Large Sample Properties of Separable Nonlinear Least Squares Estimators

Kaushik Mahata and Torsten Söderström, *Fellow, IEEE*

***Abstract*—In this paper, the large sample properties of the separable nonlinear least squares algorithm are investigated. Unlike the previous results in the literature, the data are assumed to be complex valued, and the whiteness assumption on the measurement noise sequence has been relaxed. Convergence properties of the parameter estimates are established. Asymptotic accuracy analysis has been carried out, in which the assumptions used are relatively weaker than the assumptions in the previous related works. It is shown under quite general conditions that the parameter estimates are asymptotically circular. Conditions for asymptotic complex normality are also established. Next, a bound on the deviation of the asymptotic covariance matrix from the Cramér–Rao bound (CRB) is derived. Finally, a sufficient condition for the nonlinear least squares estimate to achieve the Cramér–Rao lower bound is established. The results presented in this paper are general and can be applied to any specific application where separable nonlinear least squares is employed.**

***Index Terms*—Asymptotic analysis, consistency, Cramér–Rao bound, nonlinear least squares, variable projection problem.**

I. Introduction

A. Background

Consider a complex scalar valued sequence $\{x_t\}_{t>0}$, which is governed by the model

$$x_t = f_t(\boldsymbol{\theta}_0, \mathbf{c}_0) = \boldsymbol{\psi}_t^*(\boldsymbol{\theta}_0)\mathbf{c}_0, \quad t > 0 \tag{1}$$

where $\boldsymbol{\psi}_t(\boldsymbol{\theta})$ is a $p \times 1$ vector-valued nonlinear function of the unknown complex valued parameter vector $\boldsymbol{\theta}$ of dimension $n \times 1$. Note that we use $\boldsymbol{\psi}_t^*(\boldsymbol{\theta})$ to denote the conjugate transpose of $\boldsymbol{\psi}_t(\boldsymbol{\theta})$. As a consequence of (1), each of the members of the complex scalar valued sequence of functions $\{f_t(\boldsymbol{\theta}, \mathbf{c})\}_{t>0}$ is linear in $\mathbf{c}$ and nonlinear in $\boldsymbol{\theta}$. The parameter $\mathbf{c}$ is assumed to be a complex-valued vector of dimension $p \times 1$ with bounded norm. Here, we are concerned with the problem of estimating the parameters $\boldsymbol{\theta}_0$ and $\mathbf{c}_0$ from N samples of the noise corrupted measurements of the sequence $\{x_t\}_{t>0}$. The observed sequence is given by

$$z_t = x_t + \tilde{x}_t \tag{2}$$

Manuscript received May 15, 2003; revised August 20, 2003. This work was supported in part by Swedish Research Council for Engineering Sciences under Contract 2000-587. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jian Li.

K. Mahata is with the Centre for Complex Dynamic Systems and Control, University of Newcastle, Callaghan, NSW-2308, Australia (e-mail: kush@ee.newcastle.edu.au).

T. Söderström is with the Systems and Control Group, Department of Information Technology, Uppsala University, SE-751 05 Uppsala, Sweden (e-mail: Torsten.Soderstrom@it.uu.se).

Digital Object Identifier 10.1109/TSP.2004.827227

where $\{\tilde{x}_t\}_{t>0}$ is the additive measurement noise sequence. Let us introduce the notations

$$\mathbf{x}_N = [x_1 \quad \dots \quad x_N]^\top \tag{3}$$

$$\mathbf{f}_N(\boldsymbol{\theta}, \mathbf{c}) = [f_1(\boldsymbol{\theta}, \mathbf{c}) \quad \dots \quad f_N(\boldsymbol{\theta}, \mathbf{c})]^\top \tag{4}$$

$$\boldsymbol{\Psi}_N(\boldsymbol{\theta}) = [\boldsymbol{\psi}_1(\boldsymbol{\theta}) \quad \dots \quad \boldsymbol{\psi}_N(\boldsymbol{\theta})]^* \tag{5}$$

so that the $N \times 1$ complex vector-valued function $\mathbf{f}_N(\boldsymbol{\theta}, \mathbf{c})$ can be written using (1) as

$$\mathbf{x}_N = \mathbf{f}_N(\boldsymbol{\theta}, \mathbf{c}) = \boldsymbol{\Psi}_N(\boldsymbol{\theta})\mathbf{c}. \tag{6}$$

We will maintain

$$\tilde{\mathbf{x}}_N = [\tilde{x}_1 \quad \dots \quad \tilde{x}_N]^\top \tag{7}$$

so that

$$\mathbf{z}_N := \mathbf{x}_N + \tilde{\mathbf{x}}_N = \boldsymbol{\Psi}_N(\boldsymbol{\theta})\mathbf{c} + \tilde{\mathbf{x}}_N. \tag{8}$$

One way to estimate the true parameter $\boldsymbol{\theta}_0$ from the observations is to solve (8) in a least squares sense, i.e., to seek for the global minimum point of the loss function

$$\hat{L}_N(\boldsymbol{\theta}, \mathbf{c}) := N^{-1} \{\mathbf{z}_N - \mathbf{f}_N(\boldsymbol{\theta}, \mathbf{c})\}^* \{\mathbf{z}_N - \mathbf{f}_N(\boldsymbol{\theta}, \mathbf{c})\} \tag{9}$$

and estimate $\boldsymbol{\theta}_0$ as

$$\hat{\boldsymbol{\theta}}_N = \arg\min_{\boldsymbol{\theta}} \left[\min_{\mathbf{c}} \hat{L}_N(\boldsymbol{\theta}, \mathbf{c})\right]. \tag{10}$$

However, the optimization problem in (10) is *separable* in the sense that it can be solved for $\boldsymbol{\theta}$ and $\mathbf{c}$ separately. From the theory of linear least squares [1], [2], it follows that for a given $\boldsymbol{\theta}$, the loss function (9) can be minimized analytically with respect to $\mathbf{c}$, and the minimum is achieved at

$$\begin{aligned} \hat{\mathbf{c}}(\boldsymbol{\theta}) &:= \arg\min_{\mathbf{c}} \hat{L}_N(\boldsymbol{\theta}, \mathbf{c}) \\ &= [\boldsymbol{\Psi}_N^*(\boldsymbol{\theta})\boldsymbol{\Psi}_N(\boldsymbol{\theta})]^{-1} \boldsymbol{\Psi}_N^*(\boldsymbol{\theta})\mathbf{z}_N \\ &= \boldsymbol{\Psi}_N^\dagger(\boldsymbol{\theta})\mathbf{z}_N \end{aligned} \tag{11}$$

where $\boldsymbol{\Psi}_N^\dagger(\boldsymbol{\theta})$ is the *pseudo-inverse* of $\boldsymbol{\Psi}_N(\boldsymbol{\theta})$. Note that we assume $\boldsymbol{\Psi}_N(\boldsymbol{\theta})$ to have a full column rank p, which in general is a mild assumption. We will address this point later. Substituting (11) in (9), we have the concentrated loss function

$$\hat{\ell}_N(\boldsymbol{\theta}) := \hat{L}_N\{\boldsymbol{\theta}, \hat{\mathbf{c}}(\boldsymbol{\theta})\} = N^{-1}\mathbf{z}_N^* \boldsymbol{\Pi}_{\boldsymbol{\Psi}_N}^\perp(\boldsymbol{\theta})\mathbf{z}_N \tag{12}$$

where $\boldsymbol{\Pi}_{\boldsymbol{\Psi}_N}^\perp(\boldsymbol{\theta})$ is the orthogonal projection operator onto the null space of $\boldsymbol{\Psi}_N^*(\boldsymbol{\theta})$ given by

$$\boldsymbol{\Pi}_{\boldsymbol{\Psi}_N}^\perp(\boldsymbol{\theta}) = \mathbf{I}_N - \boldsymbol{\Psi}_N(\boldsymbol{\theta})\boldsymbol{\Psi}_N^\dagger(\boldsymbol{\theta}). \tag{13}$$

Using (12), the optimization problem in (10) reduces to

$$\hat{\boldsymbol{\theta}}_N = \arg\min_{\boldsymbol{\theta}} \hat{\ell}_N(\boldsymbol{\theta}). \tag{14}$$

The optimization problem in (14) is often referred to as a variable projection problem; see [3]. In general, such optimization problems must be solved numerically. Finally, using $\hat{\boldsymbol{\theta}}_N$, one obtains the estimate of $\mathbf{c}_0$ as

$$\hat{\mathbf{c}}_N = \hat{\mathbf{c}}(\hat{\boldsymbol{\theta}}_N). \tag{15}$$

B. Contributions and Motivation

As mentioned in [4] and references therein, the nonlinear least squares (NLLS) is ubiquitous to parameter estimation problems in signal processing. In this paper, our object is to study the statistical properties of the variable projection problem. In Section II, we carry out the consistency analysis assuming $\tilde{x}_t$ to be correlated. The additional constraints imposed on $\tilde{x}_t$ are much weaker than the whiteness assumption made in previous related works [5], [6].

In Section III, the accuracy analysis is carried out in a very general framework (the noise is correlated, parameters are complex valued, and the data are complex valued). Under very mild conditions, the circularity and asymptotic normality of the estimates are established. The new expressions presented in this section generalize many previous results in context of specific applications; see, for example, [7]–[10].

In Section IV, we will give a bound on the loss of statistical efficiency. We also establish a sufficient condition on the model for which the NLLS achieves the CRB and thereby give an alternative interpretation of the result proved in [11].

Finally, in Section V, we extend the previously derived results for real-valued data. All the results derived in this paper can be readily extended to the case where the model structure is given by

$$z_t = f_t(\mathbf{z}_{t-1}, \boldsymbol{\theta}, \mathbf{c}) + \tilde{x}_t. \tag{16}$$

II. Parameter Convergence

In this section, our aim is to establish sufficient conditions to ensure the strong convergence of the parameter estimates. In the rest of the paper, $\tilde{x}_t$ will be assumed to satisfy the following.

Assumption 1: $\{\tilde{x}_k\}_{k>0}$ is a stationary, zero mean, circular process with bounded fourth-order moment [12], [13] so that

$$\mathbf{E}\tilde{\mathbf{x}}_N = \mathbf{0}_{N\times 1},\ \mathbf{E}\left\{\tilde{\mathbf{x}}_N\tilde{\mathbf{x}}_N^*\right\} = \boldsymbol{\Lambda}_N,\ \mathbf{E}\left\{\tilde{\mathbf{x}}_N\tilde{\mathbf{x}}_N^\top\right\} = \mathbf{0}_{N\times N} \tag{17}$$

where $\boldsymbol{\Lambda}_N$ is a Hermitian and Toeplitz matrix defined in terms of the autocorrelation sequence $\{\lambda_t\}$ of $\tilde{x}_t$ as $[\boldsymbol{\Lambda}_N]_{ij} = \lambda_{i-j}$. Moreover, the sequence $\{\lambda_k\}$ is absolutely summable, i.e.,

$$\overline{\lambda} := \lim_{N\to\infty} \sum_{t=-N}^{N} |\lambda_t| \tag{18}$$

exists so that the process $\tilde{x}_t$ has a bounded spectral density. We will use $s_{\max}$ and $s_{\min}$ respectively, to denote the supremum and the infenum of the spectral density of $\tilde{x}_t$. Immediately, it follows that [14]

$$s_{\min} \le \text{eig}(\boldsymbol{\Lambda}_t) \le s_{\max}, \quad \forall t > 0 \tag{19}$$

where $\text{eig}(\boldsymbol{\Lambda}_t)$ denotes an eigenvalue of $\boldsymbol{\Lambda}_t$.

In contrast to the previous related work [5], [6], here, we allow $\tilde{x}_t$ to be correlated. This is motivated by many practical applications where the additive noise may not be white.[1] The condition on circularity in Assumption 1 is very common. For instance, we can verify the validity of this assumption when the measured data are the discrete Fourier transform of real-valued data [12] or a complex-valued signal recorded at the receiver in a typical telecommunication or array application [15]. This condition alone, however, is not sufficient to ensure the consistency of NLLS. To see that, consider the concentrated loss function $\hat{\ell}_N(\boldsymbol{\theta})$. It is readily verified from (12) and Assumption 1 that

$$\mathbf{E}\hat{\ell}_N(\boldsymbol{\theta}) = \ell_N(\boldsymbol{\theta}) + N^{-1}\text{tr}\left\{\boldsymbol{\Lambda}_N\boldsymbol{\Pi}^{\perp}_{\boldsymbol{\Psi}_N}(\boldsymbol{\theta})\right\} \tag{20}$$

where $\ell_N(\boldsymbol{\theta})$ is the noise-free concentrated loss function obtained by using $\tilde{\mathbf{x}}_t = \mathbf{0}_{N\times 1}$ in (12). Assuming that the loss function $\hat{\ell}_N(\boldsymbol{\theta})$ converges (in a stochastic sense) to its expected value as $N \to \infty$ (which is true in most of the cases), we can easily notice that in general

$$\arg\min_{\boldsymbol{\theta}} \mathbf{E}\hat{\ell}_N(\boldsymbol{\theta}) \neq \arg\min_{\boldsymbol{\theta}} \ell_N(\boldsymbol{\theta}). \tag{21}$$

The second term in the right-hand side of (20) does not in general have the same minimum point as the first term.[2] Therefore, we need a stronger constraint in the form of (18) to ensure consistency (although it is much weaker than the *whiteness* assumption). Intuitively, from (21), one would expect bias effects in finite sample cases, even if the estimates are consistent. We need a few more assumptions.

Assumption 2: The magnitude of the derivative of the function $f_t(\boldsymbol{\theta}, \mathbf{c})$ with respect to the real or imaginary part of θ_k is bounded from below for all t and for all $1 \le k \le n$, where θ_k denotes the kth component of $\boldsymbol{\theta}$.

Assumption 3: Introduce

$$\boldsymbol{\alpha} := \begin{bmatrix}\boldsymbol{\theta}^\top & \mathbf{c}^\top\end{bmatrix}^\top, \quad \boldsymbol{\alpha}_0 := \begin{bmatrix}\boldsymbol{\theta}_0^\top & \mathbf{c}_0^\top\end{bmatrix}^\top. \tag{22}$$

There exists a compact set $\mathcal{S} \subset \mathbb{C}^{n+p}$ and an integer N_0 such that whenever $\boldsymbol{\alpha} \in \mathcal{S}$

$$0 < b_{\min} \le N^{-1}\text{eig}\left\{\boldsymbol{\Psi}_N^*(\boldsymbol{\theta})\boldsymbol{\Psi}_N(\boldsymbol{\theta})\right\} \le b_{\max}, \quad \forall N \ge N_0 \tag{23}$$

for finite $b_{\min}$ and $b_{\max}$. The true parameter vector $\boldsymbol{\alpha}_0 \in \mathcal{S}$.

Assumption 2 can be seen as an identifiability condition. If Assumption 2 does not hold, the observed data fails to carry enough information about the parameter $\boldsymbol{\theta}$ for large N, even in absence of the measurement noise. However, this condition might be redundant if N is not considered large. Assumption 3 is a persistence of excitation condition [2], [16], which ensures

[1] If the noise statistics is known to the user, it is possible to prewhiten the noise. We will come back to this point later.

[2] On the other hand, if $\boldsymbol{\Lambda}_N = \lambda_0\mathbf{I}_N$, the inequality in (21) can be replaced by an equality since $\text{tr}\{\boldsymbol{\Pi}^{\perp}_{\boldsymbol{\Psi}_N}(\boldsymbol{\theta})\} = N - p$ is independent of $\boldsymbol{\theta}$.

that the information about the parameter can be extracted successfully from the observed noise-corrupted data. This means that $\boldsymbol{\Psi}_N(\boldsymbol{\theta})$ is full column rank, and $N^{-1}\mathbf{f}_N^*(\boldsymbol{\theta},\mathbf{c})\mathbf{f}_N(\boldsymbol{\theta},\mathbf{c})$ is bounded on $\mathcal{S}$ for all $N \geq N_0$. We have the following propositions.

Proposition 1: Introduce the *noise-free* loss function

$$L_N(\boldsymbol{\theta},\mathbf{c}) := N^{-1}\left\{\mathbf{x}_N - \mathbf{f}_N(\boldsymbol{\theta},\mathbf{c})\right\}^*\left\{\mathbf{x}_N - \mathbf{f}_N(\boldsymbol{\theta},\mathbf{c})\right\}. \quad (24)$$

Then, under Assumptions 1–3

$$\lim_{N\to\infty}\left\{\hat{L}_N(\boldsymbol{\theta},\mathbf{c}) - L_N(\boldsymbol{\theta},\mathbf{c})\right\} = \lambda_0 \quad (25)$$

uniformly for all $\boldsymbol{\alpha} \in \mathcal{S}$ with probability one.

Proof: Combining (24) and (9) and using definitions (3)–(8), we have

$$\hat{L}_N(\boldsymbol{\theta},\mathbf{c}) - L_N(\boldsymbol{\theta},\mathbf{c}) = N^{-1}\sum_{t=1}^{N}\left[|\tilde{x}_t|^2 + 2\mathrm{Re}\left\{\tilde{x}_t f_t^*(\boldsymbol{\theta},\mathbf{c})\right\}\right]. \quad (26)$$

Since the fourth-order moments of $\tilde{x}_t$ are bounded, using ergodicity results [2], the first term on the right-hand side of (26) converges to λ_0 with probability one. It remains to be shown that the second term on the right-hand side of (26) converges to zero with probability one. Using Kronecker's lemma [17], [18], it is sufficient to show that $\{q_N(\boldsymbol{\theta},\mathbf{c})\}$ is a Cauchy sequence on $\mathcal{S}$ with probability one, where

$$q_N(\boldsymbol{\theta},\mathbf{c}) := \sum_{t=1}^{N}\frac{2}{t}\mathrm{Re}\left\{\tilde{x}_t f_t^*(\boldsymbol{\theta},\mathbf{c})\right\}.$$

Consider $r_{kl}(\boldsymbol{\theta},\mathbf{c}) := q_{k+l}(\boldsymbol{\theta},\mathbf{c}) - q_k(\boldsymbol{\theta},\mathbf{c})$. Since $\mathbf{E}r_{kl}(\boldsymbol{\theta},\mathbf{c}) = 0$, using Chebychev's inequality [19], we have

$$\begin{aligned}&\mathrm{Prob}\left\{|r_{kl}(\boldsymbol{\theta},\mathbf{c})| \geq \epsilon\right\}\\ &\quad\leq \frac{1}{\epsilon^2}\mathbf{E}r_{kl}^2(\boldsymbol{\theta},\mathbf{c}) = \frac{2}{\epsilon^2}\boldsymbol{f}^*\boldsymbol{\Lambda}_l\boldsymbol{f}\\ &\quad\leq \frac{1}{\epsilon^2}s_{\max}\sum_{j=k+1}^{k+l}\frac{2}{j^2}\left|f_j(\boldsymbol{\theta},\mathbf{c})\right|^2 \end{aligned} \quad (27)$$

where we have used (19) in the last inequality and introduced

$$\boldsymbol{f} = \left[\frac{1}{k+1}f_{k+1}(\boldsymbol{\theta},\mathbf{c}) \quad \cdots \quad \frac{1}{k+l}f_{k+l}(\boldsymbol{\theta},\mathbf{c})\right]^\top. \quad (28)$$

Recall from Assumption 3 that $|f_t(\boldsymbol{\theta},\mathbf{c})|$ is uniformly bounded for all $\boldsymbol{\alpha} \in \mathcal{S}$. Therefore, for any given ϵ and l, it is possible to make the right-hand side of the last inequality in (27) arbitrarily small by increasing k sufficiently. Hence, $\{q_N(\boldsymbol{\theta},\mathbf{c})\}$ is a Cauchy sequence with probability one, and the proposition follows. ■

Proposition 2: Let the sequence of functions $\{\mathcal{L}_N(\theta)\}$ be uniformly convergent (as $N \to \infty$) to a continuous function $\mathcal{L}_\infty(\theta)$ on a compact set Ω, and let $\mathcal{L}_\infty(\theta)$ have a unique global minimum point at θ_*. Let $\hat{\theta}_N$ be a global minimum point of $\mathcal{L}_N(\theta)$ in Ω. Then, $\hat{\theta}_N$ converges to θ_* as $N \to \infty$.

Proof: See [20] and [21]. ■

Assumption 4: $\mathcal{S}$ is dense. The noise-free loss function $L_N(\boldsymbol{\theta},\mathbf{c})$ has a unique global minimum point $\boldsymbol{\alpha}_*(N)$ in the set $\mathcal{S}$, and $\boldsymbol{\alpha}_*(N)$ is an interior point of $\mathcal{S}$. Moreover, $\boldsymbol{\Psi}_N(\boldsymbol{\theta})$ is a continuous function of $\boldsymbol{\theta}$ on $\mathcal{S}$.

Proposition 2 is quite well known and has been used frequently as an important tool for convergence analysis. Assumption 4 is a common identifiability condition. Often, it is required that $\mathcal{S}$ be restricted to satisfy Assumption 4. The problem of estimating the sine wave frequencies from noisy observations [10] is such an example. We also point out that a necessary condition for Assumption 4 to hold asymptotically as $N \to \infty$ is Assumption 2. If the noise-free data do not obey the model (1), then we will have model error. If the model is correct, the noise-free loss function $L_N(\boldsymbol{\theta},\mathbf{c})$ will have a global minimum at $\boldsymbol{\alpha}_0$. This is not the case if the model is incorrect. Then, the minimum point of $L_N(\boldsymbol{\theta},\mathbf{c})$ would be N dependent. We are ready to state our main result in this section.

Proposition 3: Under Assumptions 1–4

$$\lim_{N\to\infty}\hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_*(N) = \mathbf{0}_{(n+p)\times 1} \quad (29)$$

almost surely, where we have denoted $\hat{\boldsymbol{\alpha}}_N := [\hat{\boldsymbol{\theta}}_N^\top \ \hat{\mathbf{c}}_N^\top]^\top$.

Proof: The proof follows by combining Propositions 1 and 2 with Assumptions 4 and 5. ■

Remarks:

- If the model is correct, $\hat{\boldsymbol{\alpha}}_N \to \boldsymbol{\alpha}_0$ almost surely.
- The proof presented here does not assume anything regarding the correctness of the model. Hence, the result in Proposition 3 is valid in the presence of model errors. Note that in presence of model errors, the true parameter vector $\boldsymbol{\alpha}_0 \neq \boldsymbol{\alpha}_*(\infty)$. That would lead to regular bias effects. However, analysis of such bias effects are beyond the scope of this paper.
- Assumption 3 implies that $\{|x_t|\}_{t>0}$ is a bounded sequence. However, by proper normalization of the loss function $\hat{L}_N(\boldsymbol{\theta},\mathbf{c})$, one can ensure the consistency of the parameter estimates even if $\limsup_{t\to\infty}|x_t|$ does not exist. The proof of consistency presented here can be accordingly modified to include that case as well. As a matter of fact, faster convergence of the parameter estimates results if $\{|x_t|\}_{t>0}$ diverges at a higher rate. However, such examples are rarely encountered in practical problems.

We point out here that if the noise statistics is known, that knowledge can be incorporated in the estimation algorithm using the framework of maximum likelihood estimation [22]. In such an approach, an alternative estimate of the parameter $\boldsymbol{\theta}$ is obtained as

$$\hat{\boldsymbol{\theta}}_N^c = \arg\min_{\boldsymbol{\theta}}\left[\min_{\mathbf{c}} N^{-1}\bar{L}_N(\boldsymbol{\theta},\mathbf{c})\right] \quad (30)$$

where $\bar{L}_N(\boldsymbol{\theta},\mathbf{c}) = \{\mathbf{z}_N - \mathbf{f}_N(\boldsymbol{\theta},\mathbf{c})\}^*\boldsymbol{\Lambda}_N^{-1}\{\mathbf{z}_N - \mathbf{f}_N(\boldsymbol{\theta},\mathbf{c})\}$. Eliminating $\mathbf{c}$ as in (10)–(12), we get

$$\hat{\boldsymbol{\theta}}_N^c = \arg\min_{\boldsymbol{\theta}} N^{-1}\{\boldsymbol{\Gamma}_N\mathbf{z}_N\}^*\boldsymbol{\Pi}^\perp_{\boldsymbol{\Gamma}_N\boldsymbol{\Psi}_N}(\boldsymbol{\theta})\{\boldsymbol{\Gamma}_N\mathbf{z}_N\} \quad (31)$$

where

$$\boldsymbol{\Lambda}_N^{-1} = \boldsymbol{\Gamma}_N^*\boldsymbol{\Gamma}_N \quad (32)$$

is the Cholesky decomposition of $\mathbf{\Lambda}_N^{-1}$. Using similar calculations as presented here, it is possible to establish the consistency of $\hat{\boldsymbol{\theta}}_N^C$ under milder conditions (Assumption 1 can be relaxed). This phenomenon can easily be explained since the modified loss function (31) involves a prewhitening step, where the data is transformed by $\mathbf{\Gamma}_N$.

III. Asymptotic Accuracy

In this section, we explore the second-order statistical properties of the estimates. For that purpose, let us introduce some further notations. Since the statistical properties of a complex-valued quantity is usually expressed in terms of the *joint* statistical properties of its real and imaginary parts, it is convenient to use an associated real-valued parameter vector. Let the parameter vectors $\boldsymbol{\theta}$, $\mathbf{c}$, and $\boldsymbol{\alpha}$ [see (22)] be expressed in terms of their real and imaginary parts as

$$\boldsymbol{\theta} = \boldsymbol{\theta}_R + \mathrm{i}\boldsymbol{\theta}_I, \quad \mathbf{c} = \mathbf{c}_R + \mathrm{i}\mathbf{c}_I, \quad \boldsymbol{\alpha} = \boldsymbol{\alpha}_R + \mathrm{i}\boldsymbol{\alpha}_I. \tag{33}$$

Following the usual convention, the joint statistical properties of the estimates $\hat{\boldsymbol{\theta}}_N$ and $\hat{\mathbf{c}}_N$ will be expressed in terms of the statistical properties of the estimates of the real-valued vectors

$$\bar{\boldsymbol{\alpha}} := \begin{bmatrix} \boldsymbol{\alpha}_R \\ \boldsymbol{\alpha}_I \end{bmatrix}, \quad \bar{\boldsymbol{\theta}} := \begin{bmatrix} \boldsymbol{\theta}_R \\ \boldsymbol{\theta}_I \end{bmatrix}, \quad \bar{\mathbf{c}} := \begin{bmatrix} \mathbf{c}_R \\ \mathbf{c}_I \end{bmatrix}. \tag{34}$$

Moreover, we will use $\bar{\boldsymbol{\alpha}}_0$, etc., to denote the true values of $\bar{\boldsymbol{\alpha}}$, etc. Note that the mapping from the complex-valued parameter vector $\boldsymbol{\theta}$, etc., to the associated real-valued parameter vector $\bar{\boldsymbol{\theta}}$ is bijective: Any function of $\boldsymbol{\theta}$ can equivalently be expressed as functions of the associated real-valued parameter vector $\bar{\boldsymbol{\theta}}$. With a slight misuse of notation, we maintain the same functional symbol to denote the *equivalent* function as well. For example, we use $L(\bar{\boldsymbol{\alpha}}) = L(\boldsymbol{\theta}, \mathbf{c})$, and so on. We need the following differentiability assumption.

Assumption 5: The function $\mathbf{\Psi}_N(\boldsymbol{\theta})$ is at least twice differentiable with respect to $\boldsymbol{\theta}_R$ and $\boldsymbol{\theta}_I$ on $\mathcal{S}$.

Proposition 4: Let $\mathcal{H}_N(\bar{\boldsymbol{\alpha}}_0)$ be the Hessian matrix and $\mathbf{g}_N(\bar{\boldsymbol{\alpha}}_0)$ be the gradient vector of the loss function $\hat{L}_N(\bar{\boldsymbol{\alpha}})$ evaluated at $\bar{\boldsymbol{\alpha}} = \bar{\boldsymbol{\alpha}}_0$. Then, under Assumptions 1–5, the asymptotic (as $N \to \infty$) estimation error $\tilde{\bar{\boldsymbol{\alpha}}}_N := \hat{\bar{\boldsymbol{\alpha}}}_N - \bar{\boldsymbol{\alpha}}_0$ is given by

$$\tilde{\bar{\boldsymbol{\alpha}}}_N = -\mathcal{H}_N^{-1}(\bar{\boldsymbol{\alpha}}_0)\mathbf{g}_N(\bar{\boldsymbol{\alpha}}_0). \tag{35}$$

Proof: Recall that $\mathbf{g}_N(\hat{\bar{\boldsymbol{\alpha}}}_N) = \mathbf{0}$. Using this fact in the Taylor's series expansion of $\hat{L}(\bar{\boldsymbol{\alpha}})$ in the neighborhood of $\bar{\boldsymbol{\alpha}}_0$ and neglecting the third and higher order terms in $\hat{\bar{\boldsymbol{\alpha}}}_N - \bar{\boldsymbol{\alpha}}_0$ (which is valid only asymptotically for large N since $\hat{\bar{\boldsymbol{\alpha}}}_N$ is a consistent estimator of $\bar{\boldsymbol{\alpha}}_0$), the proposition follows. The details of the proof are available in many related books and papers; see [2] for example. ∎

Next, the properties of the Hessian matrix $\mathcal{H}_N(\bar{\boldsymbol{\alpha}}_0)$ and the gradient vector $\mathbf{g}_N(\bar{\boldsymbol{\alpha}}_0)$ will be explored. Since we are concerned about the asymptotic distribution of $\tilde{\bar{\boldsymbol{\alpha}}}_N$, it is *not* required to establish the *almost sure* convergence of the relevant quantities (i.e., $\mathcal{H}_N(\bar{\boldsymbol{\alpha}}_0)$ and $\mathbf{g}_N(\bar{\boldsymbol{\alpha}}_0)$), but it would be sufficient to establish weak convergence *in probability*; see [17] and [18] for details. For what follows next, we use the following differentiation notations for a complex vector-valued function $\mathbf{w}(\mathbf{a})$ of a real and vector-valued parameter $\mathbf{a} = [a_1 \ \ldots \ a_m]^\top$, where we use

$$\mathbf{w}^{(k)}(\mathbf{a}) := \frac{\partial \mathbf{w}(\mathbf{a})}{\partial a_k}, \quad \mathbf{w}^{(kl)}(\mathbf{a}) := \frac{\partial^2 \mathbf{w}(\mathbf{a})}{\partial a_k \partial a_l}, \quad 1 \le k, l \le m. \tag{36}$$

Further, the matrix of first-order derivatives will be denoted by

$$\frac{\partial \mathbf{w}(\mathbf{a})}{\partial \mathbf{a}} := \left[\mathbf{w}^{(1)}(\mathbf{a}) \quad \ldots \quad \mathbf{w}^{(m)}(\mathbf{a})\right]. \tag{37}$$

Assumption 6: The following limit result holds:

$$\lim_{N\to\infty} N^{-1} \frac{\mathrm{Re}\left\{Q_N^{kl}(\bar{\boldsymbol{\alpha}})\right\}}{\left[\mathrm{Re}\left\{P_N^{kl}(\bar{\boldsymbol{\alpha}})\right\}\right]^2} = 0, \quad \forall 1 \le k, l \le 2(n+p) \tag{38}$$

where

$$\begin{aligned} P_N^{kl}(\bar{\boldsymbol{\alpha}}) &:= N^{-1}\left\{\mathbf{f}_N^{(k)*}(\bar{\boldsymbol{\alpha}})\mathbf{f}_N^{(l)}(\bar{\boldsymbol{\alpha}})\right\} \\ Q_N^{kl}(\bar{\boldsymbol{\alpha}}) &:= N^{-1}\left\{\mathbf{f}_N^{(kl)*}(\bar{\boldsymbol{\alpha}})\mathbf{f}_N^{(kl)}(\bar{\boldsymbol{\alpha}})\right\}. \end{aligned} \tag{39}$$

This assumption may appear to be restrictive. However, one can notice that neither $Q_N^{kl}(\bar{\boldsymbol{\alpha}})$ nor $P_N^{kl}(\bar{\boldsymbol{\alpha}})$ is required to be bounded by Assumption 6. As a matter of fact, most signal models that satisfy Assumptions 2 and 3 can be shown to satisfy Assumption 6 as well.[3] Recall that for consistency, it is necessary that Assumptions 2 and 3 are satisfied. In that sense, Assumption 6 is not restrictive at all. Our next proposition is a consequence of Assumption 6.

Proposition 5: Under Assumptions 1–6, the asymptotic covariance matrix of $\hat{\bar{\boldsymbol{\alpha}}}_N$ is given by

$$\begin{aligned} \mathbf{C}_{\bar{\boldsymbol{\alpha}}} = \frac{1}{2}&\left[\mathrm{Re}\left\{\nabla\mathbf{f}_N^*\nabla\mathbf{f}_N\right\}\right]^{-1} \\ &\times \left[\mathrm{Re}\left\{\nabla\mathbf{f}_N^*\mathbf{\Lambda}_N\nabla\mathbf{f}_N\right\}\right]\left[\mathrm{Re}\left\{\nabla\mathbf{f}_N^*\nabla\mathbf{f}_N\right\}\right]^{-1} \end{aligned} \tag{40}$$

where

$$\nabla\mathbf{f}_N := \frac{\partial \mathbf{f}_N(\bar{\boldsymbol{\alpha}}_0)}{\partial \bar{\boldsymbol{\alpha}}}. \tag{41}$$

is assumed to have full column rank.

Proof: Differentiating the loss function $\hat{L}_N(\bar{\boldsymbol{\alpha}})$ in (9) with respect to the kth element of $\bar{\boldsymbol{\alpha}}$ and evaluating at $\bar{\boldsymbol{\alpha}} = \bar{\boldsymbol{\alpha}}_0$, we get, using (6) and (8)

$$\begin{aligned} \hat{L}_N^{(k)}(\bar{\boldsymbol{\alpha}}_0) = -N^{-1}&\left[\mathbf{f}_N^{(k)*}(\bar{\boldsymbol{\alpha}}_0)\tilde{\mathbf{x}}_N \right. \\ &\left. + \tilde{\mathbf{x}}_N^*\mathbf{f}_N^{(k)}(\bar{\boldsymbol{\alpha}}_0)\right] \end{aligned} \tag{42}$$

$$\Rightarrow \quad \mathbf{g}_N(\bar{\boldsymbol{\alpha}}_0) = -2N^{-1}\mathrm{Re}\left(\nabla\mathbf{f}_N^*\tilde{\mathbf{x}}_N\right). \tag{43}$$

Hence, to compute the second-order moment of the gradient vector $\mathbf{g}_N(\bar{\boldsymbol{\alpha}}_0)$, using Assumption 1, we have

$$\begin{aligned} &\mathbf{E}\left\{\hat{L}_N^{(k)}(\bar{\boldsymbol{\alpha}}_0)\hat{L}_N^{(l)}(\bar{\boldsymbol{\alpha}}_0)\right\} = 2N^{-2}\mathrm{Re}\left\{\mathbf{f}_N^{(k)*}(\bar{\boldsymbol{\alpha}}_0)\mathbf{\Lambda}_N\mathbf{f}_N^{(l)}(\bar{\boldsymbol{\alpha}}_0)\right\} \\ &\Rightarrow \mathbf{E}\left\{\mathbf{g}_N(\bar{\boldsymbol{\alpha}}_0)\mathbf{g}_N^\top(\bar{\boldsymbol{\alpha}}_0)\right\} = 2N^{-2}\mathrm{Re}\left\{\nabla\mathbf{f}_N^*\mathbf{\Lambda}_N\nabla\mathbf{f}_N\right\}. \end{aligned} \tag{44}$$

Next, we consider the asymptotic properties of the Hessian matrix $\mathcal{H}_N(\bar{\boldsymbol{\alpha}}_0)$. Consider the second derivative of the loss func-

[3]To the best of the knowledge of the authors, all the signal models that satisfy Assumptions 2 and 3 satisfy Assumption 6 as well. However, a rigorous mathematical treatment of this issue is beyond the scope of this paper.

tion $\hat{L}_N(\bar{\boldsymbol{\alpha}})$ with respect to the kth and the lth element of $\bar{\boldsymbol{\alpha}}$. At $\bar{\boldsymbol{\alpha}} = \bar{\boldsymbol{\alpha}}_0$, we get, using (1), (3), and (4)

$$\hat{L}_N^{(kl)}(\bar{\boldsymbol{\alpha}}_0) = 2N^{-1}\mathrm{Re}\left\{\mathbf{f}_N^{(k)*}(\bar{\boldsymbol{\alpha}}_0)\mathbf{f}_N^{(l)}(\bar{\boldsymbol{\alpha}}_0) - \mathbf{f}_N^{(kl)*}(\bar{\boldsymbol{\alpha}}_0)\tilde{\mathbf{x}}\right\}. \tag{45}$$

$$\mathbf{E}\left\{\hat{L}_N^{(kl)}(\bar{\boldsymbol{\alpha}}_0)\right\} = 2N^{-1}\mathrm{Re}\left\{\mathbf{f}_N^{(k)*}(\bar{\boldsymbol{\alpha}}_0)\mathbf{f}_N^{(l)}(\bar{\boldsymbol{\alpha}}_0)\right\} = 2\mathrm{Re}\left\{P_N^{kl}(\bar{\boldsymbol{\alpha}}_0)\right\}. \tag{46}$$

$$\begin{aligned}\mathrm{Var}\left\{\hat{L}_N^{(kl)}(\bar{\boldsymbol{\alpha}}_0)\right\} &= 2N^{-2}\mathrm{Re}\left\{\mathbf{f}_N^{(kl)*}(\bar{\boldsymbol{\alpha}}_0)\boldsymbol{\Lambda}_N\mathbf{f}_N^{(kl)}(\bar{\boldsymbol{\alpha}}_0)\right\}\\ &\le 2N^{-2}s_{\max}\mathrm{Re}\left\{\mathbf{f}_N^{(kl)*}(\bar{\boldsymbol{\alpha}}_0)\mathbf{f}_N^{(kl)}(\bar{\boldsymbol{\alpha}}_0)\right\}\\ &= 2s_{\max}N^{-1}\mathrm{Re}\left\{Q_N^{kl}(\bar{\boldsymbol{\alpha}}_0)\right\}.\end{aligned} \tag{47}$$

Note that in the first equality in (47), we have used (17), whereas in the inequality, we have used (19). Now, using Assumption 6, for large N, we see from (46) and (47) that the mean of each element of the Hessian $\mathcal{H}_N(\bar{\boldsymbol{\alpha}}_0)$ is large compared with the standard deviation. Hence, each element of the Hessian $\mathcal{H}_N(\bar{\boldsymbol{\alpha}}_0)$ converges in mean square sense to the corresponding expected value,[4] i.e.,

$$\begin{aligned}&\hat{L}_N^{(kl)}(\bar{\boldsymbol{\alpha}}_0) \approx 2N^{-1}\mathrm{Re}\left\{\mathbf{f}_N^{(k)*}(\bar{\boldsymbol{\alpha}}_0)\mathbf{f}_N^{(l)}(\bar{\boldsymbol{\alpha}}_0)\right\}\\ &\Rightarrow \mathcal{H}_N(\bar{\boldsymbol{\alpha}}_0) \approx 2N^{-1}\mathrm{Re}\left\{\nabla\mathbf{f}_N^{*}\nabla\mathbf{f}_N\right\}\end{aligned} \tag{48}$$

for large N. Hence, from (35), (44), (48), and Proposition 4, the result follows. ■

The result in Proposition 5 was derived in [5], assuming P_N^{kl} and Q_N^{kl} to be bounded asymptotically. Note that if P_N^{kl} and Q_N^{kl} are asymptotically bounded, Assumption 6 is satisfied anyway. From that point view, Proposition 5 can be seen as a generalization of the results given in [5]. However, from an application point view, it might be too restrictive to assume P_N^{kl} and Q_N^{kl} to be asymptotically bounded, as illustrated by the following example.

Example 1: Consider the problem of estimating frequencies of complex cisoids in noise:

$$x_t = \sum_{k=1}^{p} c_k \mathrm{e}^{i\theta_k t}, \quad z_t = x_t + \tilde{x}_t \tag{49}$$

where $\{c_k\}_{k=1}^p$ are real-valued amplitudes of the sinusoids having frequencies $\{\theta_k\}_{k=1}^p$. A comparison of (49) with (1) reveals that the problem of estimating the frequencies $\{\theta_k\}_{k=1}^p$ and the associated amplitudes $\{c_k\}_{k=1}^p$ can be framed as a separable nonlinear least squares problem. Using the analysis presented so far, we can easily verify that the variable projection estimates of the associated parameters are consistent. It follows after a few steps of calculations that (using our usual notations)

$$\left[\frac{\partial\mathbf{f}_N(\boldsymbol{\theta},\mathbf{c})}{\partial\theta_k}\right]^*\left[\frac{\partial\mathbf{f}_N(\boldsymbol{\theta},\mathbf{c})}{\partial\theta_k}\right] = \mathcal{O}(N^3) \tag{50}$$

$$\left[\frac{\partial^2\mathbf{f}_N(\boldsymbol{\theta},\mathbf{c})}{\partial\theta_k^2}\right]^*\left[\frac{\partial^2\mathbf{f}_N(\boldsymbol{\theta},\mathbf{c})}{\partial\theta_k^2}\right] = \mathcal{O}(N^5). \tag{51}$$

[4]Apparently, the right-hand side of (48) may approach 0 or diverge as $N \to \infty$. If $P_N^{kl}(\bar{\boldsymbol{\alpha}}_0)$ converges to 0 and Assumption 6 is satisfied, then the standard deviation of the left-hand side of (48) converges at a faster rate than its mean. Therefore, the stochastic variation of the left-hand side of (48) can be neglected compared with the mean value. This is true even when the right-hand side of (48) diverges, because in that case, the mean diverges at a faster rate compared with the standard deviation.

Clearly, in this case, neither P_N^{kl} nor Q_N^{kl} are bounded for all k and l. However, it is readily verified that Assumption 7 is satisfied here so that an analogous asymptotic analysis as Proposition 5 can be carried out. ■

Assumption 7: The matrix-valued function $\boldsymbol{\Psi}_N(\boldsymbol{\theta})$ is an analytic function of $\boldsymbol{\theta}$.

This, again, is a mild assumption in the sense that in many practical applications, this assumption is satisfied. Note that $\mathbf{f}_N(\boldsymbol{\theta},\mathbf{c})$ is linear in $\mathbf{c}$. Using this fact combined with Assumption 7, we have $\mathbf{f}_N(\boldsymbol{\theta},\mathbf{c})$ as an analytic function of $\boldsymbol{\theta}$ and $\mathbf{c}$. We also need the following definition. Let us introduce the map $\mathcal{I}: \mathbb{C}^{m\times m} \to \mathbb{R}^{2m\times 2m}$ such that

$$\mathcal{I}(\mathbf{A}) := \begin{bmatrix}\mathrm{Re}(\mathbf{A}) & -\mathrm{Im}(\mathbf{A})\\ \mathrm{Im}(\mathbf{A}) & \mathrm{Re}(\mathbf{A})\end{bmatrix}. \tag{52}$$

Then, it is well known (see [13] and [14], for example) that $\mathcal{I}$ is an isomorphism with respect to matrix multiplication, i.e.,

$$\mathcal{I}(\mathbf{A})\mathcal{I}(\mathbf{B}) = \mathcal{I}(\mathbf{AB}). \tag{53}$$

Note that from the by property of the isomorphism, it also follows that

$$[\mathcal{I}(\mathbf{A})]^{-1} = \mathcal{I}(\mathbf{A}^{-1}). \tag{54}$$

We are now ready to state the main result of this section.

Theorem 1: Let us define

$$\boldsymbol{\Phi}_N(\bar{\boldsymbol{\alpha}}_0) := \frac{\partial\mathbf{f}_N(\bar{\boldsymbol{\alpha}}_0)}{\partial\boldsymbol{\theta}_R}. \tag{55}$$

Then, under Assumptions 1–7, the asymptotic covariance matrix $\mathbf{C}_{\bar{\boldsymbol{\alpha}}}$ is given by

$$\mathbf{C}_{\bar{\boldsymbol{\alpha}}} = \frac{1}{2}\mathcal{I}(\boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma} = \begin{bmatrix}\boldsymbol{\Sigma}_{\theta\theta} & \boldsymbol{\Sigma}_{\theta c}\\ \boldsymbol{\Sigma}_{\theta c}^* & \boldsymbol{\Sigma}_{cc}\end{bmatrix} \tag{56}$$

where

$$\boldsymbol{\Sigma}_{\theta\theta} = \left[\boldsymbol{\Phi}_N^*\boldsymbol{\Pi}_{\boldsymbol{\Psi}_N}^{\perp}\boldsymbol{\Phi}_N\right]^{-1}\boldsymbol{\Phi}_N^*\boldsymbol{\Pi}_{\boldsymbol{\Psi}_N}^{\perp}\boldsymbol{\Lambda}_N \times \boldsymbol{\Pi}_{\boldsymbol{\Psi}_N}^{\perp}\boldsymbol{\Phi}_N\left[\boldsymbol{\Phi}_N^*\boldsymbol{\Pi}_{\boldsymbol{\Psi}_N}^{\perp}\boldsymbol{\Phi}_N\right]^{-1} \tag{57}$$

$$\boldsymbol{\Sigma}_{cc} = \left[\boldsymbol{\Psi}_N^*\boldsymbol{\Pi}_{\boldsymbol{\Phi}_N}^{\perp}\boldsymbol{\Psi}_N\right]^{-1}\boldsymbol{\Psi}_N^*\boldsymbol{\Pi}_{\boldsymbol{\Phi}_N}^{\perp}\boldsymbol{\Lambda}_N \times \boldsymbol{\Pi}_{\boldsymbol{\Phi}_N}^{\perp}\boldsymbol{\Psi}_N\left[\boldsymbol{\Psi}_N^*\boldsymbol{\Pi}_{\boldsymbol{\Phi}_N}^{\perp}\boldsymbol{\Psi}_N\right]^{-1} \tag{58}$$

$$\boldsymbol{\Sigma}_{\theta c} = \left[\boldsymbol{\Phi}_N^*\boldsymbol{\Pi}_{\boldsymbol{\Psi}_N}^{\perp}\boldsymbol{\Phi}_N\right]^{-1}\boldsymbol{\Phi}_N^*\boldsymbol{\Pi}_{\boldsymbol{\Psi}_N}^{\perp}\boldsymbol{\Lambda}_N \times \boldsymbol{\Pi}_{\boldsymbol{\Phi}_N}^{\perp}\boldsymbol{\Psi}_N\left[\boldsymbol{\Psi}_N^*\boldsymbol{\Pi}_{\boldsymbol{\Phi}_N}^{\perp}\boldsymbol{\Psi}_N\right]^{-1} \tag{59}$$

and where we have omitted the arguments $\bar{\boldsymbol{\alpha}}_0$ of the matrices for simplicity.

Proof: Introduce the notation

$$\boldsymbol{\Delta}_N(\bar{\boldsymbol{\alpha}}_0) := \frac{\partial\mathbf{f}_N(\bar{\boldsymbol{\alpha}}_0)}{\partial\bar{\boldsymbol{\alpha}}_R} = \left[\boldsymbol{\Phi}_N(\bar{\boldsymbol{\alpha}}_0) \quad \boldsymbol{\Psi}_N(\bar{\boldsymbol{\alpha}}_0)\right]. \tag{60}$$

Applying Cauchy–Riemann's conditions [23] on analytic functions, we get

$$\begin{aligned}\nabla\mathbf{f}_N &= \frac{\partial\mathbf{f}_N(\bar{\boldsymbol{\alpha}}_0)}{\partial\bar{\boldsymbol{\alpha}}} = \left[\frac{\partial\mathbf{f}_N(\bar{\boldsymbol{\alpha}}_0)}{\partial\bar{\boldsymbol{\alpha}}_R} \quad \frac{\partial\mathbf{f}_N(\bar{\boldsymbol{\alpha}}_0)}{\partial\bar{\boldsymbol{\alpha}}_I}\right]\\ &= \left[\boldsymbol{\Delta}_N(\bar{\boldsymbol{\alpha}}_0) \quad \mathrm{i}\boldsymbol{\Delta}_N(\bar{\boldsymbol{\alpha}}_0)\right].\end{aligned} \tag{61}$$

In what follows, we omit the argument $\bar{\boldsymbol{\alpha}}_0$ for simplicity. From (40) and (61), it follows by straightforward calculation that

$$\begin{aligned} 2\mathbf{C}_{\bar{\boldsymbol{\alpha}}} &= [\mathcal{I}(\boldsymbol{\Delta}_N^*\boldsymbol{\Delta}_N)]^{-1}\mathcal{I}(\boldsymbol{\Delta}_N^*\boldsymbol{\Lambda}_N\boldsymbol{\Delta}_N)[\mathcal{I}(\boldsymbol{\Delta}_N^*\boldsymbol{\Delta}_N)]^{-1} \\ &= \mathcal{I}\left\{(\boldsymbol{\Delta}_N^*\boldsymbol{\Delta}_N)^{-1}\boldsymbol{\Delta}_N^*\boldsymbol{\Lambda}_N\boldsymbol{\Delta}_N(\boldsymbol{\Delta}_N^*\boldsymbol{\Delta}_N)^{-1}\right\} \\ &:= \mathcal{I}\{\boldsymbol{\Sigma}\}. \end{aligned} \tag{62}$$

In order to show the remaining part of the proposition, we apply well-known block matrix inversion results (see, for example, [1] and [14]) to get

$$(\boldsymbol{\Delta}_N^*\boldsymbol{\Delta}_N)^{-1} = \begin{bmatrix} \boldsymbol{\Phi}_N^*\boldsymbol{\Phi}_N & \boldsymbol{\Phi}_N^*\boldsymbol{\Psi}_N \\ \boldsymbol{\Psi}_N^*\boldsymbol{\Phi}_N & \boldsymbol{\Psi}_N^*\boldsymbol{\Psi}_N \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{X}_N & \mathbf{Z}_N \\ \mathbf{Z}_N^* & \mathbf{Y}_N \end{bmatrix} \tag{63}$$

where

$$\begin{aligned} \mathbf{X}_N &= \left[\boldsymbol{\Phi}_N^*\boldsymbol{\Phi}_N - \boldsymbol{\Phi}_N^*\boldsymbol{\Psi}_N\{\boldsymbol{\Psi}_N^*\boldsymbol{\Psi}_N\}^{-1}\boldsymbol{\Psi}_N^*\boldsymbol{\Phi}_N\right]^{-1} \\ &= \left[\boldsymbol{\Phi}_N^*\boldsymbol{\Pi}_{\boldsymbol{\Psi}_N}^{\perp}\boldsymbol{\Phi}_N\right]^{-1}. \end{aligned} \tag{64}$$

Similarly

$$\mathbf{Y}_N = \left[\boldsymbol{\Psi}_N^*\boldsymbol{\Pi}_{\boldsymbol{\Phi}_N}^{\perp}\boldsymbol{\Psi}_N\right]^{-1} \tag{65}$$

$$\mathbf{Z}_N = -\mathbf{X}_N\boldsymbol{\Phi}_N^*\boldsymbol{\Psi}_N\{\boldsymbol{\Psi}_N^*\boldsymbol{\Psi}_N\}^{-1} \tag{66}$$

$$\mathbf{Z}_N^* = -\mathbf{Y}_N\boldsymbol{\Psi}_N^*\boldsymbol{\Phi}_N\{\boldsymbol{\Phi}_N^*\boldsymbol{\Phi}_N\}^{-1}. \tag{67}$$

Hence, by (60) and (63)–(67), we get, after a straightforward block matrix multiplication

$$[\boldsymbol{\Delta}_N^*\boldsymbol{\Delta}_N]^{-1}\boldsymbol{\Delta}_N^* = \begin{bmatrix} \mathbf{X}_N\boldsymbol{\Phi}_N^*\boldsymbol{\Pi}_{\boldsymbol{\Psi}_N}^{\perp} \\ \mathbf{Y}_N\boldsymbol{\Psi}_N^*\boldsymbol{\Pi}_{\boldsymbol{\Phi}_N}^{\perp} \end{bmatrix}. \tag{68}$$

Now from (62)–(68), one can readily verify (57)–(59). ■

Remarks:

- It follows immediately from (56) that the complex-valued random vector $\hat{\alpha}_N$ is asymptotically circular, i.e.,

$$\mathbf{E}\hat{\boldsymbol{\alpha}}_N\hat{\boldsymbol{\alpha}}_N^* = \boldsymbol{\Sigma}, \quad \mathbf{E}\hat{\boldsymbol{\alpha}}_N\hat{\boldsymbol{\alpha}}_N^{\top} = \mathbf{0} \tag{69}$$

$$\mathbf{E}\hat{\boldsymbol{\theta}}_N\hat{\boldsymbol{\theta}}_N^* = \boldsymbol{\Sigma}_{\theta\theta}, \quad \mathbf{E}\hat{\boldsymbol{\theta}}_N\hat{\boldsymbol{\theta}}_N^{\top} = \mathbf{0} \tag{70}$$

$$\mathbf{E}\hat{\mathbf{c}}_N\hat{\mathbf{c}}_N^* = \boldsymbol{\Sigma}_{cc}, \quad \mathbf{E}\hat{\mathbf{c}}_N\hat{\mathbf{c}}_N^{\top} = \mathbf{0}. \tag{71}$$

- The asymptotic covariance matrix of the real-valued parameter vector $\hat{\bar{\boldsymbol{\theta}}}_N$ is given by $\mathbf{C}_{\bar{\boldsymbol{\theta}}} = (1/2)\mathcal{I}(\boldsymbol{\Sigma}_{\theta\theta})$, and that of $\hat{\bar{\mathbf{c}}}_N$ is given by $\mathbf{C}_{\bar{\mathbf{c}}} = (1/2)\mathcal{I}(\boldsymbol{\Sigma}_{cc})$.
- Assumption 7 is redundant if the parameter vector is real valued. The associated analysis is exactly similar but simpler. The resulting expressions are also similar, where $\hat{\boldsymbol{\alpha}}_N$ is a real-valued random vector with asymptotic covariance matrix $\mathbf{C}_{\boldsymbol{\alpha}} = (1/2)\boldsymbol{\Sigma}$. We also have $\mathbf{C}_{\boldsymbol{\theta}} = (1/2)\boldsymbol{\Sigma}_{\theta\theta}$ and $\mathbf{C}_{\mathbf{c}} = (1/2)\boldsymbol{\Sigma}_{cc}$.
- There are applications where $\boldsymbol{\theta}$ is real valued but $\mathbf{c}$ is complex valued. Using similar but more tedious calculations, one can show that the covariance matrix of the real-valued (in this case) parameter $\boldsymbol{\theta}$ is given by (under Assumptions 1–6)

$$\mathbf{C}_{\boldsymbol{\theta}} = \frac{\lambda_0}{2}\left[\mathrm{Re}\left(\boldsymbol{\Phi}_N^*\boldsymbol{\Pi}_{\boldsymbol{\Psi}_N}^{\perp}\boldsymbol{\Phi}_N\right)\right]^{-1} \tag{72}$$

when $\tilde{x}_t$ is a white noise sequence. However, when the noise is colored, the expressions take more complex forms.
- There are many applications where N is finite, but the signal-to-noise ratio is large. It can be easily seen that the expressions of the covariance matrix are the same in such cases. In this context, it can be noted that such similar expressions as (57) and (72) for the parameter variance exist in the literature for large signal-to-noise ratio; see [7] and [8].

Proposition 6: Under Assumptions 1–6, the real-valued estimate $\sqrt{N}\hat{\bar{\boldsymbol{\alpha}}}_N$ is an asymptotically jointly Gaussian random vector with covariance matrix $(N/2)\mathbf{C}_{\bar{\boldsymbol{\alpha}}}$ if $P_N^{kl}(\bar{\boldsymbol{\alpha}})$ is bounded for all $1 \le k, l \le 2(n+p)$. Furthermore, if Assumption 7 is satisfied, then $\sqrt{N}\hat{\boldsymbol{\alpha}}_N$ is asymptotically complex Gaussian distributed with covariance matrix $N\boldsymbol{\Sigma}$.

Proof: Note that [see (43)]

$$\mathbf{g}_N(\bar{\boldsymbol{\alpha}}_0) = N^{-1}\sum_{t=1}^{N} 2\mathrm{Re}\left\{\left[\frac{\partial f_t(\bar{\boldsymbol{\alpha}}_0)}{\partial\bar{\boldsymbol{\alpha}}}\right]^*\tilde{x}_t\right\}. \tag{73}$$

Since $P_N^{kl}(\bar{\boldsymbol{\alpha}})$ is bounded for all k, l, the coefficients of $\tilde{x}_t$ in the summation (73) are of bounded magnitude for all t. Therefore, it follows that $\sqrt{N}\mathbf{g}_N(\bar{\boldsymbol{\alpha}}_0)$ is asymptotically Gaussian [2]. The proposition then follows from (35) since a linear transform of a Gaussian random vector is Gaussian [19]. If Assumption 7 holds, then from Theorem 1, the complex-valued estimate $\sqrt{N}\hat{\boldsymbol{\alpha}}_N$ is complex Gaussian with covariance matrix $N\boldsymbol{\Sigma}$. ■

Remarks:

- Note that if $P_N^{kl}(\bar{\boldsymbol{\alpha}})$ is bounded, then $N\mathbf{C}_{\bar{\boldsymbol{\alpha}}}$ is finite for all N.
- If the parameters are real valued, then $\sqrt{N}\hat{\boldsymbol{\alpha}}_N$ is Gaussian with covariance matrix $N\boldsymbol{\Sigma}$.
- Proposition 6 was proved in [5] for the special case when $\tilde{x}_t$ is white and the data and the parameter are real valued.
- We point out here that the general requirements for asymptotic normality of the parameter estimates are quite stringent in the sense that in many applications, P_N^{kl} would not be bounded; see Example 1, for instance. However, we point out that the condition on boundedness of P_N^{kl} has been relaxed in [6] by imposing a few more constraints when the noise sequence $\tilde{x}_t$ is assumed to be white.

IV. Cramér–Rao Bound

In this section, we compare the results derived in the last section with the Cramér–Rao bound (CRB). Next, we derive a bound on the loss in the statistical efficiency. Finally, we derive a sufficient condition for the NLLS to achieve the CRB. First, we have the next proposition giving an expression for the CRB.

Proposition 7: Under Assumption 5, and a few more regularity conditions [22] on the distribution function of $\tilde{\mathbf{x}}_N$, the CRB of the estimation problem is given by

$$\overline{\mathbf{C}}_{\bar{\boldsymbol{\alpha}}} = \frac{1}{2}\left[\nabla\mathbf{f}_N^*\boldsymbol{\Lambda}_N^{-1}\nabla\mathbf{f}_N\right]^{-1}. \tag{74}$$

Furthermore, if Assumption 7 holds, then the CRB for $\bar{\boldsymbol{\theta}}$ is given by

$$\begin{aligned}\overline{\mathbf{C}}_{\bar{\boldsymbol{\theta}}} &= \frac{1}{2}\mathcal{I}\left(\left[\boldsymbol{\Phi}_N^*\boldsymbol{\Lambda}_N^{-1}\boldsymbol{\Phi}_N - \boldsymbol{\Phi}_N^*\boldsymbol{\Lambda}_N^{-1}\boldsymbol{\Psi}_N \right.\right. \\ &\qquad \left.\left. \times \left\{\boldsymbol{\Psi}_N^*\boldsymbol{\Lambda}_N^{-1}\boldsymbol{\Psi}_N\right\}^{-1}\boldsymbol{\Psi}_N^*\boldsymbol{\Lambda}_N^{-1}\boldsymbol{\Phi}_N\right]^{-1}\right) \\ &= \frac{1}{2}\mathcal{I}\left(\left\{[\boldsymbol{\Gamma}_N\boldsymbol{\Phi}_N]^*\boldsymbol{\Pi}^{\perp}_{\boldsymbol{\Gamma}_N\boldsymbol{\Psi}_N}[\boldsymbol{\Gamma}_N\boldsymbol{\Phi}_N]\right\}^{-1}\right) \qquad (75)\end{aligned}$$

and that of $\bar{\mathbf{c}}$ is given by

$$\begin{aligned}\overline{\mathbf{C}}_{\bar{\mathbf{c}}} &= \frac{1}{2}\mathcal{I}\left(\left[\boldsymbol{\Psi}_N^*\boldsymbol{\Lambda}_N^{-1}\boldsymbol{\Psi}_N - \boldsymbol{\Psi}_N^*\boldsymbol{\Lambda}_N^{-1}\boldsymbol{\Phi}_N \right.\right. \\ &\qquad \left.\left. \times \left\{\boldsymbol{\Phi}_N^*\boldsymbol{\Lambda}_N^{-1}\boldsymbol{\Phi}_N\right\}^{-1}\boldsymbol{\Phi}_N^*\boldsymbol{\Lambda}_N^{-1}\boldsymbol{\Psi}_N\right]^{-1}\right) \\ &= \frac{1}{2}\mathcal{I}\left(\left\{[\boldsymbol{\Gamma}_N\boldsymbol{\Psi}_N]^*\boldsymbol{\Pi}^{\perp}_{\boldsymbol{\Gamma}_N\boldsymbol{\Phi}_N}[\boldsymbol{\Gamma}_N\boldsymbol{\Psi}_N]\right\}^{-1}\right) \qquad (76)\end{aligned}$$

where $\bar{\boldsymbol{\Gamma}}_N$ is the Cholesky factor of $\boldsymbol{\Lambda}_N^{-1}$, as before [see (32)], and we have omitted the argument $\bar{\boldsymbol{\alpha}}_0$ for simplicity.

Proof: The proof for (74) can be found in [24]. For the remaining part, we see that (61) holds under Assumption 7 so that

$$2\overline{\mathbf{C}}_{\bar{\boldsymbol{\alpha}}} = \mathcal{I}\left(\left[\boldsymbol{\Delta}_N^*\boldsymbol{\Lambda}_N^{-1}\boldsymbol{\Delta}_N\right]^{-1}\right). \qquad (77)$$

Noticing the similarity of

$$\boldsymbol{\Delta}_N^*\boldsymbol{\Lambda}_N^{-1}\boldsymbol{\Delta}_N = \begin{bmatrix} \{\boldsymbol{\Gamma}_N\boldsymbol{\Phi}_N\}^*\{\boldsymbol{\Gamma}_N\boldsymbol{\Phi}_N\} & \{\boldsymbol{\Gamma}_N\boldsymbol{\Phi}_N\}^*\{\boldsymbol{\Gamma}_N\boldsymbol{\Psi}_N\} \\ \{\boldsymbol{\Gamma}_N\boldsymbol{\Psi}_N\}^*\{\boldsymbol{\Gamma}_N\boldsymbol{\Phi}_N\} & \{\boldsymbol{\Gamma}_N\boldsymbol{\Psi}_N\}^*\{\boldsymbol{\Gamma}_N\boldsymbol{\Psi}_N\} \end{bmatrix} \qquad (78)$$

with the first equality in (63), we get (75) and (76) using (64) and (65). ■

If $\{\tilde{x}_t\}_{t\geq 0}$ is a white noise sequence so that $\boldsymbol{\Lambda}_N = \lambda_0\mathbf{I}_N$ holds, then $\hat{\boldsymbol{\theta}}_N$ is a maximum likelihood estimate [1]. Therefore, we expect that the asymptotic covariance matrix of the parameter estimates to achieve the Cramér–Rao lower bound. This can be verified by comparing (74) with (40)

$$\mathbf{C}_{\bar{\boldsymbol{\alpha}}} = \overline{\mathbf{C}}_{\bar{\boldsymbol{\alpha}}} = \frac{\lambda_0}{2}\left[\nabla\mathbf{f}_N^*\nabla\mathbf{f}_N\right]^{-1}. \qquad (79)$$

However, this does not follow immediately from the theory of maximum likelihood estimation since some of the standard assumptions [22] have been considerably relaxed in the current context. For instance, the process z_t is nonstationary in general in the current context. Notice that if Assumption 6 does not hold, (40) may cease to hold any longer. However, it is an open question if there exists such an example where Assumption 6 ceases to hold even when the conditions of consistency (i.e. Assumptions 2 and 3) hold.

If the noise covariance structure is known to the user, then (31) should be used. If a similar analysis is carried out for the loss function in (31), it can be shown that the associated estimate achieves the CRB asymptotically. This is due to the prewhitening step involved in (31), as discussed before. However, in most practical applications, $\boldsymbol{\Lambda}_N$ is generally unknown to the user. In such a case, it is of interest to know the difference in the achieved accuracy by NLLS and the best achievable performance. In the next proposition, we present a bound on the loss of statistical efficiency.

Proposition 8: The following inequality holds:

$$\mathbf{C}_{\bar{\boldsymbol{\alpha}}} - \overline{\mathbf{C}}_{\bar{\boldsymbol{\alpha}}} \leq \frac{1}{2}(s_{\max} - s_{\min})\left[\nabla\mathbf{f}_N^*\nabla\mathbf{f}_N\right]^{-1} \qquad (80)$$

where for two matrices $\mathbf{A}$ and $\mathbf{B}$, we write $\mathbf{A} \geq \mathbf{B}$, if $\mathbf{A} - \mathbf{B}$ is non-negative definite.

Proof: Note from (19) that

$$\begin{aligned}&\nabla\mathbf{f}_N^*\boldsymbol{\Lambda}_N\nabla\mathbf{f}_N \leq s_{\max}\nabla\mathbf{f}_N^*\nabla\mathbf{f}_N \\ &\quad\Rightarrow\quad 2\mathbf{C}_{\bar{\boldsymbol{\alpha}}} \leq s_{\max}\left[\nabla\mathbf{f}_N^*\nabla\mathbf{f}_N\right]^{-1}. \qquad (81)\end{aligned}$$

Similarly

$$\begin{aligned}&\nabla\mathbf{f}_N^*\boldsymbol{\Lambda}_N^{-1}\nabla\mathbf{f}_N \leq s_{\min}^{-1}\nabla\mathbf{f}_N^*\nabla\mathbf{f}_N \\ &\quad\Rightarrow\quad 2\overline{\mathbf{C}}_{\bar{\boldsymbol{\alpha}}} \geq s_{\min}\left[\nabla\mathbf{f}_N^*\nabla\mathbf{f}_N\right]^{-1}. \qquad (82)\end{aligned}$$

Since the sum of two non-negative definite matrices is a non-negative definite matrix, we get (80) by combining (81) and (82). ■

The loss in efficiency is low if the spectrum of the noise sequence is flat. This observation is very common in the literature of system identification. We can also verify that for white noise where $s_{\max} = s_{\min}$, there is no loss in statistical efficiency. The bound given by Proposition 8 is applicable in quite general cases. However, it might be interesting to investigate if the estimate $\bar{\boldsymbol{\alpha}}$ can achieve the CRB even if the additive noise is not white. The following proposition gives a sufficient condition in that direction.

Proposition 9: If each of the columns of $\nabla\mathbf{f}_N$ is an asymptotic eigenvector of $\boldsymbol{\Lambda}_N$, i.e.,

$$\boldsymbol{\Lambda}_N\nabla\mathbf{f}_N = \nabla\mathbf{f}_N\mathbf{D}_N \qquad (83)$$

for some real-valued nonsingular diagonal matrix $\mathbf{D}_N$ as $N \to \infty$, then the estimate $\bar{\boldsymbol{\alpha}}$ achieves the CRB asymptotically.

Proof: Using (83) in (40), we get

$$\mathbf{C}_{\bar{\boldsymbol{\alpha}}} = \frac{1}{2}\mathbf{D}_N\left[\mathrm{Re}\left(\nabla\mathbf{f}_N^*\nabla\mathbf{f}_N\right)\right]^{-1}. \qquad (84)$$

Now, by (83), we also have

$$\boldsymbol{\Lambda}_N^{-1}\nabla\mathbf{f}_N = \nabla\mathbf{f}_N\mathbf{D}_N^{-1} \qquad (85)$$

Hence, by (74) and (85), we see that

$$\overline{\mathbf{C}}_{\bar{\boldsymbol{\alpha}}} = \frac{1}{2}\mathbf{D}_N\left[\mathrm{Re}\left(\nabla\mathbf{f}_N^*\nabla\mathbf{f}_N\right)\right]^{-1}. \qquad (86)$$

Combining (84) and (86), we prove the proposition. ■

Following the results in Proposition 9, it might be interesting to investigate the asymptotic eigenvalue distribution of a general covariance matrix $\boldsymbol{\Lambda}_N$ as $N \to \infty$. In this context, we use a remarkable result [25], [26]. Let us introduce

$$\boldsymbol{u}_N(\omega) = [1 \quad \cos(\omega) \quad \ldots \quad \cos\{\omega(N-1)\}] \qquad (87)$$

$$\boldsymbol{v}_N(\omega) = [0 \quad \sin(\omega) \quad \ldots \quad \sin\{\omega(N-1)\}]. \qquad (88)$$

Then, one can show that [25], [26] as $N \to \infty$

$$\boldsymbol{\Lambda}_N\{\boldsymbol{u}_N(\omega) + \mathrm{i}\boldsymbol{v}_N(\omega)\} = s(\omega)\{\boldsymbol{u}_N(\omega) + \mathrm{i}\boldsymbol{v}_N(\omega)\} \qquad (89)$$

where $s(\omega)$ is the spectral density of $\tilde{x}_t$ at frequency ω. This result will play the main role in Example 2 in the next section,

where we illustrate the result in Proposition 9 for a practical application.

V. Real-Valued Data

In this section, we consider the case where data are real valued. The resulting analysis is similar. However, the expressions are not exactly the same. Next, we briefly summarize the results in case the data are real valued.

Proposition 10: If the observed data are real valued, then under Assumptions 1–6 (with proper modifications), the asymptotic covariance matrix of $\hat{\bar{\boldsymbol{\alpha}}}_N$ is given by

$$\mathbf{C}_{\bar{\boldsymbol{\alpha}}} = \left[\nabla \mathbf{f}_N^\top \nabla \mathbf{f}_N\right]^{-1} \left[\nabla \mathbf{f}_N^\top \boldsymbol{\Lambda}_N \nabla \mathbf{f}_N\right] \left[\nabla \mathbf{f}_N^\top \nabla \mathbf{f}_N\right]^{-1} \quad (90)$$

where $\nabla \mathbf{f_N} := (\partial \mathbf{f}_N(\bar{\boldsymbol{\alpha}}_0)/\partial \bar{\boldsymbol{\alpha}})$. Furthermore, if Assumption 7 holds, then $\mathbf{C}_{\bar{\boldsymbol{\alpha}}}$ is block diagonal with

$$\mathbf{E}\hat{\boldsymbol{\alpha}}_{RN}\hat{\boldsymbol{\alpha}}_{RN}^\top = \mathbf{E}\hat{\boldsymbol{\alpha}}_{IN}\hat{\boldsymbol{\alpha}}_{IN}^\top = \boldsymbol{\Sigma}, \quad \mathbf{E}\hat{\boldsymbol{\alpha}}_{RN}\hat{\boldsymbol{\alpha}}_{IN}^\top = \mathbf{0} \quad (91)$$

where $\boldsymbol{\Sigma}$ is defined as in (56)–(59).

Proof: The proof will follow the exactly similar set of calculations given in Proposition 5 and Theorem 1. First, note that for the real-valued data, the gradient vector $\mathbf{g}_N(\bar{\boldsymbol{\alpha}}_0)$ is given by [see (42) and (43)]

$$\begin{aligned} \mathbf{g}_N(\bar{\boldsymbol{\alpha}}_0) &= 2N^{-1}\nabla \mathbf{f}_N^\top \tilde{\mathbf{x}}_N \\ \Rightarrow \quad \mathbf{E}\mathbf{g}_N(\bar{\boldsymbol{\alpha}}_0)\mathbf{g}_N^\top(\bar{\boldsymbol{\alpha}}_0) &= 4N^{-2}\nabla \mathbf{f}_N^\top \boldsymbol{\Lambda}_N \nabla \mathbf{f}_N. \end{aligned} \quad (92)$$

Notice that the difference between (44) and (92) is due to the fact that for the real-valued data case, the *circularity* property of $\tilde{\mathbf{x}}$ is no longer there. Using exactly the similar steps as in (45)–(47), one can readily verify using Assumption 6 (with proper modifications) that

$$\mathcal{H}_N(\bar{\boldsymbol{\alpha}}_0) = 2N^{-1}\nabla \mathbf{f}_N^\top \nabla \mathbf{f}_N. \quad (93)$$

Now combining (35), (92), and (93), we get (90). Furthermore, if the Assumption 7 is satisfied, we can repeat exactly same steps as in (61)–(68) to see that $\mathbf{C}_{\bar{\boldsymbol{\alpha}}} = \mathcal{I}(\boldsymbol{\Sigma})$. Finally, (91) follows since $\boldsymbol{\Sigma}$ is a real-valued matrix in this case. ∎

Note that the above result is in agreement with the results derived in [5] and [6]. The results in Propositions 7 and 8 are also modified if the data are real valued. The CRB is given by (see [24] for details)

$$\overline{\mathbf{C}}_{\bar{\boldsymbol{\alpha}}} = \mathcal{I}\left(\left[\boldsymbol{\Delta}_N^* \boldsymbol{\Lambda}_N^{-1} \boldsymbol{\Delta}_N\right]^{-1}\right) \quad (94)$$

and the result in Proposition 8 is modified as

$$\mathbf{C}_{\bar{\boldsymbol{\alpha}}} - \overline{\mathbf{C}}_{\bar{\boldsymbol{\alpha}}} \leq (s_{\max} - s_{\min})\left[\nabla \mathbf{f}_N^* \nabla \mathbf{f}_N\right]^{-1} \quad (95)$$

which is a fairly straightforward extension of the proof of Proposition 8, using (94) and Proposition 10. Finally, if the parameter vector $\boldsymbol{\alpha}$ is real valued, Assumption 7 is redundant. The covariance matrix of $\hat{\boldsymbol{\alpha}}_N$ is then given by $\mathbf{C}_{\boldsymbol{\alpha}} = \boldsymbol{\Sigma}$. The result in Proposition 9 also is directly applicable to the case when the data are real valued. In the following example, we illustrate Proposition 9 for real-valued data.

Example 2: In this example, we consider a sine wave in noise where

$$z_t = f_t(\omega, \phi, c) + \tilde{x}_t = c\sin(\omega t + \phi) + \tilde{x}_t, \quad t \geq 0. \quad (96)$$

Here, we are interested in estimating the real-valued parameters ω, c, and ϕ from $\{z_t\}_{t=0}^{N-1}$. Clearly, f_t is linear in c and nonlinear in ϕ and ω. In this case, we will denote $\boldsymbol{\theta} = [\omega \quad \phi]^\top$. Let us denote [see (87) and (88)]

$$\boldsymbol{t}_N = [0 \quad 1 \quad \ldots \quad N-1]^\top \quad (97)$$

$$\boldsymbol{w}_N(\omega, \phi) = \boldsymbol{u}_N(\omega)\sin(\phi) + \boldsymbol{v}_N(\omega)\cos(\phi). \quad (98)$$

Using these notations, it is readily verified that

$$\nabla \mathbf{f}_N = [c\boldsymbol{t}_N \odot \boldsymbol{w}_N(\omega, \phi_1) \quad c\boldsymbol{w}_N(\omega, \phi_1) \quad \boldsymbol{w}_N(\omega, \phi)] \quad (99)$$

where $\phi_1 = \phi + \pi/2$, and $\odot$ denotes the Hadamard product (i.e., element-wise multiplication) between two matrices. From (89), we can see that each of $\boldsymbol{u}_N(\omega)$ and $\boldsymbol{v}_N(\omega)$ are real-valued asymptotic eigenvectors of Λ_N as $N \to \infty$, with an associated eigenvalue $s(\omega)$. Thus, any linear combination of $\boldsymbol{u}_N(\omega)$ and $\boldsymbol{v}_N(\omega)$ will also be an asymptotic eigenvector of Λ_N, as $N \to \infty$. Hence, using (98), we see that

$$\boldsymbol{\Lambda}_N \boldsymbol{w}_N(\omega, \phi) = \boldsymbol{w}_N(\omega, \phi)s(\omega) \quad (100)$$

for large N and for all ϕ. Note that (100) is an $N \times 1$ vector equation, where each component of the vector on the left-hand side is a convergent infinite sum, and the corresponding component on the right-hand side is the limit of the sum as $N \to \infty$. Hence, we can differentiate (100) with respect to ω to get

$$\begin{aligned} &\boldsymbol{\Lambda}_N \left\{\boldsymbol{t}_N \odot \boldsymbol{w}_N\left(\omega, \phi + \frac{\pi}{2}\right)\right\} \\ &= \left\{\boldsymbol{t}_N \odot \boldsymbol{w}_N\left(\omega, \phi + \frac{\pi}{2}\right)\right\} s(\omega) + \boldsymbol{w}_N(\omega, \phi)\frac{\partial s(\omega)}{\partial \omega}. \end{aligned} \quad (101)$$

Now, by Assumption 2, the derivative of the spectral density $s(\omega)$ is bounded for all ω (note that at this point, we are assuming that the autocorrelation sequence λ_t decays faster than t^{-2}, which is a little stronger that Assumption 1). Therefore, on the right-hand side of (101), the norm of first term is large compared with that of the second term. Hence, for large N, one can approximate

$$\boldsymbol{\Lambda}_N \left\{\boldsymbol{t}_N \odot \boldsymbol{w}_N\left(\omega, \phi + \frac{\pi}{2}\right)\right\} = \left\{\boldsymbol{t}_N \odot \boldsymbol{w}_N\left(\omega, \phi + \frac{\pi}{2}\right)\right\} s(\omega) \quad (102)$$

for all ϕ. Combining (99), (100), and (102), we see from Proposition 9 that the estimates of $\boldsymbol{\theta}$ and c will achieve CRB asymptotically as $N \to \infty$. ∎

The results of Example 2 can easily be generalized for multiple sine waves in noise as well as in case where the sinusoids are complex valued. We also point out here that the same result for the complex-valued sine waves was derived in [11]. However, the method of analysis used here is easier and provides a more clear perspective on this issue. For numerical illustrations on this issue, see [11]. More illustrations can be found in [7], where NLLS is dealt in a general context.

VI. Conclusions

In this paper, we attempted to carry out a complete asymptotic analysis of the separable nonlinear least squares algorithm. Throughout the analysis, we have maintained rather mild assumptions on the data model and the measurement noise sequence. The consistency analysis presented here can be seen as a generalization of previously proved results but with the weaker

assumptions that the data are complex valued and that the additive noise can be colored. In the accuracy analysis, the results of Proposition 5 have been proved under relaxed assumptions (Assumption 6). It has also been shown in Theorem 1 that if the functions involved are analytic functions of the parameters, the estimates are asymptotically circular. Asymptotic normality has been established. Next, the accuracy expressions were compared with the CRB (Proposition 7), and we have given a general bound on the loss in statistical efficiency in Proposition 8. Finally, in Proposition 9, we have established a sufficient condition for the NLLS to achieve the CRB. If the measurement noise is white, NLLS achieves the CRB under rather mild conditions.

It is interesting to investigate the existence of examples where Assumption 6 ceases to hold in spite of Assumptions 2 and 3 being satisfied. In such a case, NLLS would be a consistent maximum likelihood estimate that may be unable to achieve asymptotic efficiency.

REFERENCES

[1] J. M. Mendel, *Lessons in Estimation Theory for Signal Processing, Communications and Control*. Englewood Cliffs, NJ: Prentice-Hall, 1995.

[2] T. Söderström and P. Stoica, *System Identification*. Hemel Hempstead, U.K.: Prentice-Hall Int., 1989.

[3] G. H. Golub and V. Pereyra, "The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate," *SIAM J. Numerical Anal.*, vol. 10, no. 2, pp. 413–432, 1973.

[4] ——, "Separable nonlinear least squares: The variable projection method and its applications," *Inverse Problems*, vol. 19, no. 2, pp. R1–R26, Apr. 2003.

[5] R. I. Jennrich, "Asymptotic properties of the nonlinear least squares estimators," *Ann. Math. Statist.*, vol. 40, no. 2, pp. 633–643, 1969.

[6] C. F. Wu, "Asymptotic theory of nonlinear least squares estimation," *Ann. Statist.*, vol. 9, no. 3, pp. 501–513, 1981.

[7] J. Ängeby, M. Viberg, and T. Gustafsson, "Non-linear instanteneous least squares method and its high SNR analysis," in *Proc. ICASSP*, vol. 3, Phoenix, AZ, Mar. 1999, pp. 1277–1280.

[8] K. Mahata, S. Mousavi, T. Söderström, M. Mossberg, U. Valdek, and L. Hillström, "On the use of flexural wave propagation experiments for identification of complex modulus," *IEEE Trans. Contr. Syst. Technol.*, vol. 11, pp. 863–874, Nov. 2003.

[9] M. Mossberg, L. Hillström, and T. Söderström, "Non-parametric identification of viscoelastic materials from wave propagation experiments," *Automatica*, vol. 37, no. 4, pp. 511–521, Apr. 2001.

[10] P. Stoica, R. Moses, B. Friedlander, and T. Söderström, "Maximum likelihood estimation of the parameters of multiple sinusoids from noisy measurements," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 378–392, Mar. 1989.

[11] P. Stoica, A. Jakobsson, and J. Li, "Cisoid parameter estimation in the colored noise case—asymptotic Cramer-Rao bound, maximum likelihood, and nonlinear least-squares," *IEEE Trans. Signal Processing*, vol. 45, pp. 2048–2059, July 1997.

[12] D. R. Brillinger, *Time Series: Data Analysis and Theory*. New York: Holt, Rinehart, and Winston, 1975.

[13] N. R. Goodman, "Statistical analysis based on certain multivariate complex distribution (an introduction)," *Ann. Math. Statist.*, vol. 34, no. 1, pp. 152–177, 1963.

[14] T. Söderström, *Discrete-Time Stochastic Systems*, Second ed. London, U.K.: Springer-Verlag, 2002.

[15] H. L. Van Trees, *Detection, Estimation, and Modulation Theory: Optimum Array Processing*. New York: Wiley, 2002, vol. 4.

[16] L. Ljung, *System Identification—Theory for the User*, Second ed. Upper Saddle River, NJ: Prentice-Hall, 1999.

[17] P. Billingsley, *Probability and Measure*, Third ed. New York: Wiley, 1995.

[18] R. G. Laha and V. K. Rohatgi, *Probability Theory*. New York: Wiley, 1979.

[19] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. New York: McGraw-Hill, 1991.

[20] B. Porat, *Digital Processing of Random Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1994.

[21] T. Söderström, "Convergence properties of the generalized least squares identification algorithm," *Automatica*, vol. 10, pp. 617–626, 1974.

[22] M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics*. London, U.K.: Griffin, 1966, vol. 2.

[23] R. V. Churchill and J. W. Brown, *Complex Variables and Applications*, Fourth ed: McGraw-Hill, 1989.

[24] P. Stoica and R. Moses, *Introduction to Spectral Analysis*. Upper Saddle River, NJ: Prentice-Hall, 1997.

[25] R. M. Gray, "On the asymptotic eigenvalue distribution of Toeplitz matrices," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 725–730, Nov. 1972.

[26] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ: Prentice-Hall, 1993.

Kaushik Mahata received the B.E degree in electrical engineering from Jadavpur University, Kolkata, India, in 1996 and the M.E. degree in signal processing from the Indian Institute of Science, Bangalore, in 2000. He received the Ph.D. degree from Uppsala University, Uppsala, Sweden, in 2003.

Currently, he is a research academic with the Centre for Complex Dynamic Systems and Control, University of Newcastle, Callaghan, Australia. His research interest includes system identification, control and simulation of distributed parameter systems, and spectrum analysis.

Torsten Söderström (F'92) was born in Malmö, Sweden, in 1945. He received the M.Sc. degree (civilingenjör) in engineering physics in 1969 and the Ph.D. degree in automatic control in 1973, both from Lund Institute of Technology, Lund, Sweden.

Since 1974, he has been with Uppsala University, Uppsala, Sweden, where he is a professor of automatic control. From 1975 to 1998, he was the head of the Systems and Control Group, which now is a part of Department of Information Technology. His main research interests are in the fields of system identification, signal processing, and control of mechanical systems. He is the author or coauthor of many technical papers and four books, the most recent being *Discrete-Time Stochastic Systems* (New York: Springer-Verlag, 2002, Second ed). Since 1992, he has been an editor of *Automatica* for the area of system parameter estimation.

Dr. Söderström was, with coauthors, given an *Automatica* Paper Prize Award in 1981.