Proceedings of the
47th IEEE Conference on Decision and Control
Cancun, Mexico, Dec. 9-11, 2008

Some Convergence Properties of Multi-Step Prediction Error Identification Criteria

Marcello Farina and Luigi Piroddi, *Member, IEEE*

Abstract— **Multi-step prediction error identification methods are preferred over plain one-step ahead prediction error ones in application contexts (*e.g.,* predictive control) where model accuracy is required over a wide horizon. For sufficiently high prediction horizons, their properties can be shown to be conveniently related to those of output error methods, for which several important issues (*e.g.*, uniqueness of estimation, robustness with respect to the noise model) have been characterized in the literature. The convergence properties of such criteria with respect to the prediction horizon are analyzed.**

I. Introduction

The prediction error minimization (PEM) approach is the core of black-box identification [16], [8], in view of the availability of simple identification algorithms for many model classes and of a significant wealth of theoretical results (*e.g.*, the characterization of bias and variance of the model estimation). Exact correspondence both of the input-output model and of the noise model to the data generating mechanism is necessary for unbiased estimation. In addition, in non ideal conditions, long-term prediction accuracy is not generally achieved by PEM methods and usage of models identified with this approach is not recommended for simulation, system analysis, or control design purposes.

On the other hand, simulation error minimization (SEM) methods[1], while computationally costly, display several desirable properties that overcome the mentioned problems. Most notably the SEM approach can provide unbiased estimates of the input-output model, regardless of the real system noise model structure [15], [16]. Also it is generally capable of obtaining more accurate simulation models. From a numerical minimization standpoint, the shape of the PEM cost function is found to be flatter than the corresponding SEM one near the global minimum [2], [10], which may negatively affect the convergence properties of gradient-based identification algorithms.

An intermediate step between the PEM and the SEM approach is represented by *k*-steps ahead single-step (SSPEM) and multi-step (MSPEM) prediction error minimization methods, that improve the simulation accuracy by extending the prediction horizon and, at the same time, are less costly from the computational side. These approaches are only seldom used, and often confined to specific application contexts. For example, SSPEM methods are sometimes used in time-series analysis (see [1], [7] for some recent works) and in the predictive control context[2] [5], [13], [11] for better long-range prediction performance.

In this paper we first review some major properties of the SEM estimator. Briefly, SEM identification guarantees correctness and consistency of the estimation of the process model, given standard persistence of excitation conditions on the input signal and in the case of matching structure of the process model. The SEM approach also results in a more balanced frequency weighting of the estimation in the under-parameterized case, as opposed to PEM. Secondly, we consider the properties of the SSPEM and MSPEM criteria as the prediction horizon is extended. In this regard, [3], [4] have shown that in the MSPEM estimation of an under-parameterized process model the bias generally shifts from a strong weighting in the high frequency range (in PEM case) to a more uniform frequency weighting, so that MSPEM methods should help obtaining process models that are more accurate in the low and mid frequency ranges. This can be now more clearly interpreted in view of the relation between the MSPEM and SEM criteria. In fact, we prove that, as the prediction horizon increases, both the SSPEM and MSPEM criteria tend to SEM one. This has the important implication that, for sufficiently high prediction horizons, the multi-step approaches are expected to inherit the properties of the SEM, *e.g.*, the unbiasedness even in the absence of a noise model.

The different convergence properties of the two criteria are also discussed. In this regard, the SSPEM approach is shown to tend to the SEM estimates more quickly but less smoothly than the MSPEM one (see also [5]). For this reason, a weighted version of the MSPEM (briefly denoted WM-SPEM) is also introduced, on the grounds that by suitably modulating the weighting function one can achieve both smooth and rapid convergence. These results may provide the basis for the development of computationally viable methods for SEM identification, as an alternative to direct minimization of the simulation error and iterating over the prediction horizon.

This work was supported by the MIUR project Innovative Techniques for Identification and Adaptive Control of Industrial Systems”

M. Farina and L. Piroddi are with the Dipartimento di Elettronica e Informazione, Politecnico di Milano, {farina,piroddi}@elet.polimi.it

[1]In the model identification context, the SEM approach is more frequently denoted output error minimization (OEM) [16]. In the neural networks community it is instead referred to as parallel (output error) training, as opposed to series-parallel (equation error) [9]. Note also that simulation is sometimes denoted free-run model simulation or model prediction.

[2]In that context, multi-step ahead prediction methods are also denoted MRI (Model predictive control Relevant Identification) [3], [4], [6] and LRPI [14], [12] (long range prediction identification) methods.

978-1-4244-3124-3/08/$25.00 ©2008 IEEE

II. Preliminary notation and definitions

A. Linear external representation systems

Let the data generating system $\mathscr{S}$ be defined by a linear discrete-time equation, in the conventional form:

$$y(t) = \tilde{y}(t) + n(t), \text{ where} \tag{1a}$$
$$\tilde{y}(t) = G(z)u(t) \text{ and} \tag{1b}$$
$$n(t) = H(z)e(t) \tag{1c}$$

denote the process and noise model, respectively. Here, $G(z)$ and $H(z)$ are suitable transfer functions in the domain of the complex variable z (*i.e.* $z^{-1}x(t) \equiv x(t-1)$), and $e(\cdot) \sim WN(0,\lambda^2)$ is a white noise process. Notice that (1) encompasses the well known FIR, ARX, ARMAX and OE model classes. In this work, we assume that the objective of the identification process is the estimation of the process model alone. In this respect:

$$G(z) = \frac{N_G(z)}{D_G(z)} = \frac{b_0^o + b_1^o z^{-1} + \ldots + b_{n_b}^o z^{-n_b}}{1 + a_1^o z^{-1} + \ldots + a_{n_a}^o z^{-n_a}},$$

where the polynomials $N_G(z)$ and $D_G(z)$ are assumed to be coprime. The *estimator model* is expressed as:

$$\hat{y}(t/0) = \hat{G}(z)u(t), \text{ where} \tag{2}$$

$$\hat{G}(z) = \frac{\hat{N}_G(z)}{\hat{D}_G(z)} = \frac{b_0 + b_1 z^{-1} + \ldots + b_{\hat{n}_b} z^{-\hat{n}_b}}{1 + a_1 z^{-1} + \ldots + a_{\hat{n}_a} z^{-\hat{n}_a}}$$

Remark that in general $\hat{n}_a \neq n_a$ and $\hat{n}_b \neq n_b$. In the sequel, it is assumed that both $G(z)$ and $\hat{G}(z)$ are asymptotically stable transfer functions. Models (1b) and (2) are deterministic systems, whose output can be computed by iteration of their recursive equations, given the appropriate initial conditions on past output values (the input signal is known):

$$\tilde{\mathbf{Y}}_0 = \begin{bmatrix} \tilde{y}(-n_a+1) \\ \vdots \\ \tilde{y}(0) \end{bmatrix}, \; \hat{\mathbf{Y}}_0 = \begin{bmatrix} \hat{y}(-\hat{n}_a+1) \\ \vdots \\ \hat{y}(0) \end{bmatrix}$$

The system and model parameters are synthetically expressed in vector form as $\theta = [a_1 \ldots a_{\hat{n}_a}\, b_0 \ldots b_{\hat{n}_b}]^T$ and $\theta^o = [a_1^o \ldots a_{n_a}^o\, b_0^o \ldots b_{n_b}^o]^T$, respectively.

B. Criteria and algorithms

In the following, the mentioned PEM, SEM, SSPEM, MSPEM and WMSPEM identification criteria are defined and compared. For notation purposes, we denote as $\hat{y}(t_F/t_I)$ the predictor of y at time t_F, given the values of $u(\cdot)$ and $y(\cdot)$ up to time t_I (dependent on θ). The estimation error is defined as follows:

$$\varepsilon(t_F/t_I) = y(t_F) - \hat{y}(t_F/t_I) = \tilde{y}(t_F) - \hat{y}(t_F/t_I) + n(t_F). \tag{3}$$

Mean square error cost functions are used throughout the paper. The k-steps ahead SSPEM cost function is defined as

$$J_P^N(k) = \frac{1}{N-k+1}\sum_{t=k}^{N} \varepsilon(t/t-k)^2. \tag{4a}$$

N denotes the number of data available. The PEM criterion is $J_P^N(1)$. Similarly, the SEM cost function is defined as

$$J_S^N = \frac{1}{N}\sum_{t=1}^{N} \varepsilon(t/0)^2. \tag{4b}$$

The multi-step PEM (MSPEM) criterion with maximum prediction horizon k is defined as the average of the first k SSPEM cost functions:

$$J_{MP}^N(k) = \frac{1}{k}\sum_{i=1}^{k} J_P^{N-k+i}(i). \tag{4c}$$

We also define a weighted version of the multi-step PEM (WMSPEM) criterion, with maximum prediction horizon k and "forgetting factor" λ, such that $0 < \lambda < 1$:

$$J_{WMP}^N(k) = \frac{1-\lambda}{1-\lambda^k}\sum_{i=1}^{k} \lambda^{k-i} J_P^{N-k+i}(i). \tag{4d}$$

In (4d), the multiplicative term $\frac{1-\lambda}{1-\lambda^k}$ is introduced for normalization reasons.

Asymptotic versions of such criteria are also defined. Assuming that signal $u(t)$ is *quasi-stationary* [8], the asymptotic SSPEM, SEM, MSPEM (the MRI criterion used in [3], [4]) and WMSPEM cost functions are:

$$\bar{J}_P(k) = \bar{\mathbb{E}}\left[\varepsilon(t/t-k)^2\right], \tag{5a}$$
$$\bar{J}_S = \bar{\mathbb{E}}\left[\varepsilon(t/0)^2\right], \tag{5b}$$
$$\bar{J}_{MP}(k) = \frac{1}{k}\sum_{i=1}^{k} \bar{J}_P(i), \tag{5c}$$
$$\bar{J}_{WMP}(k) = \frac{1-\lambda}{1-\lambda^k}\sum_{i=1}^{k} \lambda^{k-i}\bar{J}_P(i), \tag{5d}$$

respectively, where the operator $\bar{\mathbb{E}}[\cdot]$ is defined as:

$$\bar{\mathbb{E}}[f(t)] = \lim_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N}\mathbb{E}[f(t)],$$

and $\mathbb{E}[\cdot]$ denotes the statistical expectation value operator.

Provided that both system (1) and model (2) are asymptotically stable, and that $u(t)$ and $y(t)$ are jointly quasi-stationary, the following statements hold [15], [16], [8]:

$$J_P^N(k) \xrightarrow{N\to\infty} \bar{J}_P(k), \tag{6a}$$
$$J_S^N \xrightarrow{N\to\infty} \bar{J}_S. \tag{6b}$$

These properties, in turn, trivially imply:

$$J_{MP}^N(k) \xrightarrow{N\to\infty} \bar{J}_{MP}(k), \tag{6c}$$
$$J_{WMP}^N(k) \xrightarrow{N\to\infty} \bar{J}_{WMP}(k). \tag{6d}$$

III. Properties of the SEM approach

The properties of the SEM identification approach have been widely investigated in the literature, see, e.g., [15], [16], [8]. In this section we recall some major known results.

In the case of perfect process model matching, the parameter estimate θ_N which minimizes the sampled criterion (4b) converges to θ^o as $N \to \infty$, under suitable excitation conditions. This can be derived by recalling (6b) and by proving that θ, minimizing (5b), is equal to θ^o. In fact, the simulation error $\varepsilon(t/0)$ is computed as:

$$\varepsilon(t/0) = \tilde{y}(t) - \hat{y}(t/0) + n(t), \tag{7}$$

where $\tilde{y}(t)$ is the generic response of system (1b). Given the uncorrelation of $n(\cdot)$ and $\tilde{y}(\cdot)$ (since $n(\cdot)$ and $u(\cdot)$ are uncorrelated), the SEM criterion is reformulated as

$$\bar{J}_S = \tilde{J}_S + \mathbb{E}\left[n(t)^2\right], \text{ where} \tag{8}$$

$$\tilde{J}_S = \bar{\mathbb{E}}\left[(\tilde{y}(t) - \hat{y}(t|0))^2\right]. \tag{9}$$

Remark that $\tilde{J}_S$ is the component of $\bar{J}_S$ which depends on the unknown parameters, *i.e.*, the parameterizations minimizing $\tilde{J}_S$ are also optimal for $\bar{J}_S$. Now, observe that:

$$\tilde{J}_S = \bar{\mathbb{E}}\left[\left((G(z) - \hat{G}(z))u(t)\right)^2\right]. \tag{10}$$

Therefore, it is apparent that, in the studied case of model and system structures matching, $\tilde{J}_S(\theta^o) = 0$, so that $\theta^o \in \Theta$, where $\Theta = \{\bar{\theta} | \bar{\theta} = \underset{\theta}{\arg\min}\, \bar{J}_S = \underset{\theta}{\arg\min}\, \tilde{J}_S\}$.

In [15] it is shown that the only stationary point of $\bar{J}_S$ is θ^o if $u(t)$ is a white noise or an ARMA signal of "sufficiently" low order, compared with the order of $G(z)$. More in general, the exact parameterization is actually obtained as a result of the identification process if $u(t)$ is persistently exciting of order $n_a + n_b + 1$, see [8]. In that case, $\theta = \theta^o$.

Regarding the uncertainty of the estimates, see [16]:

$$\sqrt{N}(\theta_N - \theta^o) \overset{N\to\infty}{\longrightarrow} \mathcal{N}(0, P_{SEM}),$$

where P_{SEM} is the related covariance matrix.

Finally, the under-parameterized case is also of practical interest, and has been studied in the framework of MPC relevant identification criteria, see [4], [13], [14]. It turns out that the model identified by the SEM approach is the "best model" in the frequency range of the input signal bandwidth. In fact, the frequency domain equivalent of expression (10), obtained by means of Parceval's theorem, is

$$\tilde{J}_S = \frac{1}{2\pi}\int_{-\pi}^{\pi} \left|G(e^{j\omega}) - \hat{G}(e^{j\omega})\right|^2 \Phi_u(\omega)d\omega \tag{11}$$

where $\Phi_u(\omega)$ denotes the spectrum of the input signal $u(t)$. Notice that in this case the estimation error is weighted only by the frequency characteristics of the input signal, whereas in the PEM case it is affected by the noise model as well.

IV. Convergence properties of the SSPEM, MSPEM and WMPEM criteria

It is of major interest in practical applications to analyze the estimation correctness properties of the employed criteria when the noise structure is unknown. Unlike the SEM, the PEM approach does not yield unbiased estimates if the noise model does not match that of the system. Here we show that the SSPEM, MSPEM and WMPEM criteria provide unbiased estimates as $k \to \infty$. Therefore, any of them could be used for robust input/output model identification. The choice is mainly a matter of computational efficiency of the respective algorithms. In this respect there is a significant difference between the SSPEM, MSPEM and WMPEM criteria (all basically involving the same set of predictor models, thus having comparable computational complexity for a given k, but different convergence properties), and the SEM, which requires a lengthy iterative optimization procedure.

A. Calculation of the estimation error

Let $\tilde{y}(t+k/t)$ be the value of $\tilde{y}(t+k)$ computed by iterating k times the recursive equation (1b), as a function of data $\tilde{y}^t$ (values of $\tilde{y}$ up to time t) and u^{t+k} (values of u up to time $t+k$). One can obtain:

$$\tilde{y}(t+k/t) = R_k(z)\,\tilde{y}(t) + E_k(z)N_G(z)u(t+k)$$

where $R_k(z)$ and $E_k(z)$ solve the diophantine equation:

$$1 = D_G(z)E_k(z) + R_k(z)\,z^{-k}$$

Analogously, the k-steps predictor of $y(\cdot)$ can be expressed as a function of the data y^t and u^{t+k} as:

$$\begin{aligned}\hat{y}(t+k/t) &= \hat{R}_k(z)\,y(t) + \hat{E}_k(z)\hat{N}_G(z)u(t+k)\\ &= \hat{R}_k(z)\,\tilde{y}(t) + \hat{E}_k(z)\hat{N}_G(z)u(t+k) + v(t),\end{aligned}$$

where $v(t) = \hat{R}_k(z)\,n(t)$, and $\hat{R}_k(z)$ and $\hat{E}_k(z)$ solve

$$1 = \hat{D}_G(z)\hat{E}_k(z) + \hat{R}_k(z)\,z^{-k}. \tag{12}$$

Remark that, being $n(t)$ a zero mean stochastic process, $v(t)$ is a zero mean stochastic stationary process as well.

Lemma 1: Polynomials $\hat{E}_k(z)$ and $\hat{R}_k(z)$ can be reformulated, in matrix notation, as:

$$\hat{E}_k(z) = \mathbf{C}\sum_{i=0}^{k-1}\mathbf{A}^i\mathbf{B}\,z^{-i} \tag{13a}$$

$$\hat{R}_k(z) = \mathbf{C}\mathbf{A}^k\begin{bmatrix} z^{-n_a+1}\\ \vdots\\ 1\end{bmatrix} \tag{13b}$$

where $\mathbf{C} = \begin{bmatrix}0 & \dots & 0 & 1\end{bmatrix}$, $\mathbf{B} = \mathbf{C}^T$ and

$$\mathbf{A} = \begin{bmatrix} 0 & & I_{\hat{n}_a-1} & \\ \vdots & & & \\ 0 & & & \\ -a_{\hat{n}_a} & -a_{\hat{n}_a-1} & \dots & -a_1\end{bmatrix}.$$

Proof: Consider the linear model of the form:

$$\begin{aligned}\mathbf{x}(t+1) &= \mathbf{A}\mathbf{x}(t) + \mathbf{B}\,u(t+1)\\ y(t) &= \mathbf{C}x(t)\end{aligned}$$

where $\mathbf{x} \in \mathbf{R}^{n_a}$. The transfer function $\mathbf{C}(I - \mathbf{A}z^{-1})^{-1}\mathbf{B}$ of such system equals $1/\hat{D}_G(z)$. Furthermore, since $x_1(t) = y(t - n_a + 1)$, ..., $x_n(t) = y(t)$, it holds that:

$$(I - \mathbf{A}z^{-1})^{-1}\mathbf{B} = \mathbf{C}(I - \mathbf{A}z^{-1})^{-1}\mathbf{B}\begin{bmatrix} z^{-n_a+1}\\ \vdots\\ 1\end{bmatrix} = \frac{1}{\hat{D}_G(z)}\begin{bmatrix} z^{-n_a+1}\\ \vdots\\ 1\end{bmatrix} \tag{14}$$

Now, observing that

$$\begin{aligned}\mathbf{C}(I - \mathbf{A}z^{-1})^{-1}\mathbf{B} &= \mathbf{C}\textstyle\sum_{i=0}^{\infty}\mathbf{A}^i\mathbf{B}\,z^{-i}\\ &= \mathbf{C}\textstyle\sum_{i=0}^{k-1}\mathbf{A}^i\mathbf{B}\,z^{-i} + \mathbf{C}\mathbf{A}^k z^{-k}(I - \mathbf{A}z^{-1})^{-1}\mathbf{B}\\ &= \mathbf{C}\textstyle\sum_{i=0}^{k-1}\mathbf{A}^i\mathbf{B}\,z^{-i} + \mathbf{C}\mathbf{A}^k z^{-k}\frac{1}{\hat{D}_G(z)}\begin{bmatrix} z^{-n_a+1}\\ \dots\\ 1\end{bmatrix}\\ &= \hat{E}_k(z) + \frac{\hat{R}_k(z)}{\hat{D}_G(z)}\,z^{-k}\end{aligned}$$

equation (12) is verified. ■

In view of eq. (13b), $v(t)$ can be reformulated as follows:

$$v(t)=\sum_{i=1}^{n_a}\mathbf{C}\mathbf{A}^k\mathbf{e}_i\,n(t-n_a+i) \tag{15}$$

where $\mathbf{e}_i$ represents the i-th canonical basis vector. The k-steps prediction error is therefore:

$$\begin{aligned}\varepsilon(t+k/t)=&\tilde{\varepsilon}(t+k/t)+n(t+k)-v(t),\ \text{where}\\ \tilde{\varepsilon}(t+k/t)=&(R_k(z)-\hat{R}_k(z))\,\tilde{y}(t)+\\ &+(E_k(z)N_G(z)-\hat{E}_k(z)\hat{N}_G(z))u(t+k)\\ =&\,(G(z)-\hat{G}(z))\hat{E}_k(z)\hat{D}_G(z)u(t+k)\end{aligned} \tag{16a}$$

B. Convergence results

The following result shows the convergence of the asymptotic SSPEM criterion to the asymptotic SEM criterion and yields a bounding function for the convergence transient.

Theorem 1: Let $\mathscr{S}$ be a linear external representation system with additive noise model of type (1). Let $\mathscr{M}$ be a model of structure (2). Let $\bar{J}_P(k)$ be the SSPEM criterion (5a) and let $\bar{J}_S$ be the asymptotic SEM criterion (5b). Assume that $e(\cdot)$ is uncorrelated with input signal $u(\cdot)$. Then, there exist constants $M_P>0$ and $0<\alpha<1$ such that:

$$\frac{\left|\bar{J}_P(k)-\bar{J}_S\right|}{\bar{J}_S}\leq M_P\alpha^k, \tag{17}$$

Proof: Since $u(t)$ and $e(t)$ are uncorrelated and observing that $\mathbb{E}[n(t)]=0$, the cost function $\bar{J}_P(k)$ (4a) is:

$$\bar{J}_P(k)=\tilde{J}_P(k)+\Delta_k^N, \tag{18}$$

where $\tilde{J}_P(k)=\mathbb{E}\left[\tilde{\varepsilon}(t+k/t)^2\right]$ and $\Delta_k^N=\mathbb{E}\left[(n(t+k)-v(t))^2\right]$. Furthermore, since $n(t)$ is a stationary stochastic process the last term can be rewritten as:

$$\Delta_k^N=\mathbb{E}\left[n(t+k)^2\right]+\Delta_k^{(1)}+\Delta_k^{(2)} \tag{19a}$$

$$\Delta_k^{(1)}=\mathbb{E}\left[v(t)^2\right] \tag{19b}$$

$$\Delta_k^{(2)}=-2\,\mathbb{E}\left[n(t+k)\,v(t)\right] \tag{19c}$$

Recalling (8), by (18) and (19a):

$$\begin{aligned}\left|\bar{J}_P(k)-\bar{J}_S\right|&\leq\left|\tilde{J}_P(k)-\tilde{J}_S\right|+\left|\Delta_k^N-\mathbb{E}\left[n(t)^2\right]\right|\\&\leq\left|\tilde{J}_P(k)-\tilde{J}_S\right|+\left|\Delta_k^{(1)}\right|+\left|\Delta_k^{(2)}\right|\end{aligned} \tag{20}$$

Recall the equation (12). Using Parceval's theorem, (11) and (16a), the first term of (20) can be reformulated as:

$$\begin{aligned}|\tilde{J}_P(k)-\tilde{J}_S|&=\frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\left|\hat{E}_k(e^{j\omega})\hat{D}_G(e^{j\omega})\right|^2-1\right|\times\\&\times\left|G(e^{j\omega})-\hat{G}(e^{j\omega})\right|^2\Phi_u(\omega)d\omega\\&\leq\frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\left|\hat{R}_k(e^{j\omega})\right|^2+2\left|\hat{R}_k(e^{j\omega})\right|\right]\times\\&\times\left|G(e^{j\omega})-\hat{G}(e^{j\omega})\right|^2\Phi_u(\omega)d\omega.\end{aligned} \tag{21}$$

Observe that, by equation (13b),

$$|\hat{R}_k(e^{j\omega})|\leq\|\mathbf{C}\mathbf{A}^k\|_\infty\leq\|\mathbf{A}^k\|_\infty,$$

where the infinity norm of a matrix $\mathbf{M}\in\mathbb{R}^{h\times k}$ is given by:

$$\|\mathbf{M}\|_\infty=\max_{i=1,\ldots,h}\sum_{j=1}^{k}\left|\mathbf{M}_{i,j}\right|.$$

Since matrix $\mathbf{A}$ is Shur, there exists a constant β_∞ such that:

$$\|\mathbf{A}^k\|_\infty\leq\beta_\infty\,\alpha^k \tag{22}$$

and α is the spectral radius of matrix $\mathbf{A}$. Therefore:

$$|\tilde{J}_P(k)-\tilde{J}_S|\leq(\beta_\infty^2\,\alpha^k+2\beta_\infty)\,\tilde{J}_S\,\alpha^k. \tag{23}$$

Then, by (15) and (19b),

$$\begin{aligned}|\Delta_k^{(1)}|&=\left|\mathbb{E}\left[\sum_{i=1}^{n_a}\mathbf{C}\mathbf{A}^k\mathbf{e}_i\,n(t-n_a+i)\sum_{j=1}^{n_a}n(t-n_a+j)\mathbf{e}_j^T(\mathbf{C}\mathbf{A}^k)^T\right]\right|\\&=\left|\mathbf{C}\mathbf{A}^k\sum_{i=1}^{n_a}\sum_{j=1}^{n_a}\mathbf{e}_i\mathbf{e}_j^T\,\mathbb{E}\left[n(t-n_a+i)n(t-n_a+j)\right](\mathbf{C}\mathbf{A}^k)^T\right|\\&\leq|\mathbf{C}\mathbf{A}^k\sum_{i=1}^{n_a}\sum_{j=1}^{n_a}\mathbf{e}_i\mathbf{e}_j^T(\mathbf{C}\mathbf{A}^k)^T|\,\gamma_n(0)\\&\leq\|\mathbf{C}\mathbf{A}^k\|_\infty^2\gamma_n(0)\leq\|\mathbf{A}^k\|_\infty^2\gamma_n(0)\leq\beta_\infty^2\alpha^{2k}\gamma_n(0)\end{aligned} \tag{24}$$

where expression (22) has been used again, and $\gamma_n(\tau)$ is the covariance function of $n(t)$. On the other hand, expression (19c) can be reformulated as:

$$\Delta_k^{(2)}=-2\sum_{i=1}^{n_a}\mathbf{C}\mathbf{A}^k\mathbf{e}_i\,\gamma_n(k+n_a-i)$$

Its absolute value is

$$\begin{aligned}|\Delta_k^{(2)}|&\leq2\,|\sum_{j=1}^{n_a}\mathbf{C}\mathbf{A}^k\mathbf{e}_j|\,\gamma_n(0)\\&\leq2\|\mathbf{C}\mathbf{A}^k\|_\infty\gamma_n(0)\leq2\|\mathbf{A}^k\|_\infty\gamma_n(0)\\&\leq2\gamma_n(0)\beta_\infty\alpha^k.\end{aligned} \tag{25}$$

From equations (20), (23), (24) and (25):

$$\begin{aligned}\left|\bar{J}_P(k)-\bar{J}_S\right|&\leq(\beta_\infty^2\,\alpha^k+2\beta_\infty)\,(\tilde{J}_S+\gamma_n(0))\,\alpha^k\\&\leq(\beta_\infty^2+2\beta_\infty)\,\bar{J}_S\,\alpha^k\end{aligned}$$

which proves (17) with $M_P=\beta_\infty^2+2\beta_\infty$. ■

Corollary 1: Let $\mathscr{S}$ be a linear external representation system with additive noise model, of type (1). Let $\mathscr{M}$ be a model of structure (2). Let $\bar{J}_{MP}(k)$ and $\bar{J}_{WMP}(k)$ be the asymptotic MSPEM and WMSPEM criteria (5c) and (5d) respectively and let $\bar{J}_S$ be the SEM one (5b). Assume that $e(\cdot)$ is uncorrelated with input signal $u(\cdot)$. Then there exist constants $M_P>0$ and $0<\alpha<1$ such that:

$$\frac{\left|\bar{J}_{MP}(k)-\bar{J}_S\right|}{\bar{J}_S}\leq\frac{1}{k}\frac{\alpha}{1-\alpha}M_P, \tag{26a}$$

$$\frac{\left|\bar{J}_{WMP}(k)-\bar{J}_S\right|}{\bar{J}_S}\leq\alpha\frac{(\lambda^k-\alpha^k)}{\lambda-\alpha}M_P, \tag{26b}$$

λ being the "forgetting factor" appearing in criterion (5d).

Proof: From (17):

$$\begin{aligned}\left|\bar{J}_{MP}(k)-\bar{J}_S\right|&\leq\left|\frac{1}{k}\sum_{i=1}^{k}\bar{J}_P(i)-\bar{J}_S\right|\leq\frac{1}{k}\sum_{i=1}^{k}\left|\bar{J}_P(i)-\bar{J}_S\right|\\&\leq\frac{1}{k}\sum_{i=1}^{k}M_P\bar{J}_S\alpha^i\leq\frac{1}{k}\frac{\alpha(1-\alpha^k)}{1-\alpha}M_P\bar{J}_S\\&\leq\frac{1}{k}\frac{\alpha}{1-\alpha}M_P\bar{J}_S.\end{aligned}$$

Similarly:

$$\begin{aligned}\left|\bar{J}_{WMP}(k)-\bar{J}_S\right| &\le \frac{1-\lambda}{1-\lambda^k}\left|\sum_{i=1}^{k}\lambda^{k-i}(\bar{J}_P(i)-\bar{J}_S)\right| \\ &\le \sum_{i=1}^{k}\lambda^{k-i}\left|\bar{J}_P(i)-\bar{J}_S\right| \\ &\le \lambda^k\sum_{i=1}^{k}M_P(\frac{\alpha}{\lambda})^i\bar{J}_S \le \alpha\frac{\lambda^k-\alpha^k}{\lambda-\alpha}M_P\bar{J}_S\end{aligned}$$

and we obtain the statement (26b). ∎

The previous results guarantee that, for sufficiently long prediction horizons k, the SSPEM, MSPEM and WMSPEM criteria lead to "sufficiently good" identification results, regardless of the noise structure. Notice that the real parameter vector θ^o minimizes $\bar{J}_S$, whereas, in general, it does not minimize any multi-stage criteria if k is finite. Furthermore, $\left|\bar{J}_P(k)-\bar{J}_S\right|/\bar{J}_S$ and $\left|\bar{J}_{WP}(k)-\bar{J}_S\right|/\bar{J}_S$ have an exponential decay, whereas $\left|\bar{J}_{MP}(k)-\bar{J}_S\right|/\bar{J}_S$ has an hyperbolic decay (see Figure 1). The SSPEM yields a faster convergence (with respect to k) than both the MSPEM and WMSPEM criteria. However, choosing suitable values for λ, the convergence of the WMSPEM criterion can be significantly enhanced. This can be exploited to obtain a much smoother convergence than the SSPEM without sacrificing the rate of convergence. Figure 1 compares the bounding functions appearing in equations (17), (26a) and (26b), for different values of the "forgetting factor" λ, and assuming $M_P = 1$ and $\alpha = 0.3$.

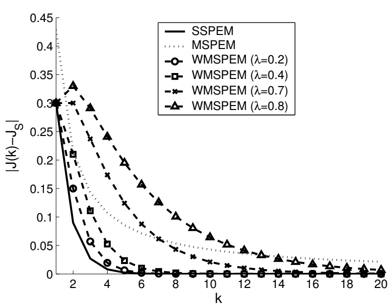

Fig. 1. Bounding functions for the convergence of the SSPEM, MSPEM and WMSPEM ($\lambda = 0.2, 0.4, 0.7, 0.8$) criteria to SEM.

V. EXAMPLE

Consider the data generation mechanism $\mathscr{S}$:

$$\tilde{y}(t) = -a^o\tilde{y}(t-1) + b_0^o u(t) + b_1^o u(t-1) \quad (27a)$$
$$y(t) = \tilde{y}(t) + n(t) \quad (27b)$$

where $a^o = -0.8$, $b_0^o = 1$, and $b_1^o = 0.3$.

Assume that the noise term $n(t)$ is generated according to three different model structures resulting in an overall model of type OE, ARMAX(1,2,1), ARARX(1,2,1), respectively:

$$\begin{aligned}&\text{OE}: \ n(t) = e_1(t) - 0.8e_1(t-1),\ e_1(\cdot) \sim (0, \tfrac{\sigma^2}{1.64}) \\ &\text{ARMAX}: \ n(t) = e_2(t) - 0.5e_2(t-1),\ e_2(\cdot) \sim (0, \tfrac{\sigma^2}{1.25}) \\ &\text{ARARX}: \ n(t) = -0.5n(t-1) + e_3(t),\ e_3(\cdot) \sim (0, 0.75\sigma^2)\end{aligned} \quad (28)$$

The values of the variances of the white noise terms $e_1(\cdot)$, $e_2(\cdot)$, $e_3(\cdot)$ have been assigned in order to enforce a noise variance $\mathbb{E}\left[n(t)^2\right] = \sigma^2$ in all the three cases. An ARX model $\mathscr{M}$ is used in the identification process with a process model structure matching that of system $\mathscr{S}$ (27):

$$y(t) = -ay(t-1) + b_0u(t) + b_1u(t-1) + \varepsilon(t) \quad (29)$$

Despite the model matching condition, it can be shown that least squares estimation minimizing the PEM criterion yields biased estimation for all cases (OE, ARMAX, ARARX). Consider now the minimization of the SEM criterion, the system being excited by a white noise input signal $u(t) \sim WN(0, \lambda_u^2)$. According to the results in Section III this criterion is minimized if and only if $a = a^o$, $b_0 = b_0^o$ and $b_1 = b_0^1$. This can be evidenced by computing (with Monte-Carlo simulation) the sampled probability distribution functions of the parameter estimates obtained with the SEM approach. For this purpose, the data generated according to equations (27), using the three noise structures in (28) with $\sigma^2 = 1$, are considered. The estimation is performed on 1000 realizations for each of the different processes (OE, ARMAX, ARARX). To test the consistency and correctness properties of the SEM estimator, the simulation experiment is repeated for different values of the number of samples, see Figure 2. In accordance with [15], there is no bias for any noise model structure and the estimation uncertainty decreases as the number of samples grows.

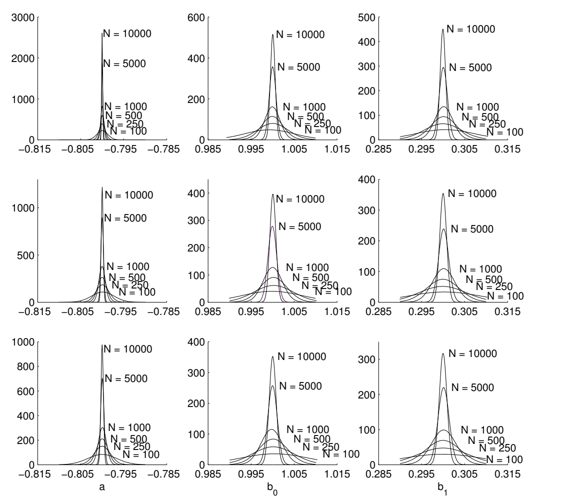

Fig. 2. Probability density functions of the estimates a, b_0, b_1 (first, second and third column respectively) obtained by the SEM criterion computed by Monte Carlo simulation in the three noise model cases (OE, ARMAX, ARARX, shown in the first, second and third row respectively) computed for different values of N (100, 250, 500, 1000, 5000, 10000).

Next, we test the SSPEM, MSPEM and WMSPEM approaches on the same example to show that they lead to unbiased estimates as the prediction horizon k increases. For this purpose we concentrate on a single realization of the ARARX model structure, and perform parameter estimation based on the SEM, SSPEM, MSPEM and WMSPEM ($\lambda = 0.8$) criteria, for $k = 1, \ldots, 50$. Here $N = 1000$, and the noise variance equals $\sigma^2 = 1$. Apparently, (see Figure

3), the SSPEM, MSPEM and WMSPEM estimates tend to the SEM ones as k increases. Since the SEM estimates are asymptotically unbiased, the same applies to the SSPEM and MSPEM estimates for $k \to \infty$. Figure 3 also illustrates the convergence properties of the criteria for increasing k. More precisely, MSPEM has a slower but smoother convergence (with respect to the maximum prediction horizon k) than SSPEM, thanks to its inherent averaging of predictions at different horizons, see (17), (26a) and (26b).

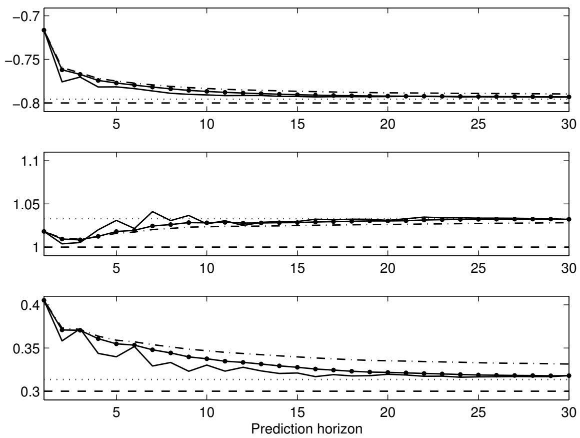

Fig. 3. Estimates of a^o, b_0^o, b_1^o (first, second and third row respectively) obtained by the SEM criterion (dotted line), by SSPEM (continuous line), by MSPEM (dash-dot line) and by WMSPEM (continuous line with circles), computed on a single realization of the process in the ARARX model case

A test is performed on the extended horizon prediction criteria, aimed at analyzing the estimation uncertainty as a function of k. The sampled probability density functions of the parameter estimates obtained with the SSPEM and MSPEM criteria, for $k = 1, 5, 10$ are computed. The usual Monte-Carlo method is employed for this purpose, with 1000 realizations of system (27) with an OE noise structure and $N = 1000$. As expected, the uncertainty for both the SSPEM and MSPEM criteria tends towards the uncertainty related to SEM estimation as k increases (see Figure 4).

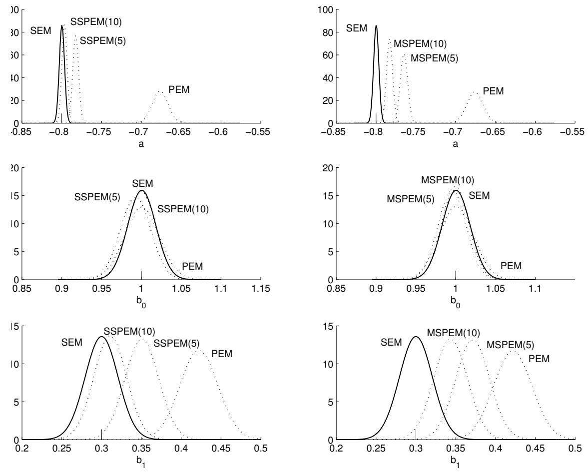

Fig. 4. Probability density functions of the estimates of a^o, b_0^o, b_1^o (first, second and third row respectively) obtained by the SEM (solid line), SSPEM (left column, dotted line), MSPEM (right column, dotted line) criteria computed by Monte Carlo simulation in the case of OE noise structure computed for $k = 1, 5, 10$, being k the number of steps ahead.

VI. CONCLUSIONS

In this paper the properties of multi-step identification criteria are analyzed. Indeed, criteria with extended prediction horizon are shown to inherit the correctness and consistency properties of the output error minimization (or simulation error minimization) approach, for sufficiently high values of the prediction horizon. The convergence properties of these criteria are investigated and a weighted version of the multi-step ahead prediction criterion is formulated to achieve the best compromise in terms of smoothness and rapidity of convergence. An example is provided to illustrate the stated results. An identification approach based on multi-step criteria is envisaged as an iterative method for output error estimation.

REFERENCES

[1] R. Chen, L. Yang, and C. Hafner. Nonparametric multi-step ahead prediction in time series analysis. *Journal of the Royal Statistical Society, Series B*, 66(3):669–686, 2004.

[2] P. Connally, K. Li, and G. W. Irwin. Prediction- and simulation-error based perceptron training: Solution space analysis and a novel combined training scheme. *Neurocomputing*, 70(4–6):819–827, 2007.

[3] R. B. Gopaluni, R. S. Patwardhan, and S. L. Shah. Bias distribution in MPC relevant identification. In *IFAC World Congress*, pages 2196–2201, Barcelona (Spain), 2002.

[4] R. B. Gopaluni, R. S. Patwardhan, and S. L. Shah. MPC relevant identification–tuning the noise model. *Journal of Process Control*, 14(6):699–714, 2004.

[5] S. B. Jorgensen J. B. Jorgensen. Comparison of prediction-error-modelling criteria. In *American Control Conference*, pages 140–146, New York (USA), 2007.

[6] S. B. Jorgensen J. B. Jorgensen. MPC-relevant prediction-error identification. In *American Control Conference*, pages 128–133, New York (USA), 2007.

[7] K. Judd and M. Small. Towards long-term prediction. *Physica D*, 136(1–2):31–44, 2000.

[8] L. Ljung. *System Identification: Theory for the User*. Prentice-Hall, Englewood Cliffs, NJ, 1999.

[9] K. S. Narendra and K. Parthasarathy. Identification and control of dynamical systems using neural networks. *IEEE Transactions on Neural Networks*, 1(1):4–27, 1990.

[10] L. Piroddi and W. Spinelli. Structure selection for polynomial NARX models based on simulation error minimization. In *13th IFAC Symposium on System Identification*, pages 371–376, Rotterdam (The Netherlands), 2003.

[11] J. A. Rossiter and B. Kouvaritakis. Modelling and implicit modelling for predictive control. *International Journal of Control*, 74(11):1085–1095, 2001.

[12] J. A. Rossiter, B. Kouvaritakis, and L. Huaccho Huatuco. The benefits of implicit modelling for predictive control. In *39th IEEE Conference on Decision and Control*, pages 183–188, Sydney (Australia), 2000.

[13] D. S. Shook, C. Mohtadi, and S. L. Shah. Identification for long-range predictive control. *IEE Proceedings Part D, Control Theory and Applications*, 138(1):75–84, 1991.

[14] D. S. Shook, C. Mohtadi, and S. L. Shah. A control-relevant identification strategy for GPC. *IEEE Transactions on Automatic Control*, 37(7):975–980, 1992.

[15] T. Söderström and P. Stoica. Some properties of the output error method. *Automatica*, 18(1):93–99, 1982.

[16] T. Söderström and P. Stoica. *System Identification*. Prentice-Hall, London, 1989.