Second-Order Volterra System Identification With Noisy Input–Output Measurements

Umut Ozertem, *Member, IEEE*, and Deniz Erdogmus, *Member, IEEE*

***Abstract*—System identification with noisy input–output measurements has been dominantly addressed through the optimization of the mean-squared-error criterion (MSE), especially in adaptive filtering. MSE is known to provide models that approximate the conditional expectation of the target output given the input; however, when the input signal is also contaminated by noise—a frequent occurrence—MSE yields biased estimates of the model parameters with the severity of the bias dependent on the noise power. This drawback has been addressed in various ways, including errors-in-variables techniques. Recently, error whitening criterion (EWC) and associated adaptation algorithms were proposed to address this issue in linear system identification. We extend the applicability of the main concept behind EWC to the unbiased identification of order-2 Volterra series models of nonlinear dynamical systems. The extension does not apply to higher order Volterra models. The main contribution of this letter is a statistical criterion that can be utilized to identify analytically the true parameters of an order-2 Volterra model from noisy input–output data. We also support the theoretical results with simulations; however online learning algorithms that can be derived for the proposed criterion will not be addressed.**

***Index Terms*—Discrete-time order-2 Volterra model, error whitening criterion, errors-in-variables, instrumental variables, system identification.**

I. Introduction

THE mean-squared-error (MSE) criterion has found widespread use in model fitting to noisy data, most relevantly through Wiener and Kolmogorov's theories on stochastic process modeling, which led to the current adaptive filtering theory [1]. However, in the presence of additive noise in the input data, the bias that the power of this noise imposes on the Wiener solution for an optimal linear filter has been an important topic of research, which has induced many alternative and interesting solutions, some of which are reviewed below. An important consideration in designing a solution to this problem is to avoid making certain strong assumptions about the signal distributions; although a Bayesian approach could be employed to determine the best solution under certain parametric density model assumptions for the processes involved, these solutions do not necessarily lend themselves to convenient real-time adaptation, and their success is dependent on the suitability of the density model that is assumed.

Principal subspace Wiener filtering is a trivial modification of the least-squares regression solution through principal component analysis (PCA)-based dimensionality reduction to improve signal-to-noise ratio (SNR) before employing the Wiener solution [1]. This approach has clear shortcomings, for instance, it is only applicable to high SNR situations. Total least squares (TLS) is a powerful technique that addresses the problem of input-noise-induced bias in linear regression [2], [3]. This method requires a singular value decomposition, therefore computationally efficient algorithms are possible; however, it is applicable under strict equality constraints for noise powers. The instrumental variables (IV) method, which evolved from the errors-in-variables literature in statistics, relies on exploiting information from the nonzero lags of the autocorrelation function of the signals and solves the problem of unbiased linear system identification in white noise [4]. Error whitening criterion is a recent improvement that uses the IV concept and generates numerically more stable algorithms for adaptation of linear filters [5]. The IV and EWC solutions can be modified nontrivially by simultaneously learning prewhitening filters in situations where the noise processes are not white [4].

Volterra series and their variants have traditionally been utilized for modeling nonlinear dynamical systems. Due to its linear-in-parameters nature, this representation exhibits similarities to linear regression and therefore provides us with a good starting point towards extending the solutions discussed above which are designed for linear systems to the problem of unbiased identification of nonlinear systems using noisy input and output measurements. System identification literature is rich in Volterra series identification papers, where the noiseless input and noisy output [i.e., the input $\mathbf{x}$ and the output $\mathbf{y} = V(\mathbf{x}) + \mathbf{n}$] are available for modeling. For example, Ogunfunmi and Chang employ Wiener filter theory and LMS to second-order Volterra series identification [6]. Koukoulas and Kalouptsidis present a more elaborate approach for the same problem using the cross-spectrum measurements of the noiseless input, and the noisy output [7]. However, the more realistic case, where both input and output measurements are corrupted by noise, is not addressed in Volterra system identification literature.

We propose a statistical objective measure in the spirit of IV and EWC that can be evaluated using samples acquired from the noisy input output processes, yet provide an unbiased estimate of the true system parameters of an order-2 Volterra model. We will demonstrate that with this criterion, an analytical solution for the optimal parameters can be derived. Online algorithms to calculate the solutions will not be presented due to limited space. Deriving stochastic gradient (LMS-variant) algorithms to solve for optimal solution of the proposed criterion is relatively straightforward, and a brief derivation will be provided in the Appendix.

Manuscript received August 30, 2007; revised June 09, 2008. Current version published December 12, 2008. This work was supported in part by the NSF under grants ECS-0524835 and ECS-0622239. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Wei Xing Zheng.

U. Ozertem is with Yahoo! Inc., Sunnyvale, CA 95054 USA (e-mail: umut@yahoo-inc.com).

D. Erdogmus is with the Electrical and Computer Engineering Department, Northeastern University, Boston, MA 02215 USA (e-mail: derdogmus@ieee.org).

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

Digital Object Identifier 10.1109/LSP.2008.2008478

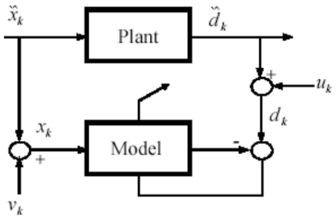

Fig. 1. Block diagram of the system identification problem.

II. Volterra System Identification With Noisy Measurements

Given the noisy input signal samples x_k and the noisy desired output signal samples d_k for our model, we are trying to estimate the model parameters for the unknown order-2 Volterra system, which is driven by the noiseless input signal $\tilde{x}_k$ and produces the noiseless output signal $\tilde{d}_k$. Letting u_k be the noise in the output measurements and v_k be the noise in the input measurements, the block diagram of the system identification problem becomes as shown in Fig. 1. For the unknown plant, the discrete Volterra series model is

$$d_t = h^0 + \sum_{k=0}^{L_1-1} h_k^1 x_{t-k} + \sum_{k_1=0}^{L_{21}-1} \sum_{k_2=0}^{L_{22}-1} h_{k_1 k_2}^2 x_{t-k_1} x_{t-k_2} \quad (1)$$

where h^0, h^1, and h^2 are the sets of model coefficients (convolution kernels) of order zero, one, and two, respectively. Without loss of generality, we assume that $h^0 = 0$ in the rest of this letter. The bias term h^0 can simply be obtained by $h^0 = E(d) - E(y)$ after adapting the other coefficients.

Observing (1), one can see that the Volterra series representation is linear in parameters. To write the equivalent linear model, one should put all the model coefficients and the corresponding input polynomials into vectors. This yields

$$d_t = \mathbf{h}^T \mathbf{x}_t \quad (2)$$

where the vector form of the coefficients is defined as

$$\mathbf{h} = \left[h_0^1 \dots h_{L_1-1}^1, h_0^2 h_0^2 \dots h_{L_{21}-1}^2 h_{L_{22}-1}^2\right]^T \quad (3)$$

and the vector form of all other scalars is defined using the same convention of ordering the components in the vector. Using the definitions of vectors according to the convention in (3), it is straightforward to see that (1) and (2) are equivalent. Now consider the system identification scenario. The noisy desired output is given as

$$d_t = \mathbf{h}^T \tilde{\mathbf{x}}_t + u_t \quad (4)$$

where u_t is the noise in the output measurements. Similarly, the noisy input signal is

$$x_t = \tilde{x}_t + v_t \quad (5)$$

where v_t is the noise in the input measurements. Using the noisy input signal x_t given in (6), we define the model output

$$y_t = \mathbf{w}^T \mathbf{x}_t. \quad (6)$$

Here, $\mathbf{w}$ defines the coefficients of the adaptive Volterra model and the vector forms of the coefficients w, and the noise signal v_t are defined in the same convention of (3).

Assumptions:

- $\tilde{x}_t$, u_t, and v_t are ergodic, strictly stationary, independent stochastic processes.
- $\tilde{x}_t$ is colored, and u_t and v_t are strictly white.
- $E(u_t) = 0$, $E(v_t) = 0$.
- Volterra model order is known.

The strict stationarity and whiteness of the processes can be relaxed to restrict joint moments only up to order 4 without loss of applicability. As stated, item 2 is equivalent to assuming u_t and v_t samples are iid. We exploit the cross-correlation of the noisy desired output d_t, and the model output y_t at lag-Δ

$$J_\Delta(\mathbf{w}) = E(d_t y_{t-\Delta}). \quad (7)$$

We introduce the objective function L to be minimized as the squared difference of two cross-correlation functions at opposite lags summed over m lags, where $m \geq (n-1)$ with n being the dimension of $\mathbf{w}$

$$\begin{aligned} L_\Delta(\mathbf{w}) &= (J_\Delta(\mathbf{w}) - J_{-\Delta}(\mathbf{w}))^2 \\ L(\mathbf{w}) &= \sum_{i=1}^{m} L_{\Delta_i}(\mathbf{w}). \end{aligned} \quad (8)$$

In the Appendix, we prove that the minimizer of $L(\mathbf{w})$ has to be parallel to $\mathbf{h}$; that is $\mathbf{w}_* = \alpha \mathbf{h}$, $\alpha \in \Re$. Overall, the criterion is quadratic in $\mathbf{w}$ and this can be seen easily by substituting (6) in (8): $L(\mathbf{w}) = \mathbf{w}^T \mathbf{P} \mathbf{w}$, where

$$\begin{aligned} \mathbf{p}_\Delta &= E(d_t \mathbf{x}_{t-\Delta} - d_{t-\Delta} \mathbf{x}_t) \\ \mathbf{P} &= \sum_{i=1}^{m} \mathbf{p}_{\Delta_i} \mathbf{p}_{\Delta_i}^T. \end{aligned} \quad (9)$$

The criterion L becomes zero when $\mathbf{w}$ is in the null space of $\mathbf{P}$ and the rank of this matrix (with theoretical expectations) satisfies $rank(\mathbf{P}) \leq \min(m, n-1)$ if the model order matches that of the true system. Hence, given input–output training data, one needs to identify the minor component of $\mathbf{P}$, which is the space spanned by $\mathbf{h}$ if $rank(\mathbf{P}) = (n-1)$.

To determine the correct scale, one should investigate two different lags of the cross-correlation function in (7) and use the ergodicity and stationarity assumptions once again to cancel out the terms that depend on the noise terms. We can find the arbitrary scale α as

$$\alpha = \frac{J_{\Delta_1}(\alpha \mathbf{h}) - J_{\Delta_2}(\alpha \mathbf{h})}{E(d_t d_{t-\Delta_1}) - E(d_t d_{t-\Delta_2})} \quad (10)$$

where $\Delta_1 \neq \pm \Delta_2$. With a derivation similar to the one in the Appendix, one can explicitly demonstrate that (10) can be written in terms of noisefree data statistics. Due to lack of space, we omit this derivation

$$\alpha = \frac{\alpha \mathbf{h}^T \left(E\left(\tilde{\mathbf{x}}_t \tilde{\mathbf{x}}_{t-\Delta_1}^T\right) - E\left(\tilde{\mathbf{x}}_t \tilde{\mathbf{x}}_{t-\Delta_2}^T\right)\right) \mathbf{h}}{\mathbf{h}^T \left(E\left(\tilde{\mathbf{x}}_t \tilde{\mathbf{x}}_{t-\Delta_1}^T\right) - E\left(\tilde{\mathbf{x}}_t \tilde{\mathbf{x}}_{t-\Delta_2}^T\right)\right) \mathbf{h}}. \quad (11)$$

Note that the derivation is based on the *true* values of the correlation and cross-correlation functions. In fact, the underlying true correlation functions are not available in real applications, and when all the expected values above are estimated from the data samples, the contribution of noise terms diminish to zero with increasing number of samples asymptotically according to the law of large numbers. In the experimental results section,

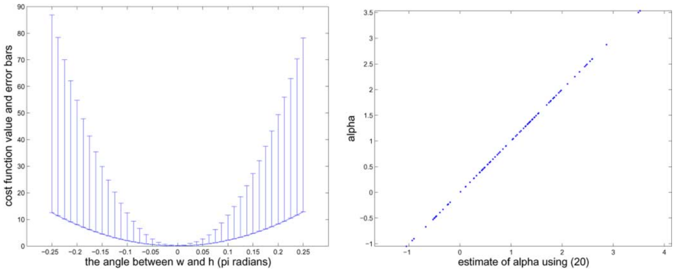

Fig. 2. Cost function value versus the angle between $\mathbf{w}$ and $\mathbf{h}$ (top), and the scatter plot of α and its estimator given in (10) (bottom).

we will utilize sample averages and also investigate the dependency on finite sample sizes for a particular example to observe that the estimation variance is inversely proportional to sample size.

III. Experimental Results

Here, we will observe two characteristics of the proposed criterion: minimization of (8) at weights that are aligned with the true parameter vector, and the asymptotic convergence of the criterion to zero with increasing sample size. We assume that the order of the system is known, and we generate a random Volterra system of order two. For this random system parameterized with $\mathbf{h}$, we plot a cross-section of the criterion L_Δ in a two-dimensional subspace that include $\mathbf{h}$ for 40 different $\mathbf{w}$ values selected and arranged such that the set of chosen $\mathbf{w}$ values linearly span the range of angles with $\mathbf{h}$ between $(-\pi/4, \pi/4)$. The set contains a weight vector that is aligned with the true parameter vector (zero-angle). To decrease the effects of finite size sample estimators, we run this experiment for 100 Monte Carlo simulations using 10 000 training samples for each run, while keeping the correlation function of the noiseless input signal and the unidentified system the same. The results are presented in Fig. 2, showing the expected parabolic cross-section of the criterion. In Fig. 2(a), the average value of the cost function versus the angle between $\mathbf{w}$ and $\mathbf{h}$ is shown along with the corresponding 90% confidence level error bars. One can see that the minimum is obtained when $\mathbf{w}$ and $\mathbf{h}$ are aligned. Another observation is that the robustness of the system identification scheme increases around the optimizer. The magnitude of the weight vector is estimated using (10) over the same 100 Monte Carlo runs of 10 000 training samples each. The scatter plot of α versus its estimate is given in Fig. 2(b), which demonstrates excellent correlation in the estimation of this parameter.

The effect of sample size on the estimated cost is studied with Monte Carlo simulations. Previous experiments demonstrate that the variance of the criterion at the line spanned by the true parameter vector is significantly small that at other weight vectors. For a second-order Volterra system, we performed 100 Monte Carlo trials with varying number of samples in training set (from 10 to 10^5) and evaluated the criterion at true parameter vector. Fig. 3 shows the average minimum criterion versus sample size. The linear decay in the log-log scale indicates the $1/N$ dependency of the statistical variance of (7).

The effect of finite sample sizes should also be considered in different SNR levels. To find $\mathbf{w}$, here we used the smallest eigenvector (with a corresponding eigenvalue on the order of 10^{-14} to 10^{-8}) of $\mathbf{P}$ to find the null-space of this rank $n-1$ matrix, and the bias term h^0 is obtained using $h^0 = E(d) - E(y)$ after adapting the other coefficients. We fixed the SNR of output measurements to 15 dB, and varied the SNR of the input measurements from 5 to 20 dB. Results of 100 Monte Carlo simulations are presented in Fig. 4.

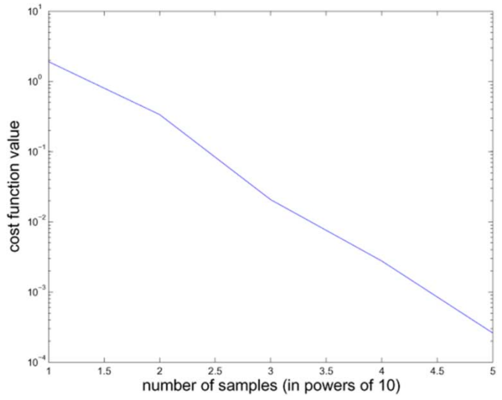

Fig. 3. Cost function value at $\mathbf{w} = \alpha\mathbf{h}$ versus number of samples.

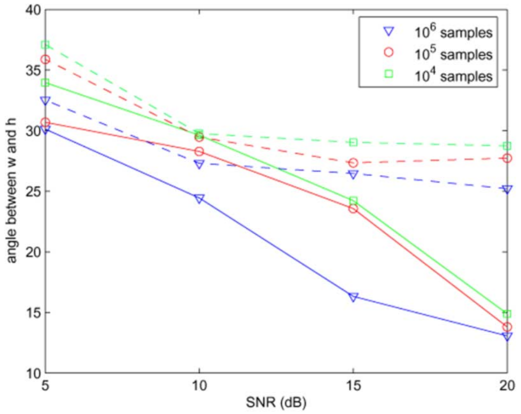

Fig. 4. Angle (degrees) between the actual and estimated parameter vectors for the proposed approach (solid lines) and least squares method (dashed lines) versus SNR.

Fig. 4 shows the angle between the estimated coefficient vector $\mathbf{w}$ and the actual coefficient vector $\mathbf{h}$ for the proposed method (solid lines) and least squares (dashed lines) for different orders of samples. For high SNR scenarios, higher number of samples does not significantly increase the estimation performance. However, as the SNR level decreases, much more samples are required to keep the performance.

IV. Discussion

In this letter, we address the problem of order-2 Volterra dynamic system model identification problem under the conditions of noisy input and noisy output measurements, exploiting the assumption that input and output noise processes are white and the true excitation input is colored. Traditionally and typically, the presence of noise in the input measurements is ignored in nonlinear model regression, which leads to biased solutions for model parameter estimates. Existing techniques in adaptive filtering that attempt to correct for this bias limit their attention to linear FIR filters. In this letter, we present a statistical criterion, which enables one to identify the true parameters of a nonlinear order-2 Volterra model from noisy data with reasonable assumptions of independence and stationarity regarding the stochastic processes involved. The extension of the technique to other generalized linear models, including higher order Volterra models, is the topic of future research. The fundamental departure in the set of assumptions exploited in the presented approach compared to those employed in Bayesian regression is the utilization of colored-excitation input versus iid sample requirements of typical Bayesian modeling approaches, which lead to simplification of likelihood calculations. In this letter, we explicitly exploit the fact that noise is distinguished from signals of interest by its temporal correlation structure.

This letter extends the unbiased model identification capability from linear (order-1) Volterra models to order-2; however, in its current form, it cannot be proven to work for higher order Volterra models. In typical nonlinear signal processing applications where Volterra models are used (e.g., modeling nonlinear amplifiers), typically low-order Volterra models are used (up to order-3). Since the model order is fixed *a priori* and sample training data is typically available, the bias-variance trade-off is not generally considered in these applications. The proposed criterion offers an analytical solution for the weight vector, expressed in the form of a linear system of equations; therefore, the implementation of LMS variants that solve for the optimal Volterra weights online can be easily developed. We provide the derivation of this approach in Appendix B, but due to limited space, we leave further discussion on this topic to a future publication.

APPENDIX A

Here we show that for an order-2 Volterra system and adaptive model and sufficiently large m [that is $m \geq (n-1)$], $\mathbf{w}$ is in the linear span of $\mathbf{h}$ if and only if $L(\mathbf{w})$ becomes zero.

($\Rightarrow$) Let $\mathbf{w} = \alpha\mathbf{h}$. For this case, we have

$$\begin{aligned} L_\Delta(\mathbf{w}) = \mathbf{h}^T E\left(\widetilde{\mathbf{x}}_t \mathbf{x}_{t-\Delta}^T - \widetilde{\mathbf{x}}_{t-\Delta}\mathbf{x}_t^T\right)\mathbf{w} \\ +\mathbf{w}^T E(u_t\mathbf{x}_{t-\Delta} - u_{t-\Delta}\mathbf{x}_t). \end{aligned} \quad (12)$$

Note that since $\mathbf{x}_t$ and u_t are independent and the latter is zero mean, the second term on the right-hand side vanishes to zero for any $\mathbf{w}$. The matrix in the first term can be decomposed into three components: 1) terms that depend only on the statistics of $\widetilde{x}_t$, 2) terms that depend only on the statistics of v_t, and 3) terms that depend on the mixed statistics of these two processes. The terms of types 2) and 3) vanish due to independence, ergodicity, stationarity, and zero-mean assumptions on v_t. To see this result, note that

$$\begin{aligned} \mathbf{x}_t &= \begin{aligned}&[\ldots(\widetilde{x}_{t-i}+v_{t-i})\ldots| \\ &\ldots(\widetilde{x}_{t-j}+v_{t-j})(\widetilde{x}_{t-k}+v_{t-k})\ldots]\end{aligned} \\ &= \begin{bmatrix} [\ldots\widetilde{x}_{t-i}\ldots|\ldots\widetilde{x}_{t-j}\widetilde{x}_{t-k}\ldots] \\ \quad +[\ldots v_{t-i}\ldots|\ldots v_{t-j}v_{t-k}\ldots] \\ \qquad +[\ldots 0\ldots|\ldots\widetilde{x}_{t-j}v_{t-k}\ldots] \end{bmatrix} \end{aligned} \quad (13)$$

where $i = 0,\ldots,L_1-1, j = 0,\ldots,L_{21}-1, k = 0,\ldots,L_{22-1}$. The expectation in the first term of (12) consists of three types of terms that correspond to the description above and indicated by the three terms on the right-hand side of (13) in the same order. The terms of type 1) remain as nonzero contributors. The terms of type 2) become zero after the subtraction of Δ and $-\Delta$ terms due to stationarity of v_t. The terms of type 3) become zero because of stationarity of v_t as well as its independence from $\widetilde{x}_t$ and it having zero mean. Consequently, the expression for L_Δ reduces to

$$L_\Delta(\mathbf{w}) = (\mathbf{h}^T(E(\widetilde{\mathbf{x}}_t\widetilde{\mathbf{x}}_{t-\Delta}^T) - E(\widetilde{\mathbf{x}}_{t-\Delta}\widetilde{\mathbf{x}}_t^T))\mathbf{w})^2. \quad (14)$$

Due to symmetry, $L_\Delta(\alpha\mathbf{h}) = 0$. Hence, $L(\alpha\mathbf{h}) = 0$.

($\Rightarrow$) Let $L(\mathbf{w}) = 0$. This implies that $\mathbf{P}$ has at least one zero-eigenvalue. It is assumed that m is sufficiently large, so we let $rank(\mathbf{P}) = (n-1)$. This implies that $\mathbf{w}$ must lie in the null-space of $\mathbf{P}$. From the previous part of the proof, we know that the line spanned by $\mathbf{h}$ is always contained in the null space of $\mathbf{P}$, and since the exact dimensionality of the null space is one, we must have $\mathbf{w} = \alpha\mathbf{h}$. QED

There is no upper bound on m other than computational constraints. When the adaptive weights are parallel to the true parameters of the unknown second-order Volterra system, all terms simultaneously diminish to zero, and including more delays improves the statistical variance of the weight estimates.

APPENDIX B

Based on the proposed criterion, one can derive a stochastic gradient algorithm, if an online training is necessary. Rewriting the cost function in (8) and taking its derivative with respect to $\mathbf{w}$, one obtains

$$\begin{aligned} L(\mathbf{w}) &= \sum_{i=1}^{m} L_{\Delta_i}(\mathbf{w}) \\ \frac{\partial\mathbf{L}}{\partial\mathbf{w}} &= 2\mathbf{P}\mathbf{w}. \end{aligned} \quad (15)$$

Hence, the update equation should be

$$\begin{aligned} \mathbf{w} &\leftarrow \mathbf{w} - \eta\frac{\partial\mathbf{L}}{\partial\mathbf{w}} \\ \mathbf{w} &\leftarrow \mathbf{w} - \eta\sum_{i=1}^{m}\mathbf{p}_{\Delta_i}\mathbf{p}_{\Delta_i}^T\mathbf{w} \end{aligned} \quad (16)$$

where η is the step size. To obtain the stochastic gradient-type algorithm, one should drop the expected value operators in $\mathbf{p}_{\Delta_i}$ and use the instantaneous values of d_t, $\mathbf{x}_{t-\Delta}$, $d_{t-\Delta}$, and $\mathbf{x}_t$. This yields

$$\mathbf{w} \leftarrow \mathbf{w} - \eta\sum_{i=1}^{m}(d_t\mathbf{x}_{t-\Delta} - d_{t-\Delta}\mathbf{x}_t)(d_t\mathbf{x}_{t-\Delta} - d_{t-\Delta}\mathbf{x}_t)^T\mathbf{w}. \quad (17)$$

ACKNOWLEDGMENT

The authors would like to thank P. P. Pokharel for valuable discussions.

REFERENCES

[1] S. Haykin, *Adaptive Filter Theory*. Upper Saddle River, NJ: Prentice-Hall, 1996.

[2] J. A. Cadzow, "Total least squares, matrix enhancement, and signal processing," *Digit. Signal Process.*, vol. 4, pp. 21–39, 1994.

[3] P. Lammerling, "Structured total least squares: Analysis, algorithms, and applications," Ph.D. dissertation, Katholeike Univ, Leuven, Belgium, 1999.

[4] T. Soderstorm and P. Stoica, *System Identification*. London, U.K.: Prentice-Hall, 1989.

[5] Y. N. Rao, D. Erdogmus, and J. C. Principe, "Error whitening criterion for adaptive filtering: Theory and algorithms," *IEEE Trans. Signal Process.*, vol. 53, no. 3, pp. 1057–1069, Mar. 2005.

[6] T. Ogunfunmi and S. L. Chang, "Second-order adaptive Volterra system identification based on discrete nonlinear Wiener model," *Proc. Inst. Elect. Eng., Vis., Image, Signal Process.*, vol. 148, no. 1, pp. 21–29, Feb. 2001.

[7] P. Koukoulas and N. Kalouptsidis, "Second order Volterra system identification," in *Proc. 8th IEEE Signal Processing Workshop on Statistical Signal and Array Processing (SSAP '96)*, 1996, pp. 383–383.