf{U}$ at time $n = 1500$ retaining its orthonormal structure in order to examine the capability of the algorithms to follow abrupt changes in the subspace. In the second, $\mathbf{U}$ is kept constant, but we replace the estimate $\mathbf{U}_{1500}$ in all algorithms with the same, nonorthonormal matrix. By forcing loss of orthonormality, we examine the capability of each algorithm to recover this characteristic.

Figs. 4 and 5 depict the projection error power and the corresponding orthonormality error power for the first scenario. We can see in Fig. 4 that as long as the orthonormality error is significantly smaller than the estimation error, the DPM-based algorithms exhibit the same indistinguishable performance in the projection error. Focusing on Fig. 5, we observe that DPM, FDPM, and FOOJA satisfy orthonormality within machine accuracy (DPM orthonormalizes at every iteration), FRANS on the other hand is slowly and persistently drifting away from this desirable property. After 3000 iterations, FRANS's orthonormality error has degraded by more than 160 dB. In fact, if we continue the simulation long enough, FRANS eventually becomes unstable, providing meaningless subspace estimates. The reason that the divergence from orthonormality does not manifest itself sooner in the projection error power, is due to the high numerical accuracy environment under which we performed our simulations (Matlab). It is clear that in a less accurate environment this divergence would have been transferred to the projection error power much faster. As far as HFRANS is concerned, it is a more stable version than FRANS. However, by carefully observing Fig. 5 we realize that the algorithm still diverges (slowly) from orthonormality, a fact that is also reported in [30].

In the second scenario, depicted in Figs. 6 and 7, we deliberately destroy orthonormality in the estimate $\mathbf{U}_{1500}$. Similar remarks as in the previous example can be made for the first

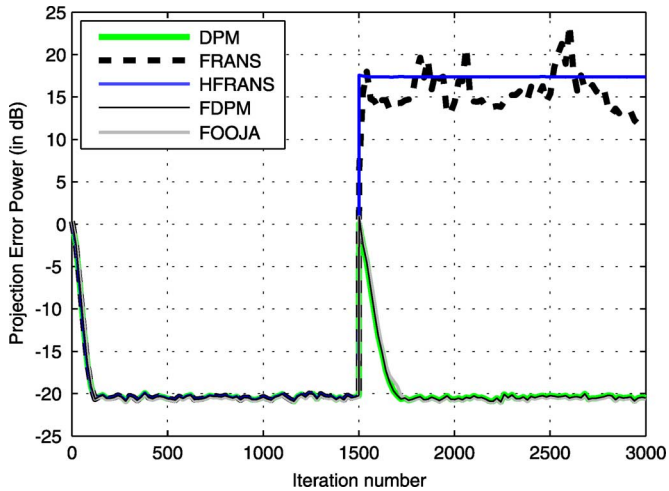


Fig. 6. Performance of the noise subspace tracking schemes, nonorthonormal change.

half of the execution. When however we focus on the part after the abrupt change, we realize that DPM has no problem since it orthonormalizes at every iteration, while FDPM regains orthonormality very quickly. On the other hand, neither FRANS nor HFRANS can recover from the change and the orthonormality error is immediately passed onto its subspace estimates, destroying them completely.

V. Conclusion

In this paper, we have considered the problem of adaptive subspace tracking. Our principal contribution consists in developing a fast and numerically stable orthonormalization technique for the DPM algorithm [12]. The proposed scheme reduces DPM's computational complexity from $O(NL^2)$ to $O(NL)$ meaning that the corresponding algorithm belongs to the low complexity class of subspace tracking techniques. The advantage of our method is its simple structure and its capability to switch from signal subspace to noise subspace estimation with a simple change of sign in its step size. As far as noise subspace tracking is concerned, we should point out that our algorithm is the *first* $O(NL)$ complexity scheme

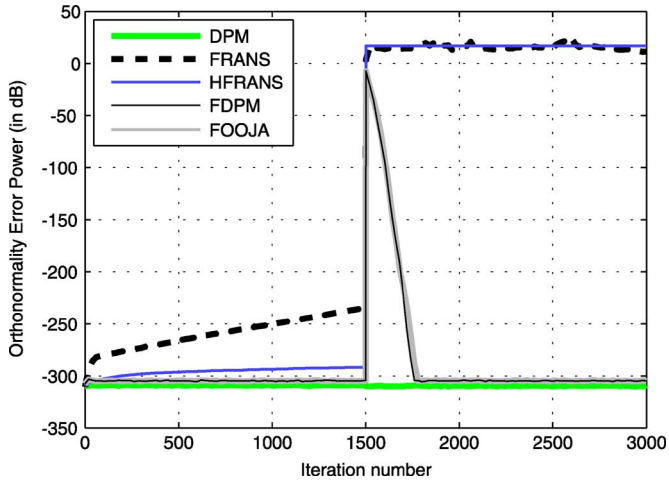


Fig. 7. Deviation from orthonormality of the noise subspace tracking schemes, nonorthonormal change.

which is numerically stable (attaining orthonormality within machine accuracy). Finally, regarding convergence towards orthonormality, our algorithm exhibits the fastest rate among all existing $O(NL)$ subspace tracking algorithms. The convergence characteristics of our method were first predicted by a simple theoretical analysis which was based on available results from Stochastic Approximation theory, with our theoretical findings being in absolute agreement with our simulations.

Appendix I
Proof of Theorem 1 and Lemma 2

In order to prove Theorem 1 and Lemma 2 and also Theorems 2 and 3, we are going to rely on the *Stochastic Approximation Theory*. We summarize the basic results we are going to need contained in [34, pp. 101–108]. Consider the following adaptation:

$$\Theta_n = \Theta_{n-1} - \mu H(\Theta_{n-1}, X_n) \tag{38}$$

where $\{\Theta_n\}$ denotes the adaptive estimates, $\{X_n\}$ is a stationary random data sequence, and $H(\cdot,\cdot)$ is a vector function. If we call $h(\Theta) = \mathbb{E}_X[H(\Theta, X_n)]$, where the expectation is taken *only with respect to the data* X_n and Θ is assumed deterministic, then for the mean trajectory $\bar{\Theta}_n = \mathbb{E}[\Theta_n]$ we have the following adaptation:

$$\bar{\Theta}_n \approx \bar{\Theta}_{n-1} - \mu h(\bar{\Theta}_{n-1}). \tag{39}$$

The *equilibrium points* of this adaptation are the solutions of

$$h(\Theta_\star) = 0 \tag{40}$$

which are the possible states where the adaptation in (38) can converge to (in the mean). Regarding stability of an equilibrium point $\Theta_\star$ we examine it from the *local* point of view by assuming that $\bar{\Theta}_n = \Theta_\star + \mathcal{D}_n$ where $\mathcal{D}_n$ is small. This allows for the linearization of (39) around $\Theta_\star$ as follows:

$$\mathcal{D}_n \approx \mathcal{D}_{n-1} - \mu h_\Theta(\Theta_\star)\mathcal{D}_{n-1} \tag{41}$$

where $h_\Theta(\Theta_\star)$ denotes the Jacobian of $h(\Theta)$ evaluated at the equilibrium $\Theta_\star$. The equilibrium $\Theta_\star$ is (locally) stable if the matrix $\mathbf{I} - \mu h_\Theta(\Theta_\star)$ has all its eigenvalues strictly inside the unit circle. This, for sufficiently small μ happens whenever all the eigenvalues of $h_\Theta(\Theta_\star)$ have *positive real part*. If at least one eigenvalue has negative real part then $\Theta_\star$ is (locally) unstable. From (41), we conclude that, close to the equilibrium, the mean error trajectory behaves as the following matrix power: $\bar{\Theta}_n - \Theta_\star \sim [\mathbf{I} - \mu h_\Theta(\Theta_\star)]^n$. The main contribution to this matrix power comes from the largest, in magnitude, eigenvalue of the matrix $\mathbf{I} - \mu h_\Theta(\Theta_\star)$. Therefore, for sufficiently small μ, we can write $\bar{\Theta}_n - \Theta_\star = O([1 - \mu \min_i \mathrm{Re}(\lambda_i)]^n)$, where λ_i the eigenvalues of $h_\Theta(\Theta_\star)$ and $\mathrm{Re}(\cdot)$ denotes real part. This suggests that the (local) exponential convergence rate towards $\Theta_\star$ is equal to $\mu \min_i\{\mathrm{Re}(\lambda_i)\}$.

Regarding the estimation error $E_n = \Theta_n - \Theta_\star$ *at steady state*, as far as first- and second-order statistics are concerned, it follows the linear model

$$E_n \approx [\mathbf{I} - \mu h_\Theta(\Theta_\star)]\, E_{n-1} - \mu H(\Theta_\star, X_n). \tag{42}$$

If the equilibrium $\Theta_\star$ is stable, then the error covariance matrix $\mathbb{E}[E_n E_n^t]$, at steady state, is of the form $\mu\mathcal{P} + o(\mu)$, where matrix $\mathcal{P}$ satisfies the following Liapunov equation:

$$\mathcal{P} h_\Theta^t(\Theta_\star) + h_\Theta(\Theta_\star)\mathcal{P} = \sum_{n=-\infty}^{\infty} \mathbb{E}\left[H(\Theta_\star, X_0) H^t(\Theta_\star, X_n)\right]. \tag{43}$$

In the special case where $H(\Theta, X)$ is affine in Θ, that is

$$H(\Theta, X_n) = \mathbf{A}(X_n)\Theta + B(X_n) \tag{44}$$

with $\mathbf{A}(\cdot)$ a matrix function and $B(\cdot)$ a vector function, we have a much stronger result [35]. Specifically, the estimation error $E_n = \Theta_n - \Theta_\star$, with

$$\Theta_\star = -\left(\mathbb{E}\left[\mathbf{A}(X_n)\right]\right)^{-1} \mathbb{E}\left[B(X_n)\right] \tag{45}$$

the *unique* equilibrium (provided that $\mathbb{E}[\mathbf{A}(X_n)]$ is invertible), as far as mean convergence and steady-state second-order statistics are concerned, satisfies the recursion

$$E_n \approx \left(\mathbf{I} - \mu \mathbb{E}\left[\mathbf{A}(X_n)\right]\right) E_{n-1} + \mu\left\{B(X_n) + \mathbf{A}(X_n)\Theta_\star\right\}. \tag{46}$$

In this case, stability is no longer local but global and for sufficiently small μ it is assured when the eigenvalues of $\mathbb{E}[\mathbf{A}(X_n)]$ have positive real part. The previous information is sufficient for proving the desired results.

Proof of Theorem 1: Since we are interested in the projection operator and the latter is independent from the orthonormalization procedure, we apply the following orthonormalization in (9):

$$\begin{aligned}\mathbf{U}_n &= \mathrm{orthnorm}\left\{(\mathbf{I}_N \pm \mu\mathbf{R}_n)\mathbf{U}_{n-1}\right\} \\ &= (\mathbf{I}_N \pm \mu\mathbf{R}_n)\mathbf{U}_{n-1} \\ &\quad \times \left(\mathbf{U}_{n-1}^t(\mathbf{I}_N \pm \mu\mathbf{R}_n)^2\mathbf{U}_{n-1}\right)^{-1/2}.\end{aligned} \tag{47}$$

If we form the projection operator, then we have

$$\mathbf{U}_n\mathbf{U}_n^t = (\mathbf{I}_N \pm \mu\mathbf{R}_n)\mathbf{U}_{n-1}\left(\mathbf{U}_{n-1}^t(\mathbf{I}_N \pm \mu\mathbf{R}_n)^2\mathbf{U}_{n-1}\right)^{-1} \times\mathbf{U}_{n-1}^t(\mathbf{I}_N \pm \mu\mathbf{R}_n). \quad (48)$$

Denoting with $\mathbf{P}_n$ the projection operator and recalling that μ is small, we linearize with respect to the step size, retaining terms of order up to μ, this yields

$$\mathbf{P}_n \approx \mathbf{P}_{n-1} \pm \mu\{(\mathbf{I}_N - \mathbf{P}_{n-1})\mathbf{R}_n\mathbf{P}_{n-1} + \mathbf{P}_{n-1}\mathbf{R}_n(\mathbf{I}_N - \mathbf{P}_{n-1})\} \quad (49)$$

where for the last approximate equality we also used the fact that $\mathbf{U}_{n-1}$ is orthonormal. Adaptation (49) is of the form of (38). According to the general theory presented previously, the mean trajectory satisfies the recursion

$$\bar{\mathbf{P}}_n \approx \bar{\mathbf{P}}_{n-1} \pm \mu(\mathbf{I}_N - \bar{\mathbf{P}}_{n-1})\mathbf{R}\bar{\mathbf{P}}_{n-1} \pm \mu\bar{\mathbf{P}}_{n-1}\mathbf{R}(\mathbf{I}_N - \bar{\mathbf{P}}_{n-1}). \quad (50)$$

Let us now consider the (deterministic) orthogonal iteration

$$\begin{aligned}\bar{\mathbf{U}}_n &= \operatorname{orthnorm}\left\{(\mathbf{I}_N \pm \mathbf{R})\bar{\mathbf{U}}_{n-1}\right\} \\ &= (\mathbf{I}_N \pm \mu\mathbf{R})\bar{\mathbf{U}}_{n-1}\left(\bar{\mathbf{U}}_{n-1}^t(\mathbf{I}_N \pm \mu\mathbf{R})^2\bar{\mathbf{U}}_{n-1}\right)^{-1/2}. \quad (51)\end{aligned}$$

If we form the projector $\bar{\mathbf{U}}_n\bar{\mathbf{U}}_n^t$, then we can verify that, to a first-order approximation in μ, this projector satisfies the same approximate recursion as in (50). As we have seen in Section II, $\bar{\mathbf{U}}_n$ tends exponentially fast to the ideal subspace basis with a rate of the form $\mu c + o(\mu)$. We can, thus, conclude that $\bar{\mathbf{U}}_n\bar{\mathbf{U}}_n^t$ (and therefore $\bar{\mathbf{P}}_n$) converges to $\mathbf{P}_\star = \mathbf{U}\mathbf{U}^t$ with the same rate.

Regarding the error power of $\mathbf{P}_n$ at steady state, following (42), we write $\mathbf{P}_n = \mathbf{P}_\star + \mathbf{E}_n$ and linearize (49). Then, at steady state, as far as first- and second-order statistics are concerned, $\mathbf{E}_n$ satisfies the recursion

$$\begin{aligned}\mathbf{E}_n \approx \mathbf{E}_{n-1} &\pm \mu\{(\mathbf{I}_N - \mathbf{P}_\star)\mathbf{R}\mathbf{E}_{n-1} - \mathbf{E}_{n-1}\mathbf{R}\mathbf{P}_\star \\ &\quad + \mathbf{E}_{n-1}\mathbf{R}(\mathbf{I}_N - \mathbf{P}_\star) - \mathbf{P}_\star\mathbf{R}\mathbf{E}_{n-1}\} \\ &\pm \mu\{(\mathbf{I}_N - \mathbf{P}_\star)\mathbf{R}_n\mathbf{P}_\star + \mathbf{P}_\star\mathbf{R}_n(\mathbf{I}_N - \mathbf{P}_\star)\}. \quad (52)\end{aligned}$$

By considering the column-wise version $\operatorname{col}\{\mathbf{E}_n\}$ of the projection error (i.e., where the matrix elements are read in a column by column manner) and by using repeatedly the equality $\operatorname{col}\{\mathbf{A}\mathbf{B}\mathbf{C}^t\} = (\mathbf{A}\otimes\mathbf{C})\operatorname{col}\{\mathbf{B}\}$, where "$\otimes$" denotes Kronecker product, the previous recursion can be written as

$$\operatorname{col}\{\mathbf{E}_n\} = (\mathbf{I}_{N^2} \pm \mu\boldsymbol{\Omega})\operatorname{col}\{\mathbf{E}_{n-1}\} \pm \mu\mathbf{D}\operatorname{col}\{\mathbf{R}_n\} \quad (53)$$

where

$$\begin{aligned}\boldsymbol{\Omega} &= (\mathbf{I}_N - \mathbf{P}_\star)\mathbf{R}\otimes\mathbf{I}_N - \mathbf{I}_N\otimes\mathbf{R}\mathbf{P}_\star \\ &\quad + \mathbf{I}_N\otimes\mathbf{R}(\mathbf{I}_N - \mathbf{P}_\star) - \mathbf{P}_\star\mathbf{R}\otimes\mathbf{I}_N \quad (54)\\ \mathbf{D} &= (\mathbf{I}_N - \mathbf{P}_\star)\otimes\mathbf{P}_\star + \mathbf{P}_\star\otimes(\mathbf{I}_N - \mathbf{P}_\star). \quad (55)\end{aligned}$$

We note that (53) is of the same form as (42), therefore, the covariance matrix of $\operatorname{col}\{\mathbf{E}_n\}$, at steady state, is proportional to μ (to a first-order approximation). Additionally, since for a matrix $\mathbf{A}$, the square of the Frobenius norm $\|\mathbf{A}\|_{\mathrm{F}}^2$ is equal to the sum of squares of all the elements of $\mathbf{A}$, this means that $\|\mathbf{A}\|_{\mathrm{F}}^2 = \operatorname{col}^t\{\mathbf{A}\}\operatorname{col}\{\mathbf{A}\} = \operatorname{tr}\{\operatorname{col}\{\mathbf{A}\}\operatorname{col}^t\{\mathbf{A}\}\}$. We thus conclude that the projection error power is equal to the trace of the covariance matrix of $\operatorname{col}\{\mathbf{E}_n\}$ suggesting that, at steady state, the projection error power is also proportional to μ (to a first-order approximation). This concludes the proof. ■

Proof of Lemma 2: From (50) the mean trajectory of the adaptation depends only on the average $\mathbf{R}$ and is therefore independent from the sequence $\{\alpha_i\}$. Regarding the steady-state projection error power, as we have seen, it is equal to the trace of the covariance matrix of $\operatorname{col}\{\mathbf{E}_n\}$, with the latter satisfying the adaptation in (53). The covariance matrix according to (43), depends on the matrices $\boldsymbol{\Omega}$, $\mathbf{D}$ and the sum $\sum_{n=-\infty}^{\infty}\mathbb{E}[\operatorname{col}\{\mathbf{R}_0\}\operatorname{col}^t\{\mathbf{R}_n\}]$. The two matrices $\boldsymbol{\Omega}$, $\mathbf{D}$ are independent from $\{\alpha_i\}$, we will show that the same holds true for the sum. Indeed if for simplicity we denote $\boldsymbol{\Delta}_n = \mathbf{y}_n\mathbf{y}_n^t$, then we can write

$$\begin{aligned}&\sum_{n=-\infty}^{\infty}\mathbb{E}\left[\operatorname{col}\{\mathbf{R}_0\}\operatorname{col}^t\{\mathbf{R}_n\}\right] \\ &= \sum_{n=-\infty}^{\infty}\mathbb{E}\left[\operatorname{col}\left\{\sum_{i=0}^{K-1}\alpha_i\boldsymbol{\Delta}_{-i}\right\} \times\operatorname{col}^t\left\{\sum_{j=0}^{K-1}\alpha_j\boldsymbol{\Delta}_{n-j}\right\}\right] \quad (56)\\ &= \sum_{i=0}^{K-1}\sum_{j=0}^{K-1}\alpha_i\alpha_j\sum_{n=-\infty}^{\infty}\mathbb{E}\left[\operatorname{col}\{\boldsymbol{\Delta}_{-i}\}\operatorname{col}^t\{\boldsymbol{\Delta}_{n-j}\}\right] \quad (57)\\ &= \sum_{i=0}^{K-1}\sum_{j=0}^{K-1}\alpha_i\alpha_j\sum_{n=-\infty}^{\infty}\mathbb{E}\left[\operatorname{col}\{\boldsymbol{\Delta}_0\}\operatorname{col}^t\{\boldsymbol{\Delta}_{n-j+i}\}\right] \quad (58)\\ &= \sum_{i=0}^{K-1}\sum_{j=0}^{K-1}\alpha_i\alpha_j\underbrace{\sum_{n=-\infty}^{\infty}\mathbb{E}\left[\operatorname{col}\{\boldsymbol{\Delta}_0\}\operatorname{col}^t\{\boldsymbol{\Delta}_n\}\right]}_{\mathbb{Q}} \quad (59)\\ &= \mathbb{Q}\sum_{i=0}^{K-1}\sum_{j=0}^{K-1}\alpha_i\alpha_j = \mathbb{Q}\sum_{i=0}^{K-1}\alpha_i\sum_{j=0}^{K-1}\alpha_j = \mathbb{Q}. \quad (60)\end{aligned}$$

To go from (57) to (58), we used the stationarity of the data; from (58) to (59) we applied the change of indexes $n \leftarrow n - j + i$, which does not affect the limits of the third sum; finally, from (60), it becomes clear that the steady-state error covariance is independent from the sequence $\{\alpha_i\}$. ■

APPENDIX II
PROOF OF THEOREM 2

Here we consider the stability and robustness properties of FRANS by analyzing (30). Assuming small step size we can write (30) as follows:

$$\begin{aligned}\operatorname{col}\{\mathcal{E}_n\} &= \left[\mathbf{H}_n^t\otimes\mathbf{H}_n^t\right]\operatorname{col}\{\mathcal{E}_{n-1}\} + \operatorname{col}\{\mathcal{V}_n\} \quad (61)\\ &\approx \left[\mathbf{I}_{L^2} \mp \mu\left(\left[\mathbf{r}_n\mathbf{r}_n^t\right]\otimes\mathbf{I}_L + \mathbf{I}_L\otimes\left[\mathbf{r}_n\mathbf{r}_n^t\right]\right)\right] \\ &\quad\times\operatorname{col}\{\mathcal{E}_{n-1}\} + \operatorname{col}\{\mathcal{V}_n\} \quad (62)\end{aligned}$$

where we retained terms up to order μ. The last recursion is of the special affine form of (44), therefore, if we call

$$\mathbf{\Phi} = \mathbf{I}_{L^2} \mp \mu \left(\mathbb{E}\left[\mathbf{r}_n \mathbf{r}_n^t\right] \otimes \mathbf{I}_L + \mathbf{I}_L \otimes \mathbb{E}\left[\mathbf{r}_n \mathbf{r}_n^t\right] \right) \tag{63}$$

and assume that the round-off error $\mathcal{V}_n$ is zero mean the equilibrium state is equal to the zero vector. This means that (62), as far as mean trajectory and steady-state second-order statistics are concerned, according to (46) it is equivalent to

$$\mathrm{col}\{\mathcal{E}_n\} \approx \mathbf{\Phi} \mathrm{col}\{\mathcal{E}_{n-1}\} + \mathrm{col}\{\mathcal{V}_n\}. \tag{64}$$

The recursion in (64) is a classical linear system in state-space form with $\mathrm{col}\{\mathcal{E}_n\}$ the "state," $\mathrm{col}\{\mathcal{V}_n\}$ the "input" and $\mathbf{\Phi}$ the "system matrix". We know that such a system is stable if the eigenvalues of $\mathbf{\Phi}$ are all inside the unit circle and unstable if at least one eigenvalue is outside. Since $\mathbb{E}[\mathbf{r}_n \mathbf{r}_n^t]$ is positive definite if we denote with $\omega_1, \ldots, \omega_L$ its eigenvalues, then the eigenvalues of $\mathbf{\Phi}$ are equal to $1 \mp \mu\{\omega_i + \omega_j\}$ for $i, j = 1, \ldots, L$. Now notice that when we use the "$-$" sign (corresponding to signal subspace), for sufficiently small μ the eigenvalues of $\mathbf{\Phi}$ are all in the interval (0,1) meaning that the system is stable. We have exactly the opposite picture in the noise subspace when we employ the "$+$" sign. Since the eigenvalues of $\mathbf{\Phi}$ are larger than unity we obtain an unstable system.

Assuming that the round-off error $\mathcal{V}_n$ is zero mean and taking expectation in (64), we obtain the recursion of the mean trajectory which is of the form $\mathbb{E}[\mathrm{col}\{\mathcal{E}_n\}] \approx \mathbf{\Phi}\mathbb{E}[\mathrm{col}\{\mathcal{E}_{n-1}\}] \approx \mathbf{\Phi}^n \mathbb{E}[\mathrm{col}\{\mathcal{E}_0\}]$. Therefore, if we start with a nonzero error (initialize with a nonorthonormal matrix) then the average error tends to zero in the signal subspace case and it diverges in the noise subspace case. Let us now select $\mathcal{E}_0 = 0$, in other words initialize the algorithm with an orthonormal matrix. If we assume that the round-off error $\mathcal{V}_n$ and the state $\mathrm{col}\{\mathcal{E}_{n-1}\}$ are uncorrelated and call $\mathbf{F}_n$ and $\mathbf{Q}$ the covariance matrices of $\mathrm{col}\{\mathcal{E}_n\}$ and $\mathrm{col}\{\mathcal{V}_n\}$, respectively, then from (64) we obtain the recursion $\mathbf{F}_n \approx \mathbf{\Phi}\mathbf{F}_{n-1}\mathbf{\Phi} + \mathbf{Q}$. Again this recursion converges to a bounded matrix when $\mathbf{\Phi}$ is stable and diverges when $\mathbf{\Phi}$ is unstable. This means that the orthonormality error covariance $\mathbf{F}_n$ and, therefore, the orthonormality error power, in signal subspace estimation remains bounded, while in noise subspace it increases without limit.

Finally for the convergence rate, since the mean trajectory behaves as $\mathbf{\Phi}^n$ with the main contribution coming from the largest in absolute value eigenvalue of $\mathbf{\Phi}$, exactly as described in the beginning of Appendix I, the convergence rate is equal to $-\log(1 - \mu c) \approx \mu c$, where $c = 2\min_i \omega_i > 0$ in the signal subspace case and $c = 2\min_i(-\omega_i) < 0$ in the noise subspace case. ∎

Appendix III
Proof of Theorem 3

To analyze FDPM we refer to (33). This adaptation is nonlinear in $\mathcal{E}_{n-1}$, therefore we are going to demonstrate *local stability* assuming that $\mathcal{E}_{n-1}$ is small. Recalling that μ is also small, we can write

$$\mathcal{E}_n = \mathbf{H}_n^t \mathcal{E}_{n-1} \mathbf{H}_n - \mathrm{diag}\left\{\mathbf{H}_n^t \mathcal{E}_{n-1} \mathbf{H}_n\right\} + o(\mathcal{E}_{n-1}) + \mu O(\mathcal{E}_{n-1}) + \mathcal{V}_n \tag{65}$$

where the fourth term in the right-hand side (RHS) comes from the fact that $\delta_n = O(\mu)$. Since we assumed that $\mu \ll 1$ and also that the error is small, the third and forth term are negligible compared to the first two leading to the following simplification:

$$\begin{aligned}\mathcal{E}_n &\approx \mathbf{H}_n^t \mathcal{E}_{n-1} \mathbf{H}_n - \mathrm{diag}\left\{\mathbf{H}_n^t \mathcal{E}_{n-1} \mathbf{H}_n\right\} + \mathcal{V}_n \\ &= \mathbf{H}_n^t \mathcal{E}_{n-1} \mathbf{H}_n - \sum_{k=1}^{L} \mathbf{e}_k \mathbf{e}_k^t \left\{\mathbf{e}_k^t \mathbf{H}_n^t \mathcal{E}_{n-1} \mathbf{H}_n \mathbf{e}_k\right\} + \mathcal{V}_n \end{aligned} \tag{66}$$

where $\mathbf{e}_k$, similarly to $\mathbf{e}_1$, is a vector of length L with all elements equal to zero except the kth which is equal to 1. Noticing that we can write $\mathbf{e}_k^t \mathbf{H}_n^t \mathcal{E}_{n-1} \mathbf{H}_n \mathbf{e}_k = (\mathbf{e}_k^t \otimes \mathbf{e}_k^t)(\mathbf{H}_n^t \otimes \mathbf{H}_n^t)\mathrm{col}\{\mathcal{E}_{n-1}\}$, if we use this in (66), after some manipulations, we can write the recursion in vector form as

$$\begin{aligned}\mathrm{col}\{\mathcal{E}_n\} \approx &\left(\mathbf{I}_{L^2} - \sum_{k=1}^{L} (\mathbf{e}_k \otimes \mathbf{e}_k)\left(\mathbf{e}_k^t \otimes \mathbf{e}_k^t\right)\right) \\ &\times \left(\mathbf{H}_n^t \otimes \mathbf{H}_n^t\right) \mathrm{col}\{\mathcal{E}_{n-1}\} + \mathrm{col}\{\mathcal{V}_n\}. \end{aligned} \tag{67}$$

Assuming $\mathcal{V}_n$ zero mean and uncorrelated from $\mathcal{E}_{n-1}$, as far as mean convergence and steady-state second-order statistics of $\mathrm{col}\{\mathcal{E}_n\}$ are concerned, the previous relation is equivalent to

$$\begin{aligned}\mathrm{col}\{\mathcal{E}_n\} \approx &\left(\mathbf{I}_{L^2} - \sum_{k=1}^{L} (\mathbf{e}_k \otimes \mathbf{e}_k)\left(\mathbf{e}_k^t \otimes \mathbf{e}_k^t\right)\right) \mathbf{\Psi} \mathrm{col}\{\mathcal{E}_{n-1}\} \\ &+ \mathrm{col}\{\mathcal{V}_n\} \end{aligned} \tag{68}$$

where $\mathbf{\Psi} = \mathbb{E}[\mathbf{H}_n^t \otimes \mathbf{H}_n^t]$. To show stability and robustness, as in the proof of Appendix II, we only need to show that the linear system in (68) is stable, namely that all eigenvalues of the system matrix $\tilde{\mathbf{\Psi}} = \{\mathbf{I}_{L^2} - \sum_{k=1}^{L} (\mathbf{e}_k \otimes \mathbf{e}_k)(\mathbf{e}_k^t \otimes \mathbf{e}_k^t)\}\mathbf{\Psi}$ lie inside the unit circle. For this, we are going to use the following lemma whose proof can be found in [32, pp. 410, 411].

Lemma 3: If $\mathbf{Q}$ is a real symmetric matrix of size K with eigenvalues $\lambda_1, \ldots, \lambda_K$, then the eigenvalues are real; furthermore, if $\mathbf{x}$ denotes a real vector of length K we have

$$\min_i\{\lambda_i\} = \inf_{\mathbf{x} \neq 0} \frac{\mathbf{x}^t \mathbf{Q} \mathbf{x}}{\|\mathbf{x}\|^2} \leq \sup_{\mathbf{x} \neq 0} \frac{\mathbf{x}^t \mathbf{Q} \mathbf{x}}{\|\mathbf{x}\|^2} = \max_i\{\lambda_i\}. \tag{69}$$

Let us first apply this lemma to show that $\mathbf{\Psi}$ has all its eigenvalues with magnitude bounded by 1. From (29) we have that $\mathbf{H}_n$ is symmetric with eigenvalues equal to ± 1. Since the eigenvalues of $\mathbf{A} \otimes \mathbf{B}$ are all the possible products $\lambda_i(\mathbf{A})\nu_j(\mathbf{B})$, where $\lambda_i(\mathbf{A})$ and $\nu_j(\mathbf{B})$ are the eigenvalues of $\mathbf{A}$ and $\mathbf{B}$, respectively; we conclude that $\mathbf{H}_n^t \otimes \mathbf{H}_n^t$ has eigenvalues equal to ± 1 as well. Using Lemma 3 this suggests that for *any nonzero constant* vector $\mathbf{x}$, we have $-1 \leq \mathbf{x}^t \mathbf{H}_n^t \otimes \mathbf{H}_n^t \mathbf{x}/\|\mathbf{x}\|^2 \leq 1$. Taking expectation yields $-1 \leq \mathbf{x}^t \mathbf{\Psi} \mathbf{x}/\|\mathbf{x}\|^2 \leq 1$, $\forall \mathbf{x} \neq 0$. Taking infimum, then supremum over $\mathbf{x}$ and using Lemma 3, we prove that $-1 \leq \lambda_{\min}(\mathbf{\Psi}) \leq \lambda_{\max}(\mathbf{\Psi}) \leq 1$, where $\lambda_{\min}(\mathbf{\Psi})$, $\lambda_{\max}(\mathbf{\Psi})$ denote the smallest and largest eigenvalue of $\mathbf{\Psi}$.

We will now show that the eigenvalues of the desired system matrix $\tilde{\mathbf{\Psi}}$ are bounded from below and above by $\lambda_{\min}(\mathbf{\Psi})$, $\lambda_{\max}(\mathbf{\Psi})$ suggesting that $\tilde{\mathbf{\Psi}}$ is more stable that $\mathbf{\Psi}$. Note that $\mathbf{I}_{L^2} - \sum_{k=1}^{L} (\mathbf{e}_k \otimes \mathbf{e}_k)(\mathbf{e}_k^t \otimes \mathbf{e}_k^t)$ is a diagonal matrix with all its diagonal elements equal to 1 except L diagonal elements that are equal to 0. This matrix is symmetric and equal to

its square, it is, therefore, an orthogonal projection operator. This means that it can be decomposed as $\mathbf{V}\mathbf{V}^t$ where $\mathbf{V}$ is an orthonormal matrix of dimensions $L^2 \times (L^2 - L)$. Let us identify the eigenvalues of $\tilde{\mathbf{\Psi}} = \mathbf{V}\mathbf{V}^t\mathbf{\Psi}$. Because $\mathbf{V}\mathbf{V}^t$ is rank deficient, matrix $\tilde{\mathbf{\Psi}}$ has rank $L^2 - L$, meaning that L of its eigenvalues are equal to 0. We can also easily show that the remaining $L^2 - L$ eigenvalues are the same as the eigenvalues of $\mathbf{V}^t\mathbf{\Psi}\mathbf{V}$. Indeed, if $(\nu, \mathbf{z}_\nu)$ is an eigenvalue/eigenvector pair for $\mathbf{V}^T\mathbf{\Psi}\mathbf{V}$ then we can immediately verify that $(\nu, \mathbf{V}\mathbf{z}_\nu)$ is a similar pair for $\tilde{\mathbf{\Psi}}$. It is clearly the latter eigenvalues that we would like to bound. Notice now that for any vector $\mathbf{z}$ of length $L^2 - L$, the vector $\mathbf{x} = \mathbf{V}\mathbf{z}$ is of length L^2. Furthermore, due to the orthonormality of $\mathbf{V}$, we have $||\mathbf{x}|| = ||\mathbf{z}||$. With this in mind and recalling Lemma 3 we can write

$$\lambda_{\min}(\mathbf{\Psi}) = \inf_{\mathbf{x}\neq 0} \frac{\mathbf{x}^t\mathbf{\Psi}\mathbf{x}}{||\mathbf{x}||^2} \leq \inf_{\mathbf{x}=\mathbf{V}\mathbf{z},\mathbf{z}\neq 0} \frac{\mathbf{x}^t\mathbf{\Psi}\mathbf{x}}{||\mathbf{x}||^2} = \inf_{\mathbf{z}\neq 0} \frac{\mathbf{z}^t\mathbf{V}^t\mathbf{\Psi}\mathbf{V}\mathbf{z}}{||\mathbf{z}||^2} = \lambda_{\min}(\tilde{\mathbf{\Psi}}). \tag{70}$$

The previous inequality is true since the second minimization is performed over a smaller space. In exactly a similar way, we can show that $\lambda_{\max}(\tilde{\mathbf{\Psi}}) \leq \lambda_{\max}(\mathbf{\Psi})$. We thus conclude that the system matrix in (68) has all its eigenvalues bounded by 1.

Up to this point we have assured marginal stability since we have not excluded the case of an eigenvalue being $\pm$ 1. Thus, let us assume that we have a vector $\mathbf{z}_o$ that satisfies $\mathbf{z}_o^t\mathbf{V}^t\mathbf{\Psi}\mathbf{V}\mathbf{z}_o/||\mathbf{z}_o||^2 = 1$. We recall that $\mathbf{\Psi} = \mathbb{E}[\mathbf{H}_n^t \otimes \mathbf{H}_n^t]$ and, as we have argued above, $\mathbf{z}_o^t\mathbf{V}^t(\mathbf{H}_n^t \otimes \mathbf{H}_n^t)\mathbf{V}\mathbf{z}_o/||\mathbf{z}_o||^2 \leq 1$. We, therefore, conclude that the quantity $\mathbf{z}_o^t\mathbf{V}^t(\mathbf{H}_n^t \otimes \mathbf{H}_n^t)\mathbf{V}\mathbf{z}_o/||\mathbf{z}_o||^2$ is a random variable less than or equal to 1 for all data realizations with a mean equal to 1. But this is only possible when the random variable is always equal to 1 *for all realizations (or more accurately, with probability one)*. The latter outcome is also equivalent to saying that the deterministic vector $\mathbf{V}\mathbf{z}_o$ is an eigenvector to all random matrices $\mathbf{H}_n^t \otimes \mathbf{H}_n^t$, corresponding to a unit eigenvalue. Such a possibility however is easily avoidable if for example the data have some completely random (white noise) component. Under such an assumption the eigenvalues are strictly inside the unit circle assuring stability of our scheme. As we have seen, this stability property is, to a first-order approximation, independent from the step size μ. ■

ACKNOWLEDGMENT

The authors would like to express their gratitude to T. Chonavel, Professor at ENST-Bretagne, Brest, France, for providing the codes of most signal subspace tracking schemes.

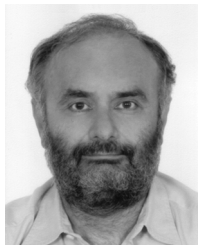
REFERENCES

[1] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, pp. 387–392, Apr. 1985.
[2] P. Comon and G. H. Golub, "Tracking a few extreme singular values and vectors in signal processing," *Proc. IEEE*, vol. 78, no. 8, Aug. 1990.
[3] X. G. Doukopoulos, "Power techniques for blind channel estimation in wireless communication systems," Ph.D. dissertation, IRISA-INRIA, Univ. Rennes 1, Rennes, France, 2004.
[4] N. L. Owsley, "Adaptive data orthogonalization," in *ICASSP'1978*, 1978, pp. 109–112.
[5] G. Xu and T. Kailath, "Fast subspace decomposition," *IEEE Trans. Signal Process.*, vol. 42, no. 3, pp. 539–551, Mar. 1994.
[6] G. Xu and T. Kailath, "Fast estimation of principal eigenspace using Lanczos algorithm," *SIAM J. Matrix Analysis Appl.*, vol. 15, no. 3, pp. 974–994, Jul. 1994.
[7] B. N. Parlett, *The Symmetric Eigenvalue Problem*. Englewood Cliffs, NJ: Prentice-Hall, 1980.
[8] G. W. Stewart, "An updating algorithm for subspace tracking," *IEEE Trans. Signal Process.*, vol. 40, no. 6, pp. 1535–1541, Jun. 1992.
[9] S. Hosur, A. H. Tewfik, and D. Boley, "ULV and generalized ULV subspace tracking adaptive algorithms," *IEEE Trans. Signal Process.*, vol. 46, no. 5, pp. 1282–1297, May 1998.
[10] F. Hansen and P. Y. Yalamov, "Computing symmetric rank-revealing decompositions via triangular factorization," *SIAM J. Matrix Analysis Appl.*, vol. 23, no. 2, pp. 443–458, 2001.
[11] E. S. Baker and R. D. DeGroat, "A correlation-based subspace tracking algorithm," *IEEE Trans. Signal Process.*, vol. 46, no. 11, pp. 3112–3116, Nov. 1998.
[12] J. F. Yang and M. Kaveh, "Adaptive eigensubspace algorithms for direction or frequency estimation and tracking," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, pp. 241–251, Feb. 1988.
[13] I. Karasalo, "Estimating the covariance matrix by signal subspace averaging," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, pp. 8–12, Feb. 1986.
[14] B. Yang, "Projection approximation subspace tracking," *IEEE Trans. Signal Process.*, vol. 43, pp. 95–107, Jan. 1995.
[15] T. Chonavel, B. Champagne, and C. Riou, "Fast adaptive eigenvalue decomposition: A maximum likelihood approach," *Signal Process.*, vol. 83, pp. 307–324, 2003.
[16] B. Champagne and Q. G. Liu, "Plane rotation-based EVD updating schemes for efficient subspace tracking," *IEEE Trans. Signal Process.*, vol. 46, pp. 1886–1900, Jul. 1998.
[17] P. Strobach, "Low-rank adaptive filters," *IEEE Trans. Signal Process.*, vol. 44, pp. 2932–2947, Dec. 1996.
[18] R. Badeau, B. David, and G. Richard, "Fast approximated power iteration subspace tracking," *IEEE Trans. Signal Process.*, vol. 53, pp. 2931–2941, Aug. 2005.
[19] X. Wang and H. V. Poor, "Blind equalization and multiuser detection in dispersive CDMA channels," *IEEE Trans. Commun.*, vol. 46, pp. 91–103, Jan. 1998.
[20] B. Muquet, M. de Courville, and P. Duhamel, "Subspace-based blind and semi-blind channel estimation for OFDM systems," *IEEE Trans. Signal Process.*, vol. 50, pp. 1699–1712, Jul. 2002.
[21] X. G. Doukopoulos and G. V. Moustakides, "Blind adaptive channel estimation in OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 5, no. 7, pp. 1716–1725, Jul. 2006.
[22] P. Strobach, "Square-root QR inverse iteration for tracking the minor subspace," *IEEE Trans. Signal Process.*, vol. 48, pp. 2994–2999, Nov. 2000.
[23] R. Badeau, B. David, and G. Richard, "YAST algorithm for minor subspace tracking," in *Proc. IEEE ICASSP'06*, Toulouse, France, May 2006, vol. III, pp. 552–555.
[24] E. Oja, "Principal components, minor components, and linear neural networks," *Neural Netw.*, vol. 5, pp. 927–935, 1992.
[25] K. Abed-Meraim, S. Attallah, A. Chkeif, and Y. Hua, "Orthogonal Oja algorithm," *IEEE Signal Process. Lett.*, vol. 7, pp. 116–119, May 2000.
[26] S. Attallah and K. Abed-Meraim, "Fast algorithms for subspace tracking," *IEEE Signal Process. Lett.*, vol. 8, pp. 203–206, July 2001.
[27] S. Attallah and K. Abed-Meraim, "Low cost adaptive algorithm for noise subspace estimation," *Electron. Lett.*, vol. 38, pp. 609–611, June 2002.
[28] X. G. Doukopoulos and G. V. Moustakides, "The fast data projection method for stable subspace tracking," in *13th Europ. Signal Process. Conf., EUSIPCO'2005*, Antalya, Turkey, Sep. 2005.
[29] S. Attalah, "The generalized Rayleigh's quotient adaptive noise subspace algorithm: A householder transformation-based implementation," *IEEE Trans. Circuits Syst. II*, vol. 53, pp. 3–7, Jan. 2006.
[30] Y. Lu, S. Attalah, G. Mathew, and K. Abed-Meraim, "Analysis of orthogonality error propagation for FRANS and HFRANS algorithms," *IEEE Trans. Signal Process.*, (under review), submitted for publication.
[31] S. Bartelmaos and K. Abed-Meraim, "Principal and minor subspace tracking: Algorithms and stability analysis," in *Proc. IEEE ICASSP'06*, Toulouse, France, May 2006, vol. III, pp. 560–563.
[32] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd ed. Baltimore , MD: The John Hopkins Univ. Press, 1993.

[33] S. Attallah, "Revisiting adaptive signal subspace estimation based on Rayleigh's quotient," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process., (ICASSP)*, 2001.
[34] A. Benveniste, M. Métivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximation*. Berlin, Germany: Springer, 1990.
[35] G. V. Moustakides, "Exponential convergence of products of random matrices, application to the study of adaptive algorithms," *Int. J. Adapt. Control Signal Process.*, vol. 2, no. 12, pp. 579–597, Dec. 1998.

Xenofon G. Doukopoulos (S'00–M'05) was born in Grevena, Greece, in 1978. He received the diploma in computer engineering from the Computer Engineering and Informatics Department (CEID), University of Patras, Greece, in 2000; the M.Sc. degree in signal and image processing systems, from the same university in 2003, and the Ph.D. degree in signal processing and telecommunications from the Institut National de Recherche en Informatique et en Automatique (INRIA), Rennes, France, in 2004.

From 2004 to 2006 he was a Postdoctoral researcher with France Telecom R&D, Cesson-SévignÉ, France, and from 2006 to 2007, with La Commissariat à l'Energie Atomique (CEA), Grenoble, France, as a research scientist. In 2007, he joined the R&D division of DibCom, Paris, France. His research interests include channel estimation and multiuser detection for wireless telecommunication systems, intercell interference cancellation techniques, signal processing for telecommunication applications, and adaptive subspace tracking techniques.



George V. Moustakides (M'82–SM'97) was born in Drama, Greece, in 1955. He received the diploma in electrical and mechanical engineering from the National Technical University of Athens, Greece, in 1979; the M.Sc. degree in systems engineering from the Moore School of Electrical Engineering, University of Pennsylvania, Philadelphia, in 1980, and the Ph.D. degree in electrical engineering and computer science from Princeton University, Princeton, NJ, in 1983.

From 1983 to 1986, he was with INRIA, France, and from 1987 to 1990, he held a research position with the Computer Technology Institute of Patras, Greece. In 1991, he joined the Computer Engineering and Informatics Department, University of Patras, as Associate Professor and in 1996, he became Professor with the same department. In 2001, he joined the Department of Computer and Communication Engineering, University of Thessaly, Volos, Greece, until 2006. Since 2007, he has been a Professor with the Department of Electrical and Computer Engineering, University of Patras. His current interests include sequential detection of changes and statistical signal processing.