A Globally Consistent Nonlinear Least Squares Estimator for Identification of Nonlinear Rational Systems ⋆

Biqiang Mu [a,c], Er-Wei Bai [b], Wei Xing Zheng [c,*], and Quanmin Zhu [d]

[a] *Key Laboratory of Systems and Control, Institute of Systems Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China*

[b] *Department of Electrical and Computer Engineering, University of Iowa, Iowa City, IA 52242, USA*

[c] *School of Computing, Engineering and Mathematics, Western Sydney University, Sydney, NSW 2751, Australia*

[d] *Department of Engineering Design and Mathematics, University of the West of England, Bristol, BS16 1QY, UK*

Abstract

This paper considers identification of nonlinear rational systems defined as the ratio of two nonlinear functions of past inputs and outputs. Despite its long history, a globally consistent identification algorithm remains illusive. This paper proposes a globally convergent identification algorithm for such nonlinear rational systems. To the best of our knowledge, this is the first globally convergent algorithm for the nonlinear rational systems. The technique employed is a two-step estimator. Though two-step estimators are known to produce consistent nonlinear least squares estimates if a $\sqrt{N}$ consistent estimate can be determined in the first step, how to find such a $\sqrt{N}$ consistent estimate in the first step for nonlinear rational systems is nontrivial and is not answered by any two-step estimators. The technical contribution of the paper is to develop a globally consistent estimator for nonlinear rational systems in the first step. This is achieved by involving model transformation, bias analysis, noise variance estimation, and bias compensation in the paper. Two simulation examples and a practical example are provided to verify the good performance of the proposed two-step estimator.

Key words: Nonlinear rational systems, nonlinear least squares estimators, two-step estimators, $\sqrt{N}$-consistent estimators, Gauss-Newton algorithms

1 Introduction

System identification aims to build a mathematical model for a system from the measured data in some optimal way. If a system is linear or is well approximated by a linear system, then linear system models are a good choice. Thus, well-developed linear identification methods introduced for example in Ljung (1999); Söderström & Stoica (1989) are available to identify the system. On the other hand, if a system shows a strong nonlinear behavior, then nonlinear system models and the corresponding nonlinear identification algorithms become necessary.

A growing number of studies have demonstrated that the nonlinear autoregressive moving average with exogenous input (NARMAX) model (Chen & Billings, 1989; Haber & Unbehauen, 1990; Leontaritis & Billings, 1985) may provide a unified representation for a wide class of nonlinear systems that include several known nonlinear systems as special cases. However, the NARMAX representation is too general and inefficient in a variety of applications where a system does have some unique structures that are ignored by the general NARMAX representation. From an engineering point of view, the available structural information should be embedded into system models as well as identification algorithms. The nonlinear rational system is such a case. The study of nonlinear rational models has a long history and has been driven by practical applications and theoretical interests. On the application side, an early reported example was the catalytic dehydration of n-hexyl alcohol model (Box & Hunter, 1965)

$$y(k) = \frac{\theta_3\theta_1 u_1(k)}{1+\theta_1 u_1(k)+\theta_2 u_2(k)},$$

where y is the rate of reaction, u_1 the partial pressure of alcohol, u_2 the partial pressure of olefin, θ_1 absorption equilibrium constant of alcohol, θ_2 absorption equilibrium constant of olefin and θ_3 effective reaction rate constant. The purpose is to estimate $\theta_i,\ i=1,2,3$ from the measurements of y, u_1 and u_2. Interested readers can find quite a few real-world nonlinear rational systems in Bates & Watts (2007).

A unique feature of the nonlinear rational system is that both the numerator and denominator are linear combinations of known nonlinear functions of measurable variables with unknown coefficients or parameters. Thus, the system is nonlinear in terms of parameters in the denominator that makes identification nontrivial. Since that early paper, nonlinear rational systems have been widely used to model various biological phenomena in life science, for example, gene expression, metabolic networks, and enzyme catalyzed reactions within systems biology (Klipp et al., 2005) and chemical ki-

⋆ This work was supported in part by the President Fund of Academy of Mathematics and Systems Science, CAS under Grant 2015-hwyxqnrc-mbq, the National Key Basic Research Program of China (973 Program) under Grant 2014CB845301, the National Science Foundation under Grant CNS-1239509, the Australian Research Council under Grant DP120104986, and the National Nature Science Foundation of China under Grants 61273188 and 61603379.

* Corresponding author: Wei Xing Zheng. Tel: +61-2-4736 0608.
Email addresses: bqmu@amss.ac.cn (Biqiang Mu), er-wei-bai@uiowa.edu (Er-Wei Bai), w.zheng@westernsydney.edu.au (Wei Xing Zheng), quan.zhu@uwe.ac.uk (Quanmin Zhu).

netics of catalytic reactions in chemical engineering (Dimitrov & Kamenski, 1991; Kamenski & Dimitrov, 1993). They have also found applications in economic systems, physics, and engineering.

On the theoretical side, it was shown in Bartosiewicz (1987); Sontag (1979) that a nonlinear system possesses a realizable and bounded polynomial response if and only if the system is a rational model. Further, Bartosiewicz (1987) established that a smooth system may be immersed into a rational system if the observation field is a finitely generated extension of $\mathbb{R}$, and stated that rational systems could be simpler and more powerful than smooth systems. In addition, the existence of rational realizations of response maps was investigated in Němcová & van Schuppen (2009, 2010). It was shown that if a response map is realized by a rational system, then there also exists a minimal rational realization of the map (Němcová & van Schuppen, 2010). These evidences from the viewpoint of theory indicate that nonlinear rational systems can well approximate a wide range of nonlinear systems and actually provide a superior performance.

A number of identification algorithms have been proposed in the literature to estimate the unknown parameters in the nonlinear rational systems. They include prediction error estimator (Billings & Chen, 1989), extended least-squares estimator (Billings & Zhu, 1991), some variants of Newton-type methods (Dimitrov & Kamenski, 1991; Heiser & Parrish, 1989), back propagation parameter estimator (Zhu, 2003) and implicit least squares parameter estimator (Zhu, 2005). However, none of the estimators mentioned above are globally convergent. The main difficulties are: (a) Nonlinear rational systems could be transformed into a system which is linear in the parameters by multiplying the denominator on both sides. However, the resultant regressor is correlated with the noise and even with white noise, the resulting least squares estimate is biased. (b) The prediction error type objective function has many local minima since nonlinear rational systems are nonlinear in the parameters. Hence, various developed nonlinear optimization algorithms based on gradient descent are only locally convergent, and these results are summarized in a survey paper (Zhu et al., 2015).

As explained before, nonlinear rational systems can be converted into a system that is linear in parameters, but directly applying the ordinary least squares (OLS) estimator will lead to a biased estimate. Realizing this problem, this paper uses a two-step estimator to derive a convergent nonlinear least squares (NLS) estimator. The two-step estimator is known to produce a convergent NLS estimate (Gourieroux & Monfort, 1995) if the initial estimate provided by the first step is $\sqrt{N}$-consistent. This means that to produce a convergent NLS estimator it is sufficient to develop a $\sqrt{N}$-consistent estimator by the available data in the first step. In the paper, by a detailed analysis for the bias of the OLS estimator, it is shown that the bias can be removed if a consistent estimate for the variance of the noise is available. Consequently, the keys are a reliable estimate of the noise variance and the subsequent compensation of noise effects. Thus, this paper first provides a consistent estimate for the noise variance by seeking the minimum positive root of a polynomial constructed with the available data and demonstrates that the search is independent of the least squares estimator. Then by substituting the consistent estimate for the noise variance into the least squares estimator produces a globally $\sqrt{N}$-consistent estimator for nonlinear rational systems. Finally, the globally convergent NLS estimate for the nonlinear rational systems defined as the ratio of two nonlinear functions (not limited to polynomials) of past inputs and outputs is obtained by the second step of the two-step estimator.

The contribution of the paper is a globally convergent identification algorithm for nonlinear rational systems. To the best our knowledge, this is the first time that a globally convergent identification algorithm is proposed for nonlinear rational systems. We comment that the idea of two-step estimators is known in the literature (Gourieroux & Monfort, 1995). However, how to find a $\sqrt{N}$ consistent estimate at the first step for nonlinear national systems is nontrivial and is not answered by any two-step estimators. Thus, the technical contribution of the paper is to develop a $\sqrt{N}$ consistent estimate for nonlinear rational systems at the first step so that it can be used as an initial estimate in the second step. We also comment that bias compensation approaches have been used in compensation of linear least squares identification algorithms (Stoica & Söderström, 1982; Zheng, 1998; Zheng & Feng, 1995) and in errors-in-variable linear system identification (Söderström, 2007; Zheng, 2002). There are however several distinct differences between the technique proposed here and the bias compensation approach used in errors-in-variables systems and linear systems (Zheng, 1998, 2002; Zheng & Feng, 1995). First, obviously the systems considered above for the bias compensation approaches are linear systems and rational systems studied herein are nonlinear systems. Second, the bias compensation approaches need to construct a multi-dimensional auxiliary vector that satisfies certain properties. This is possible because the noise is independent of the inputs. For nonlinear rational systems, however, the resulting noise is a function of the inputs. It is not clear at this point if such an auxiliary vector exists for nonlinear rational systems. In the work reported here, instead of finding an auxiliary vector, it is shown that selection of an auxiliary vector is actually unnecessary and only a consistent estimate of the noise variance is required, which is one dimensional and therefore is much more efficient. The approach proposed in this paper may be considered as an extension of the technique for linear least squares compensation (Stoica & Söderström, 1982). However, because the technique of Stoica & Söderström (1982) is for linear systems and rational systems are nonlinear systems, both noise variance estimation and noise compensation for nonlinear rational systems become more involved and nontrivial. More technical details will be provided in the relevant Section 4.1.

The rest of the paper is organized as follows. Section 2 describes nonlinear rational systems under consideration and some assumptions on the system. Section 3 introduces briefly two-step estimators and their properties. A corrected least squares (CLS) estimator, which is obtained by model transformation, bias analysis, noise variance estimation, and bias compensation, is proposed in Section 4. The CLS is proved to converge to the true parameters in the global sense under proper conditions and its asymptotical normality is also established. Two simulation examples and a practical example are provided to verify the effectiveness of the proposed method in Section 5. Some concluding remarks are made in Section 6. Some technical details and the proofs are given in the Appendix.

2 Problem formulation

The nonlinear rational system under consideration is described as follows:

$$y(k)=\frac{\sum\limits_{i=1}^{q}\beta_i f_i(k)}{g_0(k)+\sum\limits_{j=1}^{p}\alpha_j g_j(k)}+\varepsilon(k),\quad 1\le k\le N, \quad (1)$$

where $f_i(k), g_j(k), 1\le i\le q, 0\le j\le p$ are *a priori* known functions of the delayed outputs and inputs $\{y(k-1),\cdots,y(k-n_y),u(k-1),\cdots,u(k-n_u)\}$ with positive integers n_y, n_u, $\theta^*\triangleq[\alpha_1,\cdots,\alpha_p,\beta_1,\cdots,\beta_q]^T$ is the unknown parameter vector that needs to be estimated, and $\varepsilon(k)$ is the observation noise. It is worth pointing out that the estimator for the nonlinear rational system (1) developed in the paper is applicable to the static case, i.e., $f_i(k), g_j(k), 1\le i\le q, 0\le j\le p$ are known functions of some exogenous variables. It is seen that the output $y(k)$ is linear in the parameters $\{\beta_1,\cdots,\beta_q\}$ but is nonlinear in the parameters $\{\alpha_1,\cdots,\alpha_p\}$, which is also the difficulty of identifying the rational system (1).

For ease of representation, define the denominator $a(k)\triangleq g_0(k)+\sum_{j=1}^{p}\alpha_j g_j(k)$, the numerator $b(k)\triangleq\sum_{i=1}^{q}\beta_i f_i(k)$, and the true output $v(k,\theta^*)=b(k)/a(k)$. Then the system (1) can be rewritten as

$$y(k)=\frac{b(k)}{a(k)}+\varepsilon(k)=v(k,\theta^*)+\varepsilon(k). \quad (2)$$

Let us give some remarks on the system (1). First, one assumes that the coefficient corresponding to the item $g_0(k)$ in (1) is 1 due to the identifiability reason of (1). This is always possible. Let the coefficient corresponding to the term $g_0(k)$ be denoted by α_0. Then, without loss of generality, one can assume $\alpha_0\neq 0$; otherwise, one can select any other item $g_j(k), 1\le j\le p$ with $\alpha_j\neq 0$ to play the role of the item $g_0(k)$ with α_0 since at least there is a parameter $\alpha_j\neq 0$ among $\{\alpha_1,\cdots,\alpha_p\}$. Next, dividing the numerator and the denominator by α_0 leads to the representation (1). An implicit assumption on the system (1) is that $a(k)\neq 0$.

In the following, let us give the conditions on the system, input, and noise for estimating the unknown parameters.

Assumption 1 *There is no undermodelling error for the system (1), i.e., the structure of the system including $a(k)$ and $b(k)$ is known and the noise $\varepsilon(k)$ is white. Further, $E|\varepsilon(k)|^{2\delta}<\infty$ for some $\delta>2$. Let $\sigma^2\triangleq\mathrm{Var}(\varepsilon(k))$.*

Assumption 2 *The sequence $\{x(k),k\ge 1\}$ with $x(k)\triangleq[y(k-1),\cdots,y(k-n_y),u(k-1),\cdots,u(k-n_u)]$ is asymptotically stationary in the wide sense and is an α-mixing process with mixing coefficients exponentially decaying to zero. Also, $E\|x(k)\|^{2\delta}<\infty$ for some $\delta>2$.*

We make a comment on the second assumption. Note the nonlinear rational model (1) can be regarded as a special case of nonlinear autoregressive systems with exogenous input (NARX)

$$y(k)=h(y(k-1),\cdots,y(k-n_y),u(k-1),\cdots,u(k-n_u))+\varepsilon(k), \quad (3)$$

where $h(\cdot)$ is a (n_y+n_u)-dimensional nonlinear function. By Lemma 1 in Zhao et al. (2013), the chain $x(k)\triangleq[y(k-1),\cdots,y(k-n_y),u(k-1),\cdots,u(k-n_u)]$ constructed by the outputs and inputs of NARX (3) is geometrically ergodic and is an α-mixing with mixing coefficients exponentially decaying to zero if the following conditions are satisfied: 1) both the input and the noise are sequences of independent and identically distributed (i.i.d.) random variables with zero mean and finite variance; 2) the system (3) satisfies certain stability condition. Note that the random vector $x(k)$ is geometrically ergodic, that is, the distribution of $x(k)$ tends to the invariant distribution at an exponential rate. This means that there is no essential difference between the stationary assumption and asymptotical stationary assumption on $\{x(k)\}$. For derivation simplicity, assume that the process $\{x(k)\}$ in Assumption 2 is stationary in the subsequent sections. These explanations indicate that Assumption 2 is not restrictive and in fact it is a standard assumption in the nonlinear system identification literature.

3 A standard two-step estimator

Prediction error methods are a natural idea for identifying the unknown parameter vector θ^* of the system (1) as that used in Billings & Chen (1989); Dimitrov & Kamenski (1991). Define the objective function with a prediction error form as

$$Q_N(\theta)=\frac{1}{N}\sum_{k=1}^{N}(y(k)-v(k,\theta))^2. \quad (4)$$

The vector minimizing (4) on a compact subset Θ of $\mathbb{R}^{p+q}$ containing θ^* is called the nonlinear least squares (NLS) estimator for θ^* based on the observations $\{u(k),y(k),1\le k\le N\}$ and is denoted by $\widehat{\theta}_N^{\mathrm{NLS}}$. Clearly, the gradient vector of $v(k,\theta)$ is given by

$$\frac{\partial v(k,\theta)}{\partial\theta}=\Big[-\frac{b(k)}{a^2(k)}\big[g_1(k),\cdots,g_p(k)\big],\ \frac{1}{a(k)}\big[f_1(k),\cdots,f_q(k)\big]\Big]^T.$$

We first give the conditions for the convergence of the NLS estimator.

Assumption 3 *(i) $Q(\theta)\triangleq E\big(v(k,\theta)-v(k,\theta^*)\big)^2$ has a unique minimum at $\theta=\theta^*$ in the compact set Θ.*

(ii) The true parameter vector θ^* is an interior point of Θ and the matrix $M(\theta^*)$ is nonsingular, where

$$M(\theta)\triangleq E\left(\frac{\partial v(k,\theta)}{\partial\theta}\frac{\partial v(k,\theta)}{\partial\theta^T}\right).$$

The NLS estimator $\widehat{\theta}_N^{\mathrm{NLS}}$ enjoys the following consistency and asymptotical normality, which can be derived by directly adopting the steps as what presented in Jennrich (1969).

Theorem 1 *(Jennrich, 1969, Theorem 7) Let $\widehat{\theta}_N^{\mathrm{NLS}}$ be the NLS estimator of (4). Under Assumptions 1, 2 and 3 (i), we have $\widehat{\theta}_N^{\mathrm{NLS}}\to\theta^*$ with probability one as N tends to infinity. Further, if Assumption 3 (ii) also holds, then*

$$\sqrt{N}(\widehat{\theta}_N^{\mathrm{NLS}}-\theta^*)\to\mathscr{N}(0,\sigma^2M^{-1}(\theta^*)) \text{ as } N\to\infty. \quad (5)$$

The NLS estimator involves a search for the solution of non-convex objective function (4), which may lead to that the gradient-based optimization algorithm converge to a local minimum if the starting point is outside the attraction neighborhood of the true value. Thus, the gradient-based optimization algorithm is generally applied to improve the precision when a good initial estimator, which is close to the true value, has obtained since $Q_N(\theta)$ is approximately convex in a small neighborhood of the true value. Additionally, it can be expected that the number of steps required for numerical convergence of the algorithm will be smaller by starting from an initial value close to θ^*.

Thus, the finding of the NLS estimator $\widehat{\theta}_N^{\rm NLS}$ is often done in two steps (Gourieroux & Monfort, 1995):

Step 1) Determine a consistent but not necessarily precise estimate.
Step 2) Use this preliminary estimate as an initial value for some algorithm that determines the NLS estimator.

In Step 2), the Gauss-Newton (GN) or other Newton-based algorithms are commonly used for improving the accuracy of the consistent estimator obtained in Step 1). The GN algorithm has the iterative form:

$$\theta_{n+1} = \theta_n + (J^T(\theta_n)J(\theta_n))^{-1}J^T(\theta_n)(Y - v(\theta_n)), \quad (6)$$

where the initial value θ_0 is the consistent estimator obtained in Step 1), $Y = [y(1), \ldots, y(N)]^T$,

$$v(\theta_n) = [v(1,\theta_n), \cdots, v(N,\theta_n)]^T,$$
$$J(\theta_n) = \left[\frac{\partial v(1,\theta_n)}{\partial\theta}, \cdots, \frac{\partial v(N,\theta_n)}{\partial\theta}\right]^T.$$

The standard two-step estimator given above has the following attractive property.

Theorem 2 *(Lehmann & Casella, 1998) Let $\widehat{\theta}_N^{\rm NLS}$ be the NLS estimator of (4). Suppose that $\widehat{\theta}_N$ is a $\sqrt{N}$-consistent estimator of θ^*, i.e., $\widehat{\theta}_N - \theta^* = O_p(1/\sqrt{N})$. Denote the one-step GN iteration of $\widehat{\theta}_N$ by $\theta_N^{\rm GN}$, i.e.,*

$$\theta_N^{\rm GN} = \widehat{\theta}_N + (J^T(\widehat{\theta}_N)J(\widehat{\theta}_N))^{-1}J^T(\widehat{\theta}_N)(Y - v(\widehat{\theta}_N)).$$

Thus under Assumptions 1–3 we have

$$\theta_N^{\rm GN} - \widehat{\theta}_N^{\rm NLS} = o_p(1/\sqrt{N}).$$

This means that $\theta_N^{\rm GN}$ has the same asymptotic property that $\widehat{\theta}_N^{\rm NLS}$ possesses.

It is seen that a key that the two-step estimator enjoys the desired property is to find a $\sqrt{N}$-consistent estimate of θ^* in Step 1). In fact, this is also the major difficulty for solving this kind of non-convex optimization problem.

4 A $\sqrt{N}$-consistent estimator: Corrected least squares

According to the two-step estimator and Theorem 2 introduced in Section 3, to obtain the NLS estimator $\widehat{\theta}_N^{\rm NLS}$ of (4) it is sufficient to find a $\sqrt{N}$-consistent estimator for θ^*. This section will develop a $\sqrt{N}$-consistent estimator for the unknown parameters of the nonlinear rational system (1) in the global sense that involves model transformation, bias analysis, noise variance estimation, and bias compensation. This is also the major goal and contribution of the paper.

4.1 Model transformation and bias analysis

Multiplying $a(k)$ on both sides (1) leads to

$$\begin{aligned} g_0(k)y(k) &= -\sum_{j=1}^{p}\alpha_j g_j(k)y(k) + \sum_{i=1}^{q}\beta_i f_i(k) + a(k)\varepsilon(k) \\ &= \psi(k)^T\theta^* + a(k)\varepsilon(k), \end{aligned} \quad (7)$$

where the regressor vector $\psi(k) \triangleq [-g_1(k)y(k), \cdots, -g_p(k)y(k), f_1(k), \cdots, f_q(k)]^T$. The resulting vector form is given by

$$Z_N = \Psi_N\theta^* + C_N, \quad (8)$$

where $Z_N \triangleq [g_0(1)y(1), \cdots, g_0(N)y(N)]^T$, $C_N \triangleq [a(1)\varepsilon(1), \cdots, a(N)\varepsilon(N)]^T$, $\Psi_N \triangleq [\psi(1), \cdots, \psi(N)]^T$. Clearly, the equation (7) is linear in all the parameters θ^* and all of the elements of $\psi(k)$ are available at time k. Thus, the ordinary least squares (OLS) estimator of (8) assumes the form of

$$\left(\frac{1}{N}\Psi_N^T\Psi_N\right)^{-1}\left(\frac{1}{N}\Psi_N^T Z_N\right). \quad (9)$$

However, the estimator (9) is a biased estimate for θ^* since the regressor $\psi(k)$ involves $y(k)$, which is correlated with the noise term $a(k)\varepsilon(k)$.

This problem is also encountered for identification of linear systems when the regressor vector is correlated with the noise, and the bias-eliminated least squares method (BELS) is the commonly adopted and effective method (Stoica & Söderström, 1982; Zheng, 1998; Zheng & Feng, 1995) to obtain a consistent estimate. In order to compare with linear cases, allow a little abuse of repeated usage of the notation. Consider the following linear case

$$y(k) = \psi(k)^T\theta^* + \varepsilon(k), \quad (10)$$

where the regressor vector $\psi(k)$ includes the delayed output and inputs, i.e., $\psi(k) = [\mathbf{y}(k)^T\mathbf{u}(k)^T]^T = [y(k-1), \cdots, y(k-n_y), u(k-1), \cdots, u(k-n_u)]^T$, θ^* is the unknown parameter that needs to be estimated, and $\mathbf{y}(k)$ is correlated with the noise $\varepsilon(k)$ but $\mathbf{u}(k)$ is uncorrelated with $\varepsilon(k)$. Under this setting, the BELS estimator of θ^* can be obtained via

$$\theta_{\rm BELS} = \theta_{\rm LS} - \left(E\psi(k)\psi(k)^T\right)^{-1}\begin{bmatrix} E\mathbf{y}(\mathbf{k})\varepsilon(k) \\ 0 \end{bmatrix}, \quad (11)$$

where $\theta_{\rm LS} = \left(E\psi(k)\psi(k)^T\right)^{-1}E\psi(k)y(k)$. So the key to the BELS method is to obtain a consistent estimate for the bias vector $E\mathbf{y}(\mathbf{k})\varepsilon(k)$, which is done usually by selecting some appropriate auxiliary vector $\zeta(k)$ satisfying $E\zeta(k)\varepsilon(k) = 0$ and $E\bar{\psi}(k)\bar{\psi}(k)^T > 0$ with $\bar{\psi}(k) = [\psi(k)^T\zeta(k)^T]^T$. The delayed inputs are selected to produce the consistent estimate for the bias in Zheng (1998) and the known regulator in closed-loop systems plays a similar role in obtaining the consistent bias in Zheng & Feng (1995). A unified framework for the BELS estimator can be referred to Jia et al. (2011). It is seen that this kind of BELS estimators depends on the selection of the auxiliary

vector $\zeta(k)$, which has a direct impact on the consistency and the accuracy of the estimator. Clearly, the regressor vector $\psi(k)$ defined in (7) is more complicated than its counterpart defined in (10) for the linear case, in which each element of $\psi(k)$ depends on all the delayed inputs and outputs and the noise term $a(k)\varepsilon(k)$ in (7) may depend on all the past inputs due to the existence of $a(k)$. This makes the selection of the auxiliary vector $\zeta(k)$ become nontrivial.

In order to avoid the indetermination for the selection of the auxiliary vector $\zeta(k)$ introduced above, the idea of the BELS estimator used in Stoica & Söderström (1982) is adopted here. In comparison with the BELS estimator having the form of (11), the advantages of the BELS estimator in Stoica & Söderström (1982) includes: 1) there is no need to select an appropriate auxiliary vector; 2) the only thing to be done for this BELS estimator is to develop a consistent estimate for a scalar quantity (the variance of the noise) instead of a multi-dimensional bias vector.

In the following, the idea for estimating the unknown parameter θ^* of the nonlinear rational model (1) is stated by referring to Stoica & Söderström (1982).

It follows from (1) and (7) that

$$\begin{aligned} g_0(k)y(k) &= -\sum_{j=1}^{p}\alpha_j g_j(k)y(k)+\sum_{i=1}^{q}\beta_i f_i(k)+a(k)\varepsilon(k) \\ &= -\sum_{j=1}^{p}\alpha_j g_j(k)v(k,\theta^*)+\sum_{i=1}^{q}\beta_i f_i(k)+g_0(k)\varepsilon(k), \\ &= \phi(k)^T\theta^*+g_0(k)\varepsilon(k), \end{aligned} \tag{12}$$

where

$$\begin{aligned} \phi(k) \triangleq \big[&-g_1(k)v(k,\theta^*),\cdots,-g_p(k)v(k,\theta^*), \\ &f_1(k),\cdots,f_q(k)\big]^T. \end{aligned} \tag{13}$$

The corresponding vector form is given by

$$Z_N = \Phi_N\theta^* + D_N, \tag{14}$$

where $Z_N \triangleq [g_0(1)y(1),\cdots,g_0(N)y(N)]^T$, $D_N \triangleq [g_0(1)\varepsilon(1),\cdots,g_0(N)\varepsilon(N)]^T$, $\Phi_N \triangleq [\phi(1),\cdots,\phi(N)]^T$. Clearly, the least squares estimator of (14) is obtained as

$$\left(\frac{1}{N}\Phi_N^T\Phi_N\right)^{-1}\left(\frac{1}{N}\Phi_N^T Z_N\right). \tag{15}$$

It is obvious that under the persistent excitation conditions on $\phi(k)$ that will be given in Assumption 4, the least squares estimator (15) is a consistent estimate for θ^* since $E\phi(k)g_0(k)\varepsilon(k)=0$. The problem is that $\phi(k)$ is unavailable. Let us define the persistent excitation condition.

Assumption 4 *There exists an integer $N_0 > 0$ such that $\frac{1}{N}\Phi_N^T\Phi_N > 0$ for all $N > N_0$.*

We provide a remark on the condition. Let us consider the noise-free case, that is,

$$y(k) = \frac{\sum_{i=1}^{q}\beta_i f_i(k)}{g_0(k)+\sum_{j=1}^{p}\alpha_j g_j(k)}.$$

An equivalent form of the above model is

$$\begin{aligned} g_0(k)y(k) &= -\sum_{j=1}^{p}\alpha_j g_j(k)v(k,\theta^*)+\sum_{i=1}^{q}\beta_i f_i(k) \\ &= \phi(k)^T\theta^*. \end{aligned}$$

In this simple case, Assumption 4 is exactly the persistent excitation condition for identifying nonlinear rational systems. Note that any linear system is just a special case. Thus, Assumption 4 can be explained as the persistent excitation condition for the system (1).

Note that Assumption 4 is a condition guaranteeing the global identifiability of the nonlinear rational system (1), while Assumption 3i) is the counterpart that ensures the local identifiability. We have the following lemma describing their connection.

Lemma 1 *Under Assumptions 1 and 2, Assumption 4 implies Assumption 3i).*

Let us proceed again to illustrate that a consistent estimate for θ^* can be obtained from (9) if the variance of the noise is a priori known. This begins with analyzing the difference between the least squares estimators (9) and (15). First, the matrices Ψ_N and Φ_N satisfy the relationship

$$\Psi_N = \Phi_N + H_N, \tag{16}$$

where $H_N \triangleq [h(1),\cdots,h(N)]^T$ and $h(k) \triangleq [-g_1(k)\varepsilon(k), \cdots,-g_p(k)\varepsilon(k),0,\cdots,0]^T$. It follows from the definition of C_N and D_N that

$$C_N = D_N - H_N\theta^*. \tag{17}$$

Similarly, by defining $G_N \triangleq [g_0(1),\cdots,g_0(N)]^T$, $A_N \triangleq [a(1),\cdots,a(N)]^T$, and $M_N \triangleq [m(1),\cdots,m(N)]^T$ with $m(k) \triangleq [-g_1(k),\cdots,-g_p(k),0,\cdots,0]^T$, we have

$$A_N = G_N - M_N\theta^*. \tag{18}$$

Then it follows from (16) that

$$\begin{aligned} &\frac{1}{N}\Psi_N^T\Psi_N \\ &= \frac{1}{N}\Phi_N^T\Phi_N+\frac{1}{N}H_N^TH_N+\frac{1}{N}\Phi_N^TH_N+\frac{1}{N}H_N^T\Phi_N \\ &= \frac{1}{N}\Phi_N^T\Phi_N+\frac{1}{N}H_N^TH_N+O_p\Big(\frac{1}{\sqrt{N}}\Big) \\ &= \frac{1}{N}\Phi_N^T\Phi_N+\sigma^2\Big(\frac{1}{N}M_N^TM_N\Big) \\ &\quad +\frac{1}{N}\Big(H_N^TH_N-\sigma^2M_N^TM_N\Big)+O_p\Big(\frac{1}{\sqrt{N}}\Big) \\ &= \frac{1}{N}\Phi_N^T\Phi_N+\sigma^2\Big(\frac{1}{N}M_N^TM_N\Big)+O_p\Big(\frac{1}{\sqrt{N}}\Big), \end{aligned} \tag{19}$$

where $\frac{1}{N}H_N^T\Phi_N = O_p(1/\sqrt{N})$ and $\frac{1}{N}\big(H_N^TH_N - \sigma^2M_N^TM_N\big) = O_p(1/\sqrt{N})$ by Theorem A1 in the Appendix since each element of $h(k)\phi(k)^T$ and $h(k)h(k)^T - \sigma^2m(k)m(k)^T$ is a martingale difference sequence and hence is an α-mixing with mixing coefficients exponentially

decaying to zero. Similarly, by (16) and (14), we have

$$\begin{aligned}
&\frac{1}{N}\Psi_N^T Z_N \\
&= \frac{1}{N}\Phi_N^T\Phi_N\theta^* + \frac{1}{N}H_N^T D_N + \frac{1}{N}\Phi_N^T D_N + \frac{1}{N}H_N^T\Phi_N\theta^* \\
&= \frac{1}{N}\Phi_N^T\Phi_N\theta^* + \frac{1}{N}H_N^T D_N + O_p\Big(\frac{1}{\sqrt{N}}\Big) \\
&= \frac{1}{N}\Phi_N^T\Phi_N\theta^* + \sigma^2\Big(\frac{1}{N}M_N^T G_N\Big) \\
&\quad + \frac{1}{N}\Big(H_N^T D_N - \sigma^2 M_N^T G_N\Big) + O_p\Big(\frac{1}{\sqrt{N}}\Big) \\
&= \frac{1}{N}\Phi_N^T\Phi_N\theta^* + \sigma^2\Big(\frac{1}{N}M_N^T G_N\Big) + O_p\Big(\frac{1}{\sqrt{N}}\Big). \qquad (20)
\end{aligned}$$

Thus, it follows from (19) and (20) that

$$\begin{aligned}
&\Big(\frac{1}{N}\Psi_N^T\Psi_N - \sigma^2\Big(\frac{1}{N}M_N^T M_N\Big)\Big)^{-1} \\
&\quad\times\Big(\frac{1}{N}\Psi_N^T Z_N - \sigma^2\Big(\frac{1}{N}M_N^T G_N\Big)\Big) \\
&= \Big(\frac{1}{N}\Phi_N^T\Phi_N + O_p\Big(\frac{1}{\sqrt{N}}\Big)\Big)^{-1} \\
&\quad\times\Big(\frac{1}{N}\Phi_N^T\Phi_N\theta^* + O_p\Big(\frac{1}{\sqrt{N}}\Big)\Big) \xrightarrow[N\to\infty]{} \theta^*.
\end{aligned}$$

Since Ψ_N, M_N, Z_N, and G_N are available by the input $u(k)$, the output $y(k)$, and the known nonlinear functions $g_i(k), f_j(k)$ for $0 \le i \le p, 1 \le j \le q$, a consistent estimate for the parameter vector θ^* is obtained if a consistent estimate for the variance σ^2 is produced in some way. This means that the key point of estimating θ^* is to independently derive a consistent estimate for σ^2.

A difference between the BELS estimator in Stoica & Söderström (1982) and its counterpart developed in this paper should now be pointed out. Because of nonlinearity presented in rational systems, both the denominator and the numerator of the least square estimator defined in (9) for the nonlinear rational model should be compensated, while only the denominator needs to be compensated for the linear case in Stoica & Söderström (1982). Further, estimation of noise variance becomes more involved, which will be discussed below.

4.2 Noise variance estimation

The results in Stoica & Söderström (1982) for linear systems implies that a consistent estimate for the variance of the noise can be obtained by solving some generalized eigenvalue problem. This motivates us to consider whether the idea given in Stoica & Söderström (1982) is applicable to the nonlinear rational system (1). The answer is positive, but the related procedure is much complicated and needs some necessary modifications. The detailed estimation procedure for the variance σ^2 of the noise is stated as follows. Define two matrices by the available data and information

$$J_N \triangleq \frac{1}{N}\begin{bmatrix}\Psi_N^T\\ Z_N^T\end{bmatrix}\big[\Psi_N, Z_N\big] = \frac{1}{N}\begin{bmatrix}\Psi_N^T\Psi_N & \Psi_N^T Z_N\\ \Psi_N Z_N^T & Z_N^T Z_N\end{bmatrix}, \quad (21)$$

$$\Delta_N \triangleq \frac{1}{N}\begin{bmatrix}M_N^T M_N & M_N^T G_N\\ G_N^T M_N & G_N^T G_N\end{bmatrix}. \quad (22)$$

Based on J_N and Δ_N, define a function $B_N(\cdot)$ over the variable λ_N as $B_N(\lambda_N) \triangleq J_N - \lambda_N\Delta_N$. Clearly, the function $\eta(\lambda_N)$ defined as $\eta(\lambda_N) \triangleq \det(B_N(\lambda_N))$ is a polynomial of power $p+1$ over λ_N. As a result, $\eta(\lambda_N) = 0$ has $p+1$ roots and denote all the roots by $\{\lambda_N(1), \cdots, \lambda_N(p+1)\}$. Thus, it will be shown below that the smallest root gives a consistent estimate $\widehat{\lambda}_N$ of the noise variance σ^2, i.e., the estimate for σ^2 can be defined by

$$\widehat{\lambda}_N = \min\{\lambda_N(j), j = 1, \cdots, p+1\}. \quad (23)$$

Note that the definition of Δ_N used here is different from its counterpart in Stoica & Söderström (1982) for linear systems. We have the following convergence conclusion on the estimate (23).

Lemma 2 *Under Assumptions 1, 2, and 4, the noise variance estimate (23) has an explicit solution*

$$\widehat{\lambda}_N = \sum_{k=1}^{N} a(k)^2\varepsilon(k)^2 \Big/ \sum_{k=1}^{N} a(k)^2, \quad (24)$$

which converges to the noise variance σ^2 with probability one and is asymptotically normal:

$$\sqrt{N}(\widehat{\lambda}_N - \sigma^2) \xrightarrow[N\to\infty]{} \mathcal{N}(0, Ea(k)^4\mathrm{Var}(\varepsilon(k)^2)/(Ea(k)^2)^2).$$

It follows from the proof of Lemma 2 given in the Appendix that all of the roots of $\eta(\lambda_N) = 0$ are greater than or equal to zero and the estimate (23) for the noise variance σ^2 is the smallest positive root of $\eta(\lambda_N) = 0$. Thus, the solution to (23) can be conveniently obtained by a root-seeking algorithm, for example, the function `fzero` in Matlab.

4.3 Corrected least squares estimator and asymptotical normality

Based on the explanation and analysis in Section 4.1, a consistent estimate for the unknown parameter vector θ^* can be obtained if a consistent estimate $\widehat{\lambda}_N$ for the noise variance σ^2 is provided. Thus, after deriving the consistent estimate (23) for the variance σ^2, the corrected least squares (CLS) estimator for θ^* can be defined by

$$\begin{aligned}
\widehat{\theta}_N^{\mathrm{CLS}} &= \Big(\frac{1}{N}\Psi_N^T\Psi_N - \widehat{\lambda}_N\Big(\frac{1}{N}M_N^T M_N\Big)\Big)^{-1} \\
&\quad\times\Big(\frac{1}{N}\Psi_N^T Z_N - \widehat{\lambda}_N\Big(\frac{1}{N}M_N^T G_N\Big)\Big), \qquad (25)
\end{aligned}$$

where $\widehat{\lambda}_N$ is given in (23). The CLS estimator defined by (25) has the following convergence and asymptotic normality.

Theorem 3 *Under Assumptions 1, 2, and 4, the CLS estimate $\widehat{\theta}_N^{\mathrm{CLS}}$ given in (25) converges to θ^* with probability one and is asymptotically normal:*

$$\sqrt{N}(\widehat{\theta}_N^{\mathrm{CLS}} - \theta^*) \xrightarrow[N\to\infty]{} \mathcal{N}(0, \Upsilon^{-1}W\Upsilon^{-1}),$$

where $\Upsilon \triangleq E\phi(k)\phi(k)^T$ *and* $W \triangleq Ew(k)w(k)^T$ *with*

$$w(k) \triangleq \phi(k)a(k)\varepsilon(k) + \Big(m(k)a(k) - a(k)^2\frac{Em(k)a(k)}{Ea(k)^2}\Big)\big(\varepsilon(k)^2 - \sigma^2\big).$$

Theorem 3 indicates that the CLS estimate $\widehat{\theta}_N^{\rm CLS}$ given in (25) is a $\sqrt{N}$-consistent estimator of θ^*. So, according to the two-step estimator introduced in Section 3, the NLS estimator $\widehat{\theta}_N^{\rm NLS}$ of the objective function (4) is obtained.

4.4 Recursive implementation of CLS

In this subsection, we present two recursive forms related to the CLS estimator defined in (25), which are useful for practical applications.

The First Form: Clearly, the CLS estimator can be rewritten as

$$\widehat{\theta}_N^{\rm CLS} = \big(\Psi_N^T\Psi_N - \widehat{\lambda}_N M_N^T M_N\big)^{-1}\big(\Psi_N^T Z_N - \widehat{\lambda}_N M_N^T G_N\big).$$

For notational simplicity, define

$$\begin{aligned}
R_N &\triangleq \big(\Psi_N^T\Psi_N - \widehat{\lambda}_N M_N^T M_N\big)^{-1}\\
S_N &\triangleq \psi(N)\psi(N)^T + \widehat{\lambda}_{N-1}M_{N-1}^T M_{N-1} - \widehat{\lambda}_N M_N^T M_N\\
V_N &\triangleq \Psi_N^T Z_N - \widehat{\lambda}_N M_N^T G_N\\
W_N &\triangleq \psi(N)g_0(N)y(N) + \widehat{\lambda}_{N-1}M_{N-1}^T G_{N-1} - \widehat{\lambda}_N M_N^T G_N
\end{aligned}$$

Then we have

$$\begin{aligned}
R_{N+1} &= \big(\Psi_N^T\Psi_N - \widehat{\lambda}_N M_N^T M_N + \psi(N+1)\psi(N+1)^T\\
&\quad + \widehat{\lambda}_N M_N^T M_N - \widehat{\lambda}_{N+1}M_{N+1}^T M_{N+1}\big)^{-1}\\
&= \big(R_N^{-1} + S_{N+1}\big)^{-1}\\
&= R_N - R_N\big(I + S_{N+1}R_N\big)^{-1}S_{N+1}R_N,
\end{aligned}$$

where the inverse of a sum of matrices (Henderson & Searle, 1981) is used. Similarly, one derives

$$\begin{aligned}
&V_{N+1} = \Psi_{N+1}^T Z_{N+1} - \widehat{\lambda}_{N+1}M_{N+1}^T G_{N+1}\\
&= \Psi_N^T Z_N - \widehat{\lambda}_N M_N^T G_N + \psi(N+1)g_0(N+1)y(N+1)\\
&\quad + \widehat{\lambda}_N M_N^T G_N - \widehat{\lambda}_{N+1}M_{N+1}^T G_{N+1}\\
&= V_N + W_{N+1}.
\end{aligned}$$

It follows that

$$\begin{aligned}
\widehat{\theta}_{N+1}^{\rm CLS} &= R_{N+1}V_{N+1} = R_{N+1}\big(V_N + W_{N+1}\big)\\
&= R_{N+1}\big(R_N^{-1}\widehat{\theta}_N^{\rm CLS} + W_{N+1}\big)\\
&= R_{N+1}\big(\big(R_{N+1}^{-1} - S_{N+1}\big)\widehat{\theta}_N^{\rm CLS} + W_{N+1}\big)\\
&= \widehat{\theta}_N^{\rm CLS} + R_{N+1}\big(W_{N+1} - S_{N+1}\widehat{\theta}_N^{\rm CLS}\big).
\end{aligned}$$

Thus, we obtain the following recursive algorithm for the CLS estimator:

$$\begin{aligned}
\widehat{\theta}_{N+1}^{\rm CLS} &= \widehat{\theta}_N^{\rm CLS} + R_{N+1}\big(W_{N+1} - S_{N+1}\widehat{\theta}_N^{\rm CLS}\big)\\
R_{N+1} &= R_N - R_N(I + S_{N+1}R_N)^{-1}S_{N+1}R_N\\
S_{N+1} &= \psi(N+1)\psi(N+1)^T - \big(\widehat{\lambda}_{N+1} - \widehat{\lambda}_N\big)\mathscr{M}_N\\
&\quad - \widehat{\lambda}_{N+1}m(N+1)m(N+1)^T\\
\mathscr{M}_N &= \mathscr{M}_{N-1} + m(N)m(N)^T\\
W_{N+1} &\triangleq \psi(N+1)g_0(N+1)y(N+1)\\
&\quad - \big(\widehat{\lambda}_{N+1} - \widehat{\lambda}_N\big)\mathscr{G}_N - \widehat{\lambda}_{N+1}m(N+1)g_0(N+1)\\
\mathscr{G}_N &= \mathscr{G}_{N-1} + m(N)g_0(N),
\end{aligned} \tag{26}$$

where the initial values are $\widehat{\theta}_0 = 0$, $R_0 = \gamma I > 0$, $\mathscr{M}_0 = 0$, and $\mathscr{G}_0 = 0$. This algorithm is an exactly recursive implementation of the CLS estimator defined in (25).

The Second Form: To obtain another recursive form of the CLS, let us start with the following estimator:

$$\widehat{\theta}_N = \big(\Psi_N^T\Psi_N - \sigma^2 M_N^T M_N\big)^{-1}\big(\Psi_N^T Z_N - \sigma^2 M_N^T G_N\big),$$

where the noise variance estimate in (25) is replaced by its true value. To allow an abuse of notation, define

$$\begin{aligned}
R_N &\triangleq \big(\Psi_N^T\Psi_N - \sigma^2 M_N^T M_N\big)^{-1},\\
S_N &\triangleq R_{N-1}^{-1} + \psi(N)\psi(N)^T,\\
V_N &\triangleq \Psi_N^T Z_N - \sigma^2 M_N^T G_N.
\end{aligned}$$

Thus, we have

$$\begin{aligned}
R_{N+1} &= \big(\Psi_N^T\Psi_N - \sigma^2 M_N^T M_N + \psi(N+1)\psi(N+1)^T\\
&\quad - \sigma^2 m(N+1)m_(N+1)^T\big)^{-1}\\
&= \big(R_N^{-1} + \psi(N+1)\psi(N+1)^T\\
&\quad - \sigma^2 m(N+1)m(N+1)^T\big)^{-1}\\
&= \big(S_{N+1} - \sigma^2 m(N+1)m(N+1)^T\big)^{-1}\\
&= S_{N+1}^{-1} + \frac{\sigma^2 S_{N+1}^{-1}m(N+1)m(N+1)^T S_{N+1}^{-1}}{1 - \sigma^2 m(N+1)^T S_{N+1}^{-1}m(N+1)}
\end{aligned}$$

and

$$S_{N+1}^{-1} = R_N - \frac{R_N\psi(N+1)\psi(N+1)^T R_N}{1 + \psi(N+1)^T R_N\psi(N+1)},$$

where the inverse of a sum of matrices (Henderson & Searle, 1981) is used. In a similar way, we obtain

$$\begin{aligned}
&V_{N+1} = \Psi_{N+1}^T Z_{N+1} - \sigma^2 M_{N+1}^T G_{N+1}\\
&= \Psi_N^T Z_N - \sigma^2 M_N^T G_N + \psi(N+1)g_0(N+1)y(N+1)\\
&\quad - \sigma^2 m(N+1)g_0(N+1)\\
&= V_N + \psi(N+1)g_0(N+1)y(N+1)\\
&\quad - \sigma^2 m(N+1)g_0(N+1).
\end{aligned}$$

It follows that

$$\begin{aligned}
&\widehat{\theta}_{N+1} = R_{N+1}V_{N+1}\\
&= R_{N+1}\big(V_N + \psi(N+1)g_0(N+1)y(N+1)\\
&\quad - \sigma^2 m(N+1)g_0(N+1)\big)\\
&= R_{N+1}\big(R_N^{-1}\widehat{\theta}_N + \psi(N+1)g_0(N+1)y(N+1)\\
&\quad - \sigma^2 m(N+1)g_0(N+1)\big)\\
&= R_{N+1}\big[\big(R_{N+1}^{-1} - \psi(N+1)\psi(N+1)^T\\
&\quad + \sigma^2 m(N+1)m(N+1)^T\big)\widehat{\theta}_N\\
&\quad + \psi(N+1)g_0(N+1)y(N+1)
\end{aligned}$$

$$
\begin{aligned}
&\quad - \sigma^2 m(N+1) g_0(N+1)\big] \\
&= \widehat{\theta}_N + R_{N+1}\big[\psi(N+1)\big(g_0(N+1)y(N+1) \\
&\quad - \psi(N+1)^T\widehat{\theta}_N\big) - \sigma^2 m(N+1)\big(g_0(N+1) \\
&\quad - m(N+1)^T\widehat{\theta}_N\big)\big].
\end{aligned}
$$

Thus, we get another recursive algorithm:

$$
\begin{aligned}
&\widehat{\theta}_{N+1} = \widehat{\theta}_N + R_{N+1}\big[\psi(N+1) \\
&\quad \times \big(g_0(N+1)y(N+1) - \psi(N+1)^T\widehat{\theta}_N\big) \\
&-\widehat{\lambda}_{N+1} m(N+1)\big(g_0(N+1) - m(N+1)^T\widehat{\theta}_N\big)\big], \\
&R_{N+1} = S_{N+1}^{-1} + \frac{\widehat{\lambda}_{N+1} S_{N+1}^{-1} m(N+1) m(N+1)^T S_{N+1}^{-1}}{1-\widehat{\lambda}_{N+1} m(N+1)^T S_{N+1}^{-1} m(N+1)}, \\
&S_{N+1}^{-1} = R_N - \frac{R_N \psi(N+1)\psi(N+1)^T R_N}{1+\psi(N+1)^T R_N \psi(N+1)},
\end{aligned}
\tag{27}
$$

where the initial values are $\widehat{\theta}_0 = 0$ and $R_0 = \gamma I > 0$. In addition to online updating the estimate at current time based on its immediate past estimate and the currently received data, an attractive merit of the recursive algorithm (27) is that it avoids the explicit matrix inverse calculation in (26), even though it is not an exactly recursive implementation of the CLS (25).

Since it is difficult to derive a recursive scheme for the noise variance estimate $\widehat{\lambda}_N$ in (23), the needed value $\widehat{\lambda}_N$ in (26) and (27) is directly calculated by (23). Actually, this will not greatly increase the computational complexity since the noise variance estimation (23) is achieved by a root-seeking algorithm for a one-dimensional polynomial of power $p+1$.

5 Numerical examples

Example 1 This example is used to illustrate that the objective function (4) has many local minima. Thus, the Newton-based optimization algorithms may converge to a local minimum if the initial value is outside the attraction region of the true value. Consider a nonlinear rational system

$$
y(k) = \frac{3u(k-1)y(k-1)}{1-0.8y(k-1)} + \varepsilon(k), \tag{28}
$$

where $u(k), y(k)$ are the input and output, respectively, $g_0(k) = 1$, $g_1(k) = y(k-1)$, $f_1(k) = u(k-1)y(k-1)$, and the true parameter vector is $\theta^* = [-0.8, 3]^T$. The input $\{u_k\}$ is a sequence of i.i.d. random variables uniformly generated from the interval $[0, 0.6]$. The noise $\{\varepsilon_k\}$ is a sequence of i.i.d. uniform random variables in the interval $[-1, 1]$. The sample size is $N = 1000$. For ease of presentation, the opposite $-Q_N(\theta)$ of the objective function (4) is plotted on its two parameters in a large region $\{(\alpha, \beta) \in [-5, 5] \times [-5, 5]\}$. Fig. 1 shows that the objective function corresponding to the system (28) has many local minima. This phenomenon still exists even if the region of the parameters is narrowed down to $\{(\alpha, \beta) \in [-0.9, -0.7] \times [2.9, 3.1]\}$ including the true value $(-0.8, 3)$ (See Fig. 2). This means that the gradient-based optimization algorithms for solving (4) may not work well.

To compare the performance of the two-step estimator proposed in the paper with other estimators for identifying the unknown parameters in (28), we first introduce all the estimators involved here. They are the corrected least squares estimator (CLS) defined by (25) in Section 4, the two-step

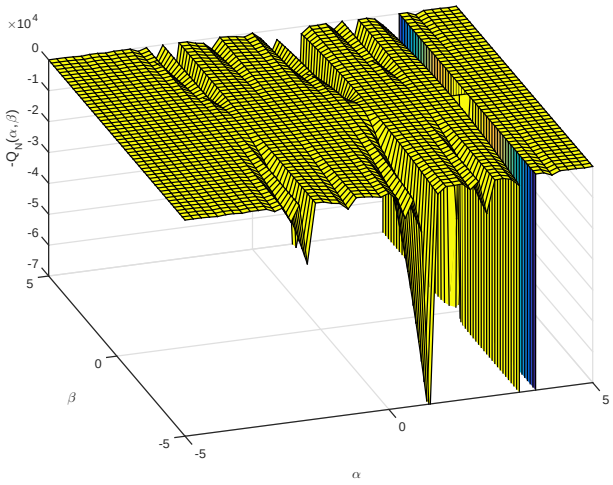

Fig. 1. The three-dimensional plot of $-Q_N(\alpha, \beta)$ corresponding to the system (28)

estimator proposed in Section 3, i.e., the Gauss-Newton algorithm with the CLS estimator serving as its initial value (CLS+GN) defined by (6), the ordinary least squares (OLS) estimator defined by (9), the Gauss-Newton algorithm with the OLS estimator serving as its initial value (OLS+GN) defined by (6), the simulated annealing algorithm (SA) for minimizing the objective function (4) which is implemented by the function `simulannealbnd` in Matlab and the initial value is set as the OLS estimator, and the genetic algorithm (GA) for minimizing the objective function (4) which is implemented by the function `ga` in Matlab and does not require to provide an initial value, respectively. To evaluate the performance of all the estimators given above, the fitness measure (FM) (Ljung, 2012)

$$
\mathrm{FM} = 100\left(1 - \frac{\|\widehat{\theta}_N - \theta^*\|_2}{\|\theta^* - \bar{\theta}^*\|_2}\right)
$$

is used, where $\widehat{\theta}_N$ represents the resulting estimate for θ^* and $\bar{\theta}^*$ is the arithmetic average of the elements of θ^*. The following results are based on 100 Monte-Carlo simulations, where the mean and the standard deviation of the signal-to-noises ratios (SNRs) calculated by the 100 runs are 14.17 dB and 0.74 dB.

To investigate the performance of all the estimators introduced above for this example, the distribution of the FM for these estimators is listed, where Table 1 gives the resulting quantiles at 10%, 25%, 50%, 75%, and 90%, respectively, and Figure 3 shows the box plot. One can first conclude from these distributions that the commonly used global optimization algorithms including the SA and GA estimators do not perform well since the true value can hardly be found by them. On the other hand, the consistent CLS estimator is superior to the biased OLS estimator. More importantly, the CLS estimator is significantly improved by the Gauss-Newton algorithm, while the OLS estimator is greatly deteriorated by the Gauss-Newton algorithm since the 10% quantile of the FM for the CLS+GN estimator is 98.10 but the 90% quantile of the FM for the OLS+GN estimator is -0.53. This also shows that the CLS estimator almost lies in the attraction neighborhood of the Gauss-Newton algorithm, but the OLS estimator does not enjoy this advantage.

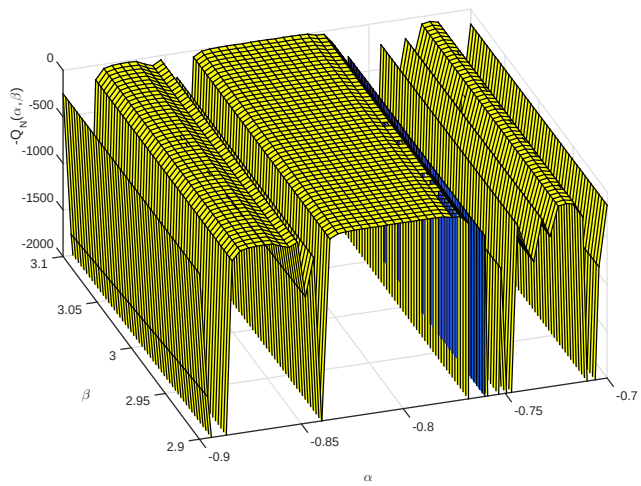

Fig. 2. The three-dimensional plot of $-Q_N(\alpha, \beta)$ in a narrower region

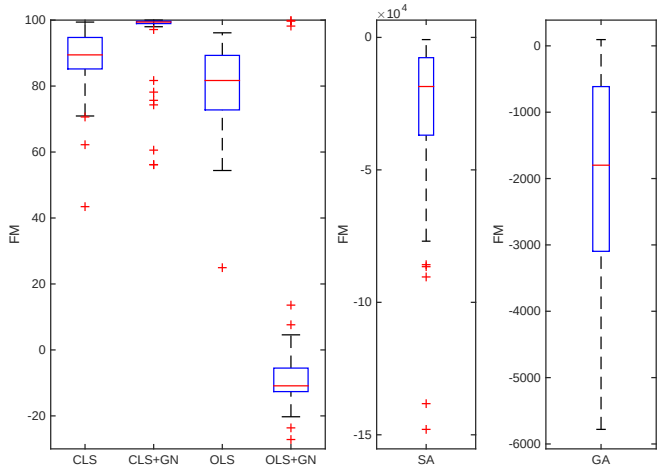

Fig. 3. The box plot of the FM for the CLS, CLS+GN, OLS, OLS+GN, SA, and GA estimators. In order to show the complete distributions of these estimators, they are displayed by some proper but different scales, respectively.

Table 1
The quantiles of all the estimators

Methods	10%	25%	50%	75%	90%
CLS	79.05	85.19	89.46	94.74	97.51
CLS+GN	98.10	98.94	99.41	99.68	99.83
OLS	68.19	72.77	81.69	89.90	93.46
OLS+GN	−15.89	−12.63	−10.89	−5.51	−0.53
SA	−62102	−36899	−18534	−7647	−3741
GA	−3899	−3097	−1799	−613.95	−13.86

Finally, a comparison of the computational complexity between the CLS estimator and the resulting two kinds of recursive CLS estimators given in Section 4.4 is also provided by considering the time spent of these estimators. For convenience, let us denote the exact recursive implementation of the CLS estimator (the first form) by RCLS and the modified recursive implementation of the CLS estimator (the second form) by MRCLS, respectively. The hardware used for this

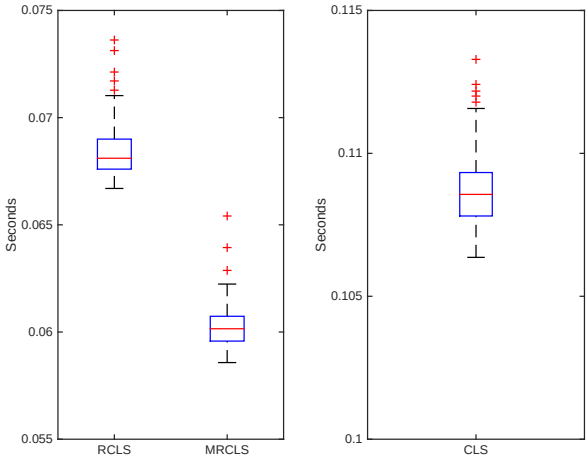

Fig. 4. The box plot of the computational complexity of the RCLS, MRCLS, and CLS estimators.

comparison includes a 3.5 GHz Intel Core i5 CPU and an 8 GB RAM while the software platform is Matlab 2014b running under OS X 10.10 operation system. Note that both the RCLS, MRCLS, and CLS estimators involve the same root-seeking step (23). Thus, it is fair to exclude the time spent by the root-seeking process for comparing the computational complexity of the CLS estimator and its recursive forms. Figure 4 plots the distributions of the three kinds of estimators. It is observed that the RCLS and MRCLS estimators can save about 37% and 45% computational time, respectively, in comparison with the CLS estimator based on their medians. Also, the standard deviation of the spent time of the MRCLS is smaller than that of the RCLS and CLS estimators.

Example 2 Consider a nonlinear rational system

$$y(k) = \frac{2y(k-1)y(k-2) + 3u(k-1)}{1 + 0.5y(k-1)^2 + u(k-1)^2} + \varepsilon(k), \quad (29)$$

where $u(k), y(k)$ are the input and output, respectively, $g_0(k) = 1$, $g_1(k) = y(k-1)^2$, $g_2(k) = u(k-1)^2$, $f_1(k) = y(k-1)y(k-2)$, $f_2(k) = u(k-1)$, and the true parameter vector is $\theta^* = [0.5, 1, 2, 3]^T$. The input $\{u_k\}$ is a sequence of i.i.d. random variables uniformly generated from the interval $[-1, 1]$. The noise $\{\varepsilon_k\}$ is a sequence of i.i.d. Gaussian random variables: $\mathcal{N}(0, \sigma^2)$.

In order to reflect the impact of the noise intensity to the estimation accuracy of θ^*, we conduct estimation under different noise levels, where the variance σ^2 of the noise is selected as 0.4^2 and 0.8^2, respectively, and the corresponding SNRs are 16.28 dB and 11.85 dB, respectively. Tables 2–3 list the estimate of the CLS and CLS+GN estimators for the sample sizes $N = 500, 2000, 5000, 10000$ under the SNRs introduced above and averaging over 100 random runs. The values in the parentheses are the resulting standard deviations. Figures 5–6 plot the distribution of the resulting fitness measures of the parameter estimation shown by box plots for the different cases described above. It is seen from these figures that the Gauss-Newton algorithm greatly improves the estimation accuracy if it starts with an estimate given by the CLS estimator.

Example 3 (A practical example) The book by Bates &

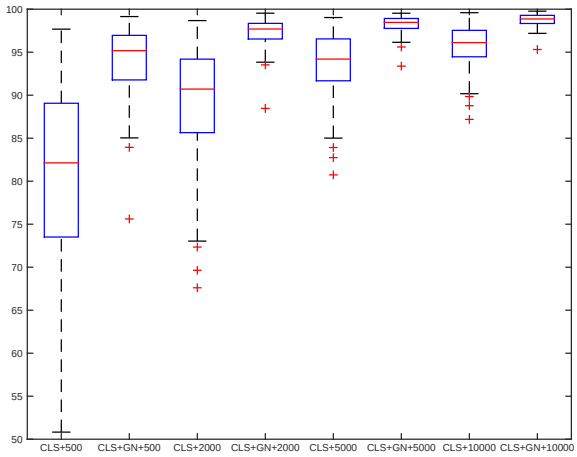

Fig. 5. Box plots of the fitness measure based on 100 random runs at $\mathrm{SNR} = 16.28$. The horizontal axis represents the adopted estimation method and the sample size while the vertical axis is the resulting fitness measure, e.g.,"CLS+500" means the estimate is obtained by the CLS when $N = 500$.

Table 2
Parameter estimation at $\mathrm{SNR} = 16.28$

True values	500	2000	5000	10000
		CLS		
0.5000	0.5137	0.5021	0.5007	0.4975
	(0.0736)	(0.0414)	(0.0264)	(0.0189)
1.0000	1.0153	1.0071	1.0072	0.9977
	(0.1745)	(0.0802)	(0.0494)	(0.0348)
2.0000	2.0320	2.0081	2.0043	1.9948
	(0.1692)	(0.0946)	(0.0615)	(0.0441)
3.0000	3.0363	3.0147	3.0136	2.9948
	(0.3757)	(0.2088)	(0.1153)	(0.0766)
FM	79.6555	89.1149	93.5884	95.6284
	(12.1309)	(6.7713)	(3.7371)	(2.5173)
		CLS+GN		
0.5000	0.5001	0.5005	0.4998	0.4994
	(0.0178)	(0.0077)	(0.0050)	(0.0038)
1.0000	0.9990	1.0041	1.0014	0.9998
	(0.0814)	(0.0321)	(0.0218)	(0.0150)
2.0000	1.9977	2.0023	2.0004	1.9991
	(0.0440)	(0.0170)	(0.0108)	(0.0085)
3.0000	2.9969	3.0024	3.0005	2.9993
	(0.1021)	(0.0474)	(0.0302)	(0.0211)
FM	93.9975	97.3350	98.2250	98.7448
	(3.9983)	(1.6554)	(0.9830)	(0.6804)

Watts (2007) contains quite a few real-world rational system examples. We consider the Michaelis-Menten model because of published experimental data. The model is for enzyme kinetics that relate the initial "velocity" y of an enzymatic reaction to the substrate concentration u through the equation

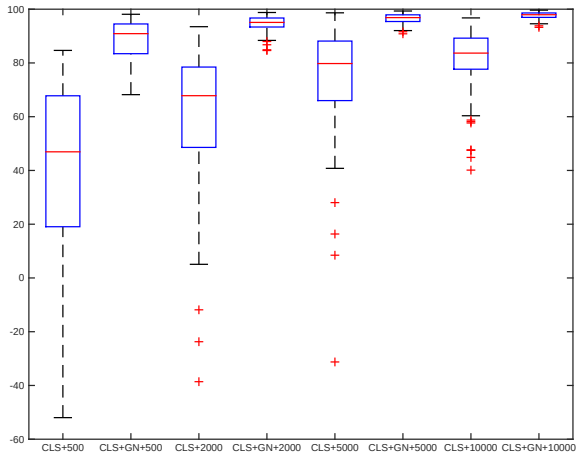

Fig. 6. Box plots of the fitness measure based on 100 random runs at $\mathrm{SNR} = 11.85$. The meanings of the horizontal and vertical axes are the same as those in Fig. 5.

Table 3
Parameter estimation at $\mathrm{SNR} = 11.85$

True values	500	2000	5000	10000
		CLS		
0.5000	0.6365	0.5770	0.5339	0.5278
	(0.2431)	(0.1719)	(0.1271)	(0.0892)
1.0000	1.0837	1.0392	1.0115	1.0102
	(0.4363)	(0.2570)	(0.1605)	(0.1097)
2.0000	2.3607	2.2059	2.0977	2.0832
	(0.5882)	(0.4064)	(0.2925)	(0.2102)
3.0000	3.2944	3.1321	3.0925	3.0458
	(1.0744)	(0.6637)	(0.4861)	(0.3138)
FM	37.6706	62.0914	74.8470	81.6704
	(38.5321)	(25.1922)	(20.0134)	(11.3487)
		CLS+GN		
0.5000	0.5052	0.5007	0.5008	0.5001
	(0.0283)	(0.0145)	(0.0091)	(0.0063)
1.0000	1.0003	0.9989	1.0016	1.0022
	(0.1243)	(0.0626)	(0.0416)	(0.0274)
2.0000	2.0072	2.0004	2.0022	2.0004
	(0.0706)	(0.0335)	(0.0218)	(0.0150)
3.0000	2.9728	2.9780	2.9939	2.9945
	(0.2001)	(0.0895)	(0.0578)	(0.0429)
FM	89.0443	94.6207	96.5450	97.5779
	(6.8676)	(2.8375)	(1.8260)	(1.3805)

$$y(k) = f(k, u, \theta) = \frac{\beta}{1 + \alpha/u(k)},$$

where β is the ultimate velocity parameter and α is the half-velocity parameter (Bates & Watts, 2007, page 33), $g_0(k) = 1, g_1(k) = 1/u(k)$, and $f_1(k) = 1$. The experiment was conducted once with enzyme treated with Puromycin and the number of the observations was 12, $\{y(k), u(k)\}_1^{12}$. The experimental data were obtained by Treloar (1974) and

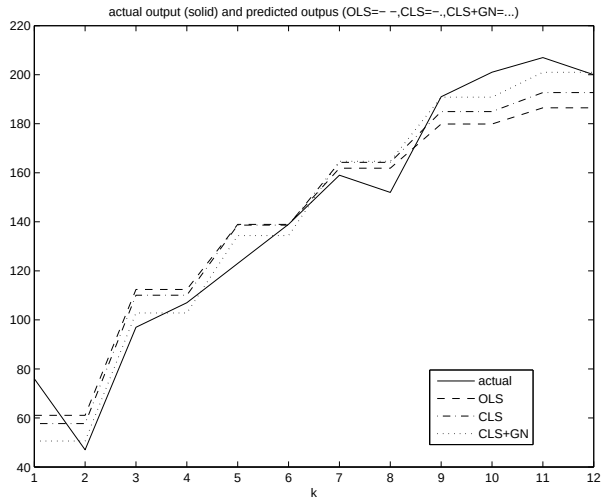

Fig. 7. Actual output (solid) $y(k)$ and the predicted outputs $\hat{y}(k)$'s by OLS (dashed), CLS (dash-dotted) and CLS+GN (dotted) estimators.

Table 4
Parameter estimation and prediction error of the relevant estimators for the practical example

Methods	$(\hat{\alpha}, \hat{\beta})$	Ave. Prediction error
OLS	(0.0435,193.8677)	13.59
CLS	(0.0498,201.4230)	11.56
CLS+GN	(0.0641,212.6837)	9.98

were reprinted on the page 269 of Bates & Watts (2007). It is important to note that because it is a real-world model, there is no "true value" or nobody knows the "true value" of (α, β). Let $(\hat{\alpha}, \hat{\beta})$ be an estimate of the "true value" (α, β) and $\hat{y}(k) = \frac{\hat{\beta}}{1+\hat{\alpha}/u(k)}$ be the predicted output based on the estimates. The quality of estimates can be measured by the averaged output errors $\sqrt{\frac{1}{12}\sum_{k=1}^{12}(y(k) - \hat{y}(k))^2}$. The estimates for the unknown parameters (α, β) by the OLS, CLS, and CLS+GN estimators as well as the corresponding average output errors are calculated for this example, as illustrated in Table 4 and Figure 7. It is easily seen that the CLS estimator performs better than the OLS estimator and moreover the CLS+GN estimator further improves the CLS. Note there are only 12 observations.

6 Conclusion

The nonlinear least squares estimator for the unknown parameters of nonlinear rational systems has been developed via a standard two-step estimator in the paper. The developed NLS estimator consists of two successive steps: 1) one provides a good initial estimator for the unknown parameter; 2) one obtains the NLS estimator for the unknown parameters by using the Gauss-Newton algorithm with the estimate obtained in Step 1 serving as the initial value. In Step 1, the CLS estimator has been proposed by model transformation, bias analysis, noise variance estimation, and bias compensation and has been proved to be a $\sqrt{N}$-consistent estimator of the unknown parameters in the global sense under some conditions. To the best of our knowledge, this is the first time that a globally consistent estimate has been provided for nonlinear rational systems. Therefore, in theory it can be guaranteed that the NLS estimator can be obtained by one-step Gauss-Newton iteration with the $\sqrt{N}$-consistent CLS estimate serving as the initial value. There exist several directions that need to be explored for future research, for example, colored noises, multi-input multi-output systems and so on.

Appendix

6.1 Auxiliary results on random sequences

For the process $\{X_k,\ k = 1, 2, \cdots\}$, denote the σ-algebra generated by $\{X_s,\ 1 \le i \le s \le j\}$ by $\mathscr{F}_i^j$. Define

$$\alpha(k) \stackrel{\triangle}{=} \sup_{n, A\in\mathscr{F}_1^n, B\in\mathscr{F}_{n+k}^{\infty}} |P(A)P(B) - P(AB)|.$$

The process $\{X_k\}$ is called α-mixing if $\alpha(k) \xrightarrow[k\to\infty]{} 0$, and the numbers $\alpha(k)$ are called the mixing coefficients of the random process $\{X_k\}$. For analyzing the convergence of the CLS estimator proposed in the paper, we need the results on the central limit theorem of α-mixing process.

Theorem A1 *(Davidson, 1994) Let $\{X_k\}$ be a stationary sequence with $EX_k = 0$ and $E|X_k|^\delta < \infty$ for some $\delta > 2$. Suppose $\{X_k\}$ is an α-mixing with exponentially decaying mixing coefficients $\alpha(k)$. Then*

$$\frac{E\big(\sum_{k=1}^N X_k\big)^2}{N} \to EX_1^2 + 2\sum_{k=2}^{\infty} E(X_1X_k) \stackrel{\triangle}{=} \chi^2.$$

Further, if $\chi^2 > 0$, then $\frac{1}{\sqrt{N}}\sum_{k=1}^N X_k \to \mathscr{N}(0, \chi^2)$. Also, there holds $\sum_{k=1}^N X_k/\sqrt{N} = O_p(1)$.

The following result is also useful for proving the asymptotic normality of the CLS estimator.

Theorem A2 *(Söderström & Stoica, 1989, Lemma B.4) Let $\{x_k\}$ be a sequence of random vectors that converges in distribution to a Gaussian vector $\mathscr{N}(0, P)$. Let $\{A_k\}$ be a sequence of random square matrices that converges in probability to nonsingular matrix A. Define $z_k = A_kx_k$. Then z_k converges in distribution to $\mathscr{N}(0, APA^T)$.*

6.2 Main proofs

Proof of Theorem 2. Since $\widehat{\theta}_N^{\rm NLS}$ is the minimum of (4) and is also a stationary point of (4) under Assumption 3, $\widehat{\theta}_N^{\rm NLS}$ satisfies the first order condition

$$\begin{aligned}&\frac{1}{N}J^T(\widehat{\theta}_N^{\rm NLS})(Y - v(\widehat{\theta}_N^{\rm NLS}))\\ &\quad=\frac{1}{N}\sum_{k=1}^N \frac{\partial v(k, \widehat{\theta}_N^{\rm NLS})}{\partial\theta}(y(k) - v(k, \widehat{\theta}_N^{\rm NLS})) = 0.\end{aligned}$$

Applying the Taylor expansion around $\widehat{\theta}_N$ derives

$$\begin{aligned}&-\frac{1}{N}J^T(\widehat{\theta}_N)(Y - v(\widehat{\theta}_N))\\ &\quad= -B(\widehat{\theta}_N^{\rm NLS} - \widehat{\theta}_N) + o_p(1/\sqrt{N}) = O_p(1/\sqrt{N}),\end{aligned} \quad (30)$$

where the assertion $\widehat{\theta}_N^{\rm NLS}-\widehat{\theta}_N=O_p(1/\sqrt{N})$ is used since both $\widehat{\theta}_N^{\rm NLS}$ and $\widehat{\theta}_N$ are $\sqrt{N}$-consistent estimators and

$$B=\frac{1}{N}J^T(\widehat{\theta}_N)J(\widehat{\theta}_N)-\frac{1}{N}\sum_{k=1}^{N}\frac{\partial^2 v(k,\widehat{\theta}_N)}{\partial\theta\partial\theta^T}(y(k)-v(k,\widehat{\theta}_N)).$$

Since $\varepsilon(k)$ is uncorrelated with $v(k,\theta^*)$ and $\widehat{\theta}_N$ is $\sqrt{N}$-consistent, B can be simplified as $B=\frac{1}{N}J^T(\widehat{\theta}_N)J(\widehat{\theta}_N)+o_p(1)$. It follows from (30) that

$$\begin{aligned}\widehat{\theta}_N^{\rm NLS}-\widehat{\theta}_N=&\Big(\frac{1}{N}J^T(\widehat{\theta}_N)J(\widehat{\theta}_N)\Big)^{-1}\Big(\frac{1}{N}J^T(\widehat{\theta}_N)(Y-v(\widehat{\theta}_N))\Big)\\&+\Big(\frac{1}{N}J^T(\widehat{\theta}_N)J(\widehat{\theta}_N)\Big)^{-1}o_p(1/\sqrt{N}).\end{aligned}$$

This means that $\widehat{\theta}_N^{\rm NLS}-\widehat{\theta}_N^{\rm GN}=o_p(1/\sqrt{N})$. ∎

Proof of Lemma 1. Suppose that Assumption 3i) does not hold. Then there exists another parameter $\widetilde{\theta}\neq\theta^*$ and $\widetilde{\theta}\in\Theta$ such that $Q(\theta)$ arrives at its minimum at $\theta=\widetilde{\theta}$. Obviously, the minimum of $Q(\theta)$ is zero. This means that $E\big(v(k,\widetilde{\theta})-v(k,\theta^*)\big)^2=0$ and further we have $v(k,\widetilde{\theta})=v(k,\theta^*)$ almost surely (a.s.). For simplicity of derivation, one assume that $g_0(k)=1$. It follow from (12) that

$$y(k)=\phi(k,\theta^*)^T\theta^*+\varepsilon(k),$$

which is a pseudo-linear regression type of (2). This implies that $\phi(k,\theta^*)^T\theta^*=\phi(k,\widetilde{\theta})^T\widetilde{\theta}$ a.s. based on the fact $v(k,\widetilde{\theta})=v(k,\theta^*)$. On the other hand, the expression of $\phi(k,\theta)$ defined in (13) shows that $\phi(k,\theta)$ depends on the parameter θ by the way of $v(k,\theta)$, so we also have $\phi(k,\theta^*)=\phi(k,\widetilde{\theta})$ a.s. This derives that $\phi(k,\theta^*)^T\theta^*=\phi(k,\theta^*)^T\widetilde{\theta}$ a.s. Multiplying $\phi(k,\theta^*)$ on both sides from left and taking expectation give

$$E\big(\phi(k,\theta^*)\phi(k,\theta^*)^T\big)\theta^*=E\big(\phi(k,\theta^*)\phi(k,\theta^*)^T\big)\widetilde{\theta}$$

This yields that $E\big(\phi(k,\theta^*)\phi(k,\theta^*)^T\big)$ is singular since $\widetilde{\theta}\neq\theta^*$ and hence Assumption 4 is violated since we have

$$\frac{1}{N}\Phi_N^T\Phi_N\to E\big(\phi(k,\theta^*)\phi(k,\theta^*)^T\big)\text{ as }N\to\infty$$

by applying the stationary of $\phi(k,\theta^*)$. This completes the proof. ∎

Proof of Lemma 2. Now, one plans to prove (24) by two steps.

Step 1: To show that $s_N\triangleq\sum_{k=1}^{N}a(k)^2\varepsilon(k)^2\Big/\sum_{k=1}^{N}a(k)^2$ is a root of the polynomial $\eta(\lambda_N)=0$. Applying the identities $C_N=Z_N-\Psi_N\theta^*$ and $A_N=G_N-M_N\theta^*$ leads to

$$\begin{aligned}&L(\lambda_N)\triangleq\begin{bmatrix}\theta^{*T}&-1\end{bmatrix}B_N(\lambda_N)\begin{bmatrix}\theta^*\\-1\end{bmatrix}\\=&\frac{1}{N}\begin{bmatrix}\theta^{*T}&-1\end{bmatrix}\begin{bmatrix}\Psi_N^T\Psi_N&\Psi_N^TZ_N\\\Psi_NZ_N^T&Z_N^TZ_N\end{bmatrix}\begin{bmatrix}\theta^*\\-1\end{bmatrix}\\&-\lambda_N\frac{1}{N}\begin{bmatrix}\theta^{*T}&-1\end{bmatrix}\begin{bmatrix}M_N^TM_N&M_N^TG_N\\G_N^TM_N&G_N^TG_N\end{bmatrix}\begin{bmatrix}\theta^*\\-1\end{bmatrix}\\=&\frac{1}{N}\begin{bmatrix}\theta^{*T}&-1\end{bmatrix}\begin{bmatrix}\Psi_N^T\Psi_N\theta^*-\Psi_N^TZ_N\\\Psi_NZ_N^T\theta^*-Z_N^TZ_N\end{bmatrix}\\&-\lambda_N\frac{1}{N}\begin{bmatrix}\theta^{*T}&-1\end{bmatrix}\begin{bmatrix}M_N^TM_N\theta^*-M_N^TG_N\\G_N^TM_N\theta^*-G_N^TG_N\end{bmatrix}\\=&\frac{1}{N}\begin{bmatrix}\theta^{*T}&-1\end{bmatrix}\begin{bmatrix}-\Psi_N^TC_N\\-Z_N^TC_N\end{bmatrix}+\lambda_N\frac{1}{N}\begin{bmatrix}\theta^{*T}&-1\end{bmatrix}\begin{bmatrix}M_N^TA_N\\G_N^TA_N\end{bmatrix}\\=&-\frac{1}{N}\big(\theta^{*T}\Psi_N^T-Z_N^T\big)C_N+\lambda_N\frac{1}{N}\big(\theta^{*T}M_N^T-G_N^T\big)A_N\\=&\frac{1}{N}C_N^TC_N-\lambda_N\frac{1}{N}A_N^TA_N.\end{aligned}$$

Clearly, $L(s_N)=0$ and $L(\lambda_N)>0$ if $\lambda_N<s_N$. Since $[\theta^{*T}\ -1]^T$ is nonzero, we must have $\det(B_N(\widehat{\lambda}_N))=0$. This implies that s_N is a root of $\eta(\lambda_N)=0$.

Step 2: To show that s_N is the smallest root of $\eta(\lambda_N)=0$. To this end, note that $B_N(\lambda_N)$ is symmetric and J_N is semi-positive definite. Let $w=[w_1^T,w_2]^T$ with $w_1\in\mathbb{R}^{p+q}$ and $w_2\in\mathbb{R}$ be any nonzero column vector linearly independent of $[\theta^{*T}\ -1]^T$. In order to reach the desired conclusion, it remains to show that $w^TB_N(\lambda_N)w>0$ for $\lambda_N\leq s_N$ since we have shown that $L(\lambda_N)>0$ if $\lambda_N<s_N$ in Step 1. Note that

$$\frac{1}{N}\begin{bmatrix}\Phi_N^T\Phi_N&\Phi_N^T\Phi_N\theta^*\\\theta^{*T}\Phi_N^T\Phi_N&\theta^{*T}\Phi_N^T\Phi_N\theta^*\end{bmatrix}\begin{bmatrix}\theta^*\\-1\end{bmatrix}=0$$

and Assumption 4, then we obtain

$$\operatorname{rank}\left\{\frac{1}{N}\begin{bmatrix}\Phi_N^T\Phi_N&\Phi_N^T\Phi_N\theta^*\\\theta^{*T}\Phi_N^T\Phi_N&\theta^{*T}\Phi_N^T\Phi_N\theta^*\end{bmatrix}\right\}=p+q.$$

This means that

$$w^T\frac{1}{N}\begin{bmatrix}\Phi_N^T\Phi_N&\Phi_N^T\Phi_N\theta^*\\\theta^{*T}\Phi_N^T\Phi_N&\theta^{*T}\Phi_N^T\Phi_N\theta^*\end{bmatrix}w>0$$

since w is linearly independent of $[\theta^{*T}\ -1]^T$. Note that

$$\begin{aligned}J_N=&\frac{1}{N}\begin{bmatrix}\Phi_N^T\Phi_N&\Phi_N^T\Phi_N\theta^*\\\theta^{*T}\Phi_N^T\Phi_N&\theta^{*T}\Phi_N^T\Phi_N\theta^*\end{bmatrix}\\&+\frac{1}{N}\begin{bmatrix}H_N^TH_N&H_N^TD_N\\D_N^TH_N&D_N^TD_N\end{bmatrix}\\&+\frac{1}{N}\begin{bmatrix}\Phi_N^TH_N+H_N^T\Phi_N&\Phi_N^TD_N+H_N^T\Phi_N\theta^*\\D_N^T\Phi_N+\theta^{*T}\Phi_N^TH_N&\theta^{*T}\Phi_N^TD_N+D_N^T\Phi_N\theta^*\end{bmatrix}.\end{aligned}$$

Clearly, we have

$$\begin{aligned}B_N(\lambda_N)=&\frac{1}{N}\begin{bmatrix}\Phi_N^T\Phi_N&\Phi_N^T\Phi_N\theta^*\\\theta^{*T}\Phi_N^T\Phi_N&\theta^{*T}\Phi_N^T\Phi_N\theta^*\end{bmatrix}\\&+\frac{1}{N}\left(\begin{bmatrix}H_N^TH_N&H_N^TD_N\\D_N^TH_N&D_N^TD_N\end{bmatrix}-\sigma^2\begin{bmatrix}M_N^TM_N&M_N^TG_N\\G_N^TM_N&G_N^TG_N\end{bmatrix}\right)\\&+\big((\sigma^2-s_N)+(s_N-\lambda_N)\big)\frac{1}{N}\begin{bmatrix}M_N^TM_N&M_N^TG_N\\G_N^TM_N&G_N^TG_N\end{bmatrix}\end{aligned}$$

$$+\frac{1}{N}\begin{bmatrix}\Phi_N^T H_N + H_N^T\Phi_N & \Phi_N^T D_N + H_N^T\Phi_N\theta^* \\ D_N^T\Phi_N + \theta^{*T}\Phi_N^T H_N & \theta^{*T}\Phi_N^T D_N + D_N^T\Phi_N\theta^*\end{bmatrix}.$$

In view of Theorem A1, we arrive at

$$\begin{aligned}
&\frac{1}{N}\left(\begin{bmatrix}H_N^T H_N & H_N^T D_N \\ D_N^T H_N & D_N^T D_N\end{bmatrix} - \sigma^2\begin{bmatrix}M_N^T M_N & M_N^T G_N \\ G_N^T M_N & G_N^T G_N\end{bmatrix}\right) \\
&\quad = O_p\Big(\frac{1}{\sqrt{N}}\Big), \\
&\frac{1}{N}\begin{bmatrix}\Phi_N^T H_N + H_N^T\Phi_N & \Phi_N^T D_N + H_N^T\Phi_N\theta^* \\ D_N^T\Phi_N + \theta^{*T}\Phi_N^T H_N & \theta^{*T}\Phi_N^T D_N + D_N^T\Phi_N\theta^*\end{bmatrix} \\
&\quad = O_p\Big(\frac{1}{\sqrt{N}}\Big), \\
&\frac{1}{N}\begin{bmatrix}M_N^T M_N & M_N^T G_N \\ G_N^T M_N & G_N^T G_N\end{bmatrix} = O_p(1), \\
&s_N - \sigma^2 = O_p\Big(\frac{1}{\sqrt{N}}\Big).
\end{aligned}$$

It follows that

$$\begin{aligned}
B_N(\lambda_N) = &\frac{1}{N}\begin{bmatrix}\Phi_N^T\Phi_N & \Phi_N^T\Phi_N\theta^* \\ \theta^{*T}\Phi_N^T\Phi_N & \theta^{*T}\Phi_N^T\Phi_N\theta^*\end{bmatrix} \\
&+ (s_N - \lambda_N)\frac{1}{N}\begin{bmatrix}M_N^T M_N & M_N^T G_N \\ G_N^T M_N & G_N^T G_N\end{bmatrix} + O_p\Big(\frac{1}{\sqrt{N}}\Big)
\end{aligned}$$

and hence

$$\begin{aligned}
w^T B_N(\lambda_N) w \geq & w^T\frac{1}{N}\begin{bmatrix}\Phi_N^T\Phi_N & \Phi_N^T\Phi_N\theta^* \\ \theta^{*T}\Phi_N^T\Phi_N & \theta^{*T}\Phi_N^T\Phi_N\theta^*\end{bmatrix} w \\
&+ O_p\Big(\frac{1}{\sqrt{N}}\Big) > 0
\end{aligned}$$

if $\lambda_N \leq s_N$.

Up to now, we have proved that s_N is the smallest root of $\eta(\lambda_N) = 0$. So, according to the definition of (23), we have $\widehat{\lambda}_N = s_N$, i.e., $\widehat{\lambda}_N = \sum_{k=1}^N a(k)^2\varepsilon(k)^2 \Big/ \sum_{k=1}^N a(k)^2$. This means

$$\widehat{\lambda}_N - \sigma^2 = \frac{\frac{1}{N}\sum_{k=1}^N a(k)^2\big(\varepsilon(k)^2 - \sigma^2\big)}{\frac{1}{N}\sum_{k=1}^N a(k)^2}.$$

Define the σ-algebra $\mathscr{F}_k \stackrel{\triangle}{=} \sigma\{\varepsilon_i, 1 \leq i \leq k\}$. Thus, the denominator $a(k)$ is measurable with respect to $\mathscr{F}_{k-1}$ and then we have

$$\begin{aligned}
&E\Big(a(k)^2\big(\varepsilon(k)^2 - \sigma^2\big) \mid \mathscr{F}_{k-1}\Big) \\
&\qquad = a(k)^2 E(\varepsilon(k)^2 - \sigma^2|\mathscr{F}_{k-1}) = 0.
\end{aligned}$$

This means that $\{a(k)^2\big(\varepsilon(k)^2 - \sigma^2\big), \mathscr{F}_k\}$ is a martingale difference sequence and hence is an α-mixing with mixing coefficients exponentially decaying to zero and

$$\begin{aligned}
&Ea(1)^4\big(\varepsilon(1)^2 - \sigma^2\big)^2 + 2\sum_{k=2}^{\infty} Ea(1)^2\big(\varepsilon(1)^2 - \sigma^2\big) \\
&\qquad \times a(k)^2\big(\varepsilon(k)^2 - \sigma^2\big) = Ea(1)^4\big(\varepsilon(1)^2 - \sigma^2\big)^2.
\end{aligned}$$

Under Assumption 2, $\{a(k)^2\}$ is an α-mixing with mixing coefficients exponentially decaying to zero. By Theorem A1, we have

$$\begin{aligned}
&\frac{1}{N}\sum_{k=1}^N a(k)^2\big(\varepsilon(k)^2 - \sigma^2\big) \to \mathscr{N}\big(0, Ea(k)^4\mathrm{Var}(\varepsilon(k)^2)\big), \\
&\frac{1}{N}\sum_{k=1}^N a(k)^2 = Ea(k)^2 + O_p\Big(\frac{1}{\sqrt{N}}\Big).
\end{aligned}$$

Finally, applying Theorem A2 yields

$$\sqrt{N}(\widehat{\lambda}_N - \sigma^2) \xrightarrow[N\to\infty]{} \mathscr{N}(0, Ea(k)^4\mathrm{Var}(\varepsilon(k)^2)/(Ea(k)^2)^2),$$

thereby completing the proof. ■

Proof of Theorem 3. Note that $Z_N = \Psi_N\theta^* + C_N$. Thus, we have

$$\begin{aligned}
&\frac{1}{N}\Psi_N^T Z_N - \widehat{\lambda}_N\Big(\frac{1}{N}M_N^T G_N\Big) \\
=&\frac{1}{N}\Psi_N^T\Psi_N\theta^* + \frac{1}{N}\Psi_N^T C_N - \widehat{\lambda}_N\Big(\frac{1}{N}M_N^T G_N\Big) \\
=&\Big(\frac{1}{N}\Psi_N^T\Psi_N - \widehat{\lambda}_N\Big(\frac{1}{N}M_N^T M_N\Big)\Big)\theta^* + \frac{1}{N}\Psi_N^T C_N \\
&+ \widehat{\lambda}_N\Big(\frac{1}{N}M_N^T M_N\Big)\theta^* - \widehat{\lambda}_N\Big(\frac{1}{N}M_N^T G_N\Big).
\end{aligned}$$

Further, using the identities $\Psi_N = \Phi_N + H_N$ and $A_N = G_N - M_N\theta^*$ derives

$$\begin{aligned}
&\sqrt{N}(\widehat{\theta}_N - \theta^*) = \left(\frac{1}{N}\Psi_N^T\Psi_N - \widehat{\lambda}_N\Big(\frac{1}{N}M_N^T M_N\Big)\right)^{-1} \\
&\times\left(\frac{1}{\sqrt{N}}\Psi_N^T C_N + \widehat{\lambda}_N\Big(\frac{1}{\sqrt{N}}M_N^T M_N\Big)\theta^* - \widehat{\lambda}_N\Big(\frac{1}{\sqrt{N}}M_N^T G_N\Big)\right) \\
=&\left(\frac{1}{N}\Psi_N^T\Psi_N - \widehat{\lambda}_N\Big(\frac{1}{N}M_N^T M_N\Big)\right)^{-1} \\
&\times\left(\frac{1}{\sqrt{N}}\Phi_N^T C_N + \frac{1}{\sqrt{N}}H_N^T C_N - \widehat{\lambda}_N\Big(\frac{1}{\sqrt{N}}M_N^T A_N\Big)\right).
\end{aligned}$$

Clearly, we have

$$\begin{aligned}
&\frac{1}{N}\Psi_N^T\Psi_N - \widehat{\lambda}_N\Big(\frac{1}{N}M_N^T M_N\Big) \\
=&\frac{1}{N}\Phi_N^T\Phi_N + \frac{1}{N}H_N^T H_N + \frac{1}{N}\Phi_N^T H_N + \frac{1}{N}H_N^T\Phi_N \\
&- \widehat{\lambda}_N\Big(\frac{1}{N}M_N^T M_N\Big) \\
=&\frac{1}{N}\Phi_N^T\Phi_N + \frac{1}{N}\Big(H_N^T H_N - \sigma^2 M_N^T M_N\Big)
\end{aligned}$$

$$+(\sigma^2-\widehat{\lambda}_N)\Big(\frac{1}{N}M_N^TM_N\Big)+\frac{1}{N}\Phi_N^TH_N+\frac{1}{N}H_N^T\Phi_N.$$

From Theorem A1 it follows that

$$\begin{aligned}
&\frac{1}{N}\Big(H_N^TH_N-\sigma^2M_N^TM_N\Big)=O_p\Big(\frac{1}{\sqrt{N}}\Big),\\
&(\sigma^2-\widehat{\lambda}_N)\Big(\frac{1}{N}M_N^TM_N\Big)=O_p\Big(\frac{1}{\sqrt{N}}\Big),\\
&\frac{1}{N}\Phi_N^TH_N=O_p\Big(\frac{1}{\sqrt{N}}\Big),\ \frac{1}{N}H_N^T\Phi_N=O_p\Big(\frac{1}{\sqrt{N}}\Big),\\
&\frac{1}{N}\Phi_N^T\Phi_N=E\phi(k)\phi(k)^T+O_p\Big(\frac{1}{\sqrt{N}}\Big).
\end{aligned}$$

This implies that

$$\begin{aligned}
&\frac{1}{N}\Psi_N^T\Psi_N-\widehat{\lambda}_N\Big(\frac{1}{N}M_N^TM_N\Big)\\
&\qquad=E\phi(k)\phi(k)^T+O_p\Big(\frac{1}{\sqrt{N}}\Big). \qquad (31)
\end{aligned}$$

By a straightforward calculation, we have

$$\begin{aligned}
&\frac{1}{\sqrt{N}}\Phi_N^TC_N+\frac{1}{\sqrt{N}}H_N^TC_N-\widehat{\lambda}_N\Big(\frac{1}{\sqrt{N}}M_N^TA_N\Big)\\
=&\frac{1}{\sqrt{N}}\Phi_N^TC_N+\frac{1}{\sqrt{N}}\Big(H_N^TC_N-\sigma^2M_N^TA_N\Big)\\
&+(\sigma^2-\widehat{\lambda}_N)\frac{1}{\sqrt{N}}\Big(M_N^TA_N-EM_N^TA_N\Big)\\
&+(\sigma^2-\widehat{\lambda}_N)\frac{1}{\sqrt{N}}EM_N^TA_N.
\end{aligned}$$

Theorem A1 derives

$$\begin{aligned}
&\widehat{\lambda}_N-\sigma^2=O_p\Big(\frac{1}{\sqrt{N}}\Big),\\
&\frac{1}{\sqrt{N}}\Big(M_N^TA_N-EM_N^TA_N\Big)=O_p(1).
\end{aligned}$$

This means that

$$(\sigma^2-\widehat{\lambda}_N)\frac{1}{\sqrt{N}}\Big(M_N^TA_N-EM_N^TA_N\Big)=O_p\Big(\frac{1}{\sqrt{N}}\Big).$$

It follows that

$$\begin{aligned}
&\frac{1}{\sqrt{N}}\Phi_N^TC_N+\frac{1}{\sqrt{N}}H_N^TC_N-\widehat{\lambda}_N\Big(\frac{1}{\sqrt{N}}M_N^TA_N\Big)\\
=&\frac{1}{\sqrt{N}}\Phi_N^TC_N+\frac{1}{\sqrt{N}}\Big(H_N^TC_N-\sigma^2M_N^TA_N\Big)\\
&+(\sigma^2-\widehat{\lambda}_N)\frac{1}{\sqrt{N}}EM_N^TA_N+O_p\Big(\frac{1}{\sqrt{N}}\Big)\\
=&\frac{1}{\sqrt{N}}\Big(\Phi_N^TC_N+\big(H_N^TC_N-\sigma^2M_N^TA_N\big)\\
&+\big(EM_N^TA_N\big)\big(\sigma^2A_N^TA_N-C_N^TC_N\big)/A_N^TA_N\Big)\\
&+O_p\Big(\frac{1}{\sqrt{N}}\Big).
\end{aligned}$$

Note that

$$\begin{aligned}
&\frac{1}{\sqrt{N}}\left(\Big(\frac{EM_N^TA_N}{A_N^TA_N}-\frac{EM_N^TA_N}{EA_N^TA_N}\Big)\big(\sigma^2A_N^TA_N-C_N^TC_N\big)\right)\\
=&\Big(\frac{EM_N^TA_N}{A_N^TA_N}-\frac{EM_N^TA_N}{EA_N^TA_N}\Big)\left(\frac{1}{\sqrt{N}}\big(\sigma^2A_N^TA_N-C_N^TC_N\big)\right)\\
=&\left(\frac{\big(\frac{1}{N}(EA_N^TA_N-A_N^TA_N)\big)EM_N^TA_N}{\big(\frac{1}{N}A_N^TA_N\big)EA_N^TA_N}\right)\\
&\times\left(\frac{1}{\sqrt{N}}\big(\sigma^2A_N^TA_N-C_N^TC_N\big)\right).
\end{aligned}$$

By Theorem A1, we get

$$\begin{aligned}
&\frac{1}{N}\big(EA_N^TA_N-A_N^TA_N\big)=O_p\Big(\frac{1}{\sqrt{N}}\Big),\\
&\frac{1}{N}A_N^TA_N=Ea(k)a(k)^T+O_p\Big(\frac{1}{\sqrt{N}}\Big)=O_p(1),\\
&\frac{EM_N^TA_N}{EA_N^TA_N}=\frac{\frac{1}{N}EM_N^TA_N}{\frac{1}{N}EA_N^TA_N}=O_p(1),\\
&\frac{1}{\sqrt{N}}\big(\sigma^2A_N^TA_N-C_N^TC_N\big)=O_p(1).
\end{aligned}$$

This entails that

$$\begin{aligned}
&\frac{1}{\sqrt{N}}\left(\frac{EM_N^TA_N}{A_N^TA_N}\big(\sigma^2A_N^TA_N-C_N^TC_N\big)\right)\\
=&\frac{1}{\sqrt{N}}\left(\frac{EM_N^TA_N}{EA_N^TA_N}\big(\sigma^2A_N^TA_N-C_N^TC_N\big)\right)+O_p\Big(\frac{1}{\sqrt{N}}\Big).
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&\frac{1}{\sqrt{N}}\Phi_N^TC_N+\frac{1}{\sqrt{N}}H_N^TC_N-\widehat{\lambda}_N\Big(\frac{1}{\sqrt{N}}M_N^TA_N\Big)\\
=&\frac{1}{\sqrt{N}}\Big(\Phi_N^TC_N+\big(H_N^TC_N-\sigma^2M_N^TA_N\big)\\
&+\big(\sigma^2A_N^TA_N-C_N^TC_N\big)\frac{EM_N^TA_N}{EA_N^TA_N}\Big)+O_p\Big(\frac{1}{\sqrt{N}}\Big)\\
=&\frac{1}{\sqrt{N}}\Big(\sum_{k=1}^N\phi(k)a(k)\varepsilon(k)+\sum_{k=1}^Nm(k)a(k)\big(\varepsilon(k)^2-\sigma^2\big)\\
&+\frac{Em(k)a(k)}{Ea(k)^2}\sum_{k=1}^Na(k)^2\big(\sigma^2-\varepsilon(k)^2\big)\Big)+O_p\Big(\frac{1}{\sqrt{N}}\Big)\\
=&\frac{1}{\sqrt{N}}\Big(\sum_{k=1}^N\phi(k)a(k)\varepsilon(k)+\Big(m(k)a(k)-a(k)^2\frac{Em(k)a(k)}{Ea(k)^2}\Big)\\
&\times\big(\varepsilon(k)^2-\sigma^2\big)\Big)+O_p\Big(\frac{1}{\sqrt{N}}\Big).
\end{aligned}$$

Define the random vector

$$\begin{aligned}
w(k)\triangleq&\ \phi(k)a(k)\varepsilon(k)\\
&+\Big(m(k)a(k)-a(k)^2\frac{Em(k)a(k)}{Ea(k)^2}\Big)\big(\varepsilon(k)^2-\sigma^2\big)
\end{aligned}$$

$$= \begin{bmatrix} -g_1(k)v(k,\theta^*)a(k)\varepsilon(k) \\ \vdots \\ -g_p(k)v(k,\theta^*)a(k)\varepsilon(k) \\ f_1(k)a(k)\varepsilon(k) \\ \vdots \\ f_q(k)a(k)\varepsilon(k) \end{bmatrix} + \varpi(k)$$

$$= \begin{bmatrix} -g_1(k)b(k)\varepsilon(k) \\ \vdots \\ -g_p(k)b(k)\varepsilon(k) \\ f_1(k)a(k)\varepsilon(k) \\ \vdots \\ f_q(k)a(k)\varepsilon(k) \end{bmatrix} + \varpi(k)$$

where

$$\varpi(k) = \begin{bmatrix} \Big(a(k)^2 \frac{Eg_1(k)a(k)}{Ea(k)^2} - g_1(k)a(k)\Big)\big(\varepsilon(k)^2 - \sigma^2\big) \\ \vdots \\ \Big(a(k)^2 \frac{Eg_p(k)a(k)}{Ea(k)^2} - g_p(k)a(k)\Big)\big(\varepsilon(k)^2 - \sigma^2\big) \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

and the σ-algebra $\mathscr{F}_k \stackrel{\triangle}{=} \sigma\{\varepsilon_i, 1 \le i \le k\}$. Thus, the functions $g_i(k), f_j(k), 0 \le i \le p, 1 \le j \le q$ are measurable with respect to $\mathscr{F}_{k-1}$ and then we have

$$\begin{aligned} &E(w(k)|\mathscr{F}_{k-1}) = E\Big(\phi(k)a(k)\varepsilon(k) \\ &+ \Big(m(k)a(k) - a(k)^2\frac{Em(k)a(k)}{Ea(k)^2}\Big)\big(\varepsilon(k)^2 - \sigma^2\big)\Big|\mathscr{F}_{k-1}\Big) \\ &= \phi(k)a(k)E(\varepsilon(k)|\mathscr{F}_{k-1}) \\ &+ \Big(m(k)a(k) - a(k)^2\frac{Em(k)a(k)}{Ea(k)^2}\Big)E\big(\varepsilon(k)^2 - \sigma^2\big|\mathscr{F}_{k-1}\big) \\ &= 0. \end{aligned}$$

This means that $\{w(k), \mathscr{F}_k\}$ is a martingale difference sequence. Definitely, $\{w(k), \mathscr{F}_k\}$ is also an α-mixing with mixing coefficients exponentially decaying to zero and

$$Ew(1)w(1)^T + 2\sum_{k=2}^{\infty} Ew(1)w(k)^T = Ew(1)w(1)^T.$$

Applying Theorem A1 gives rise to

$$\frac{1}{\sqrt{N}}\Phi_N^T C_N + \frac{1}{\sqrt{N}}H_N^T C_N - \widehat{\lambda}_N\Big(\frac{1}{\sqrt{N}}M_N^T A_N\Big) \to \mathscr{N}(0, W).$$

Combining it with (31) and applying Theorem A2 complete the proof. ■

Acknowledgments

The authors would like to thank the Associate Editor and the anonymous reviewers for their constructive and helpful comments and suggestions to improve the quality of this paper.

References

Bartosiewicz, Z. (1987). Rational systems and observation fields. *Systems & Control Letters*, *9*, 379–386.

Bates, D. M., & Watts, D. G. (2007). *Nonlinear Regression Analysis and Its Applications*. Hoboken, NJ: John Wiley & Sons, Inc.

Billings, S. A., & Chen, S. (1989). Identification of non-linear rational systems using a prediction-error estimation algorithm. *International Journal of Systems Science*, *20*, 467–494.

Billings, S. A., & Zhu, Q. M. (1991). Rational model identification using an extended least-squares algorithm. *International Journal of Control*, *54*, 529–546.

Box, G. E. P., & Hunter, W. G. (1965). The experimental study of physical mechanisms. *Technometrics*, *7*, 23–42.

Chen, S., & Billings, S. A. (1989). Representation of nonlinear systems: The NARMAX model. *International Journal of Control*, *49*, 1013–1032.

Davidson, J. (1994). *Stochastic Limit Theory: An Introduction for Econometricians*. New York: Oxford University Press.

Dimitrov, S. D., & Kamenski, D. I. (1991). A parameter estimation method for rational functions. *Computers & Chemical Engineering*, *15*, 657–662.

Gourieroux, C., & Monfort, A. (1995). *Statistics and Econometric Models*, volume 1. Cambridge, U.K.: Cambridge University Press.

Haber, R., & Unbehauen, H. (1990). Structure identification of nonlinear dynamic systems–A survey on input/output approaches. *Automatica*, *26*, 651–677.

Heiser, R. F., & Parrish, W. R. (1989). Representing physical data with rational functions. *Industrial & Engineering Chemistry Research*, *28*, 484–489.

Henderson, H. V., & Searle, S. R. (1981). On deriving the inverse of a sum of matrices. *SIAM Review*, *23*, 53–60.

Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statistics*, *40*, 633–643.

Jia, L.-J., Tao, R., Kanae, S., Yang, Z.-J., & Wada, K. (2011). A unified framework for bias compensation based methods in correlated noise case. *IEEE Transactions on Automatic Control*, *56*, 625–629.

Kamenski, D. I., & Dimitrov, S. D. (1993). Parameter estimation in differential equations by application of rational functions. *Computers & Chemical Engineering*, *17*, 643–651.

Klipp, E., Herwig, R., Kowald, A., Wierling, C., & Lehrach, H. (2005). *Systems Biology in Practice: Concepts, Implementation and Application*. Weinheim, Germany: Wiley-VCH.

Lehmann, E. L., & Casella, G. (1998). *Theory of Point Estimation*. New York: Springer-Verlag.

Leontaritis, I. J., & Billings, S. A. (1985). Input-output parametric models for non-linear systems Part I: Deterministic non-linear systems. *International Journal of Control*, *41*, 303–328.

Ljung, L. (1999). *System Identification: Theory for the User*. Upper Saddle River, NJ: Prentice-Hall.

Ljung, L. (2012). *System Identification Toolbox for Use with MATLAB*. (8th ed.). Natick, MA: The MathWorks, Inc.

Němcová, J., & van Schuppen, J. H. (2009). Realization theory for rational systems: The existence of rational realizations. *SIAM Journal on Control and Optimization*, *48*, 2840–2856.

Němcová, J., & van Schuppen, J. H. (2010). Realization theory for rational systems: Minimal rational realizations. *Acta Applicandae Mathematicae*, *110*, 605–626.

Söderström, T. (2007). Errors-in-variables methods in system identification. *Automatica*, *43*, 939–958.

Söderström, T., & Stoica, P. (1989). *System Identification*. London: Prentice Hall International.

Sontag, E. D. (1979). *Polynomial Response Maps*. Berlin:

Springer-Verlag.

Stoica, P., & Söderström, T. (1982). Bias correction in least-squares identification. *International Journal of Control*, *35*, 449–457.

Treloar, M. A. (1974). *Effects of Puromycin on Galactosyltransferase of Golgi Membranes*. Master's Thesis, University of Toronto.

Zhao, W., Zheng, W. X., & Bai, E.-W. (2013). A recursive local linear estimator for identification of nonlinear ARX systems: Asymptotical convergence and applications. *IEEE Transactions on Automatic Control*, *58*, 3054–3069.

Zheng, W. X. (1998). On a least square based algorithm for identification of stochastic linear systems. *IEEE Transactions on Signal Processing*, *46*, 1631–1638.

Zheng, W. X. (2002). A bias correction method for identification of linear dynamic errors-in-variables models. *IEEE Transactions on Automatic Control*, *47*, 1142–1147.

Zheng, W. X., & Feng, C.-B. (1995). A bias-correction method for indirect identification of closed-loop systems. *Automatica*, *31*, 1019–1024.

Zhu, Q., Wang, Y., Zhao, D., Li, S., & Billings, S. A. (2015). Review of rational (total) nonlinear dynamic system modelling, identification, and control. *International Journal of Systems Science*, *46*, 2122–2133.

Zhu, Q. M. (2003). A back propagation algorithm to estimate the parameters of non-linear dynamic rational models. *Applied Mathematical Modelling*, *27*, 169–187.

Zhu, Q. M. (2005). An implicit least squares algorithm for nonlinear rational model parameter estimation. *Applied Mathematical Modelling*, *29*, 673–689.