Separable Least Squares Identification of Wiener Box-Jenkins Models

Ibrahim A. Aljamaan Abdullah S. Bubshait
David T. Westwick

Department of Electrical and Computer Engineering, University of Calgary, Calgary, Alberta T2N 1N4, Canada,
e-mail: iaaljama, asbubsha, dwestwic@ucalgary.ca

Abstract: This paper presents an approach for the identification of a Wiener model, a dynamic linear system followed by a static nonlinearity, in the presence of colored measurement noise. A Box-Jenkins model structure is proposed where the process model consists of a recursive digital filter followed by a polynomial nonlinearity, while the noise model is represented by another recursive digital filter. The prediction error method is implemented using a separable least squares technique in order to estimate the parameters of the linear and nonlinear elements. The parameters of the digital filters are estimated using a second-order iterative optimization method since they appear nonlinearly in the output. After each iteration, the nonlinearity is fitted using linear regression. Monte-Carlo simulation is used to validate the algorithm.

Keywords: Box-Jenkins model structure, Wiener systems, Prediction error, Separable least squares.

1. INTRODUCTION

System identification, the construction of mathematical models of dynamic systems from measured data, is becoming more important in both industrial and research applications. Usually the main objective of the identification is to estimate the model using either parametric or nonparametric methods. Unlike real applications, most of the research to date has been focused on linear models (Landau and Zito, 2006). Wiener systems, where the linear model is cascaded with a memoryless nonlinearity (Giri and Bai, 2010), can accurately model many industrial systems, such as distillation columns (Bloemen et al., 2001) and flow control valves (Vall and Radhi, 2006).

In this paper we develop an algorithm for Wiener model identification in the presence of colored measurement noise. A Box-Jenkins model structure is used where the dynamic linear part of the Wiener cascade and the measurement noise model are represented by recursive digital filters, while the nonlinearity is represented by a polynomial function. The algorithm is based on a separable least squares optimization, where the parameters are separated into two sets: linear and nonparametric (Sjöberg and Viberg, 1997; Ruhe and Wedin, 1980; Westwick and Kearney, 2001). The parameters that appear nonlinearly in the predictor output are estimated using an iterative minimization method. After each iteration, a simple least squares regression is applied to get the optimal linear parameters corresponding to the current set of nonlinear parameters. Sjöberg and Viberg (1997) showed that using a separable least squares formulation can lead to better conditioned solutions and faster convergence than is possible when solving the unseparated optimization. The goal of this paper is to develop and test an identification algorithm for a Wiener Box-Jenkins model based on the principle of Separable Least Squares (SLS) optimization.

This paper is organized as follows: the next section contains the theoretical part which contains the problem description and the development of the proposed algorithm. Results from a Monte-Carlo simulation used to validate the algorithm and demonstrate its performance are shown in Section 3. Finally, the performance of the algorithm will be discussed in the conclusion.

2. THEORY

2.1 Problem Definition

The model that is being identified in this article is a single input single output (SISO) Wiener model that has a linear dynamic element followed by a static nonlinearity, as shown in Figure 1. The static nonlinearity is represented by a polynomial function. A Box-Jenkins model structure is used for the dynamics, thus the linear element in the Wiener model is represented by an infinite impulse response (IR) digital filter, \(G(q)\) and the noise is modeled by another IR digital filter, \(H(q)\). It is assumed that input and output, \(u(t)\) and \(y(t)\), respectively, are available, but \(x(t)\), \(y_d(t)\) and \(v(t)\) cannot be measured. The system used is described by

\[ y(t) = m(x(t)) + H(q)v(t) \] (1)

where the intermediate signal is computed as

\[ x(t) = G(q)u(t) \] (2)

* This work was supported by NSERC (Canada) and Ministry of Higher Education (Saudi Arabia).
Fig. 1. Block diagram of a Box-Jenkins Wiener model where
\[ G(q) = \frac{B(q)}{F(q)} \]
and the polynomials \( B(q), F(q), C(q) \) and \( D(q) \), are given as
\[
\begin{align*}
B(q) &= b_0 + b_1 q^{-1} + \cdots + b_n q^{-n_b} \\
F(q) &= 1 + f_1 q^{-1} + \cdots + f_n q^{-n_f} \\
C(q) &= 1 + c_1 q^{-1} + \cdots + c_n q^{-n_c} \\
D(q) &= 1 + d_1 q^{-1} + \cdots + d_n q^{-n_d}
\end{align*}
\] (3)
where \( q \) is the forward shift operator.

The nonlinearity \( m(\cdot) \) is represented using a polynomial of order \( M \):
\[
y_d(t, \theta) = m(x(t)) = \sum_{i=0}^{M} \gamma(i)x^i(t)
\] (4)
where \( x(t) \) is the input of the nonlinearity and \( \gamma(i) \) are the polynomial coefficients. Other basis expansions, such as orthogonal polynomials, or spline functions, could also be used.

3. IDENTIFICATION

3.1 Prediction Error Method

The one step ahead predictor of the Wiener model model is given as
\[
\hat{y}(t|t-1, \theta) = y_d(t, \theta) + \hat{v}(t|t-1, \theta)
\] (5)
where \( \theta \) is an \( n \times 1 \) vector containing the model parameters, \( y_d(t, \theta) \) is the output of the deterministic part, and \( \hat{v}(t|t-1, \theta) \) is the one step ahead prediction of the output error as computed in Ljung (1999):
\[
\hat{v}(t|t-1, \theta) = \frac{H(q) - 1}{H(q)} v(t, \theta) = (1 - H^{-1}(q)) v(t, \theta)
\] (6)
Substituting (6) into (5) and noting that \( v(t, \theta) = y(t) - y_d(t, \theta) \) will result in an expression of the predictor for the Box-Jenkins Wiener model as
\[
\hat{y}(t|t-1, \theta) = (1 - H^{-1}(q)) y(t) + H^{-1}(q) \sum_{i=0}^{M} \gamma(i)(G(q)u(t))'
\] (7)
This expression is identical to that for a linear Box-Jenkins model structure, except that the output of the linear plant model has been transformed by the nonlinearity (4). The prediction error of the model \( \epsilon(t) \) will be given by
\[
\epsilon(t, \theta) = y(t) - \hat{y}(t|t-1, \theta) = H^{-1}(q)(y(t) - y_d(t, \theta))
\] (8)
where \( y_d(t, \theta) \) is the output of the deterministic part of the system, given in (4).

The parameters can be obtained using parametric optimization technique that is based on minimizing the cost function, \( V_N(\theta) \) defined as
\[
V_N(\theta) = \frac{1}{2N} \sum_{t=1}^{N} \epsilon^2(t, \theta)
\] (9)
Thus, the goal is to find an estimate of the parameter vector, \( \theta \), such that
\[
\theta = \arg \min_{\theta} V_N(\theta)
\] (10)

3.2 Levenberg Marquardt Optimization Method

Since, the predictor \( \hat{y}(t|t-1) \) is nonlinear in the parameters, an iterative optimization method is required. Let \( \theta^{(j)} \) be the estimate of the parameter vector after \( j \) iterations. It will be updated using:
\[
\theta^{(j+1)} = \theta^{(j)} + d^{(j)}
\] (11)
where \( d^{(j)} \) is the direction and step size that is chosen to ensure \( V_N(\theta^{(j+1)}) \leq V_N(\theta^{(j)}) \). The direction \( d^{(j)} \) can be computed by a variety of techniques (Nocedal and Wright, 1999), including several quasi-Newton methods, such as the one used in this study.

These methods are based on expressing the cost function using a second-order Taylor expansion about the current parameter estimate, \( \theta^{(j)} \):
\[
V_N(\theta^{(j)} + d^{(j)}) = V_N(\theta^{(j)}) - \frac{1}{N} (v^T J) d^{(j)} + \frac{1}{2!} d^T(\partial^2 d) H d^{(j)}
\] (12)
where \( J \) is the Jacobian defined as the first partial derivative of the predictor output with respect to \( \theta \), and the Hessian \( H \) is the second-order partial derivative of the cost function \( V_N \) with respect to \( \theta \).

The value of \( d^{(j)} \) that minimizes \( V_N(\theta) \) is found by taking the derivative of (12) with respect to \( d^{(j)} \) and equating the result to zero. Thus
\[
d^{(j)} = H^{-1} J^T \epsilon
\] (13)
The Hessian matrix, \( H \) can be computed exactly, which results in the Newton method. However the Newton approach will produce a singular Hessian when it is applied to block-structure identification problems (Westwick and Kearney, 2003), as the models contain redundant degrees of freedom. Thus, the Levenberg-Marquardt method was used since it uses an approximate Hessian that is guaranteed to be invertible:
\[
H = J^T J + \delta_j I
\]
where $\delta_j$ is used to control the step size and adjust the convergence and $I$ is an $n \times n$ identity matrix. So, the search direction will be given as

$$d^{(j)} = (J^T J + \delta_j I)^{-1} J^T \epsilon$$

### 3.3 Separable Least Squares Technique

Instead of searching for all parameters using the Levenberg Marquardt optimization method (Ljung, 1999; Nocedal and Wright, 1999) a separation of parameters is applied to update the nonlinear parameters only. Since the coefficients, $\gamma^{(i)}$ in Eq. (4), appear linearly in the predictor output, they can be computed using linear regression. Thus, the parameter vector will be separated into two parts as

$$\theta = [\theta^T_n \theta^T_t]^T$$

where $\theta_n$ is a vector containing the numerator and denominator coefficients of the filters $G(q)$ and $H(q)$, as they appear nonlinearly in the output

$$\theta_n = [b_0 b_1 \cdots b_{nb} f_1 \cdots f_{nf} c_1 \cdots c_{nc} d_1 \cdots d_{nd}]$$

where $n_b, n_f, n_c$ and $n_d$ are the numbers of filter weights corresponding to $B(q)$, $F(q)$, $C(q)$, and $D(q)$ respectively, as defined in (3). The total number, $N_n$, of the parameters that appear nonlinearly in the output, is $N_n = n_b + n_f + n_c + n_d + 1$.

Similarly, the linear parameter vector, $\theta_t$, of length $N_t = M + 1$, contains the coefficients, $\gamma^{(i)}$, of the polynomial which appear linearly in the predictor output, and can therefore be estimated using linear regression. The output of the nonlinearity, shown in Figure 1, is computed using:

$$y_d(t, \theta) = X(\theta_n) \theta_t$$

where

$$X(\theta_n, t) = [1 \ x(\theta_n, t) \ x^2(\theta_n, t) \ \cdots \ x^M(\theta_n, t)]$$

If the noise is white the solution for $\theta_t$ is given by

$$\theta_t = (X^T(\theta_n) X(\theta_n))^{-1} X^T(\theta_n) y$$

where

$$X(\theta_n) = \begin{bmatrix}
X(\theta_n, 1) \\
X(\theta_n, 2) \\
\vdots \\
X(\theta_n, N)
\end{bmatrix}$$

and $y$ is the column vector containing all measurements $y(t)$, $t \in \{1, \ldots, N\}$. However, we considered colored noise. Therefore, we need to filter the columns of $X(\theta_n)$ and the vector $y(t)$ by $H^{-1}(q)$. Thus, the solution for $\theta_t$ will be

$$\theta_t = (X_f^T(\theta_n) X_f(\theta_n))^{-1} X_f^T(\theta_n) y_f$$

where

$$y_f(t) = \frac{D(q)}{C(q)} y(t)$$

and

$$X_f(t, k) = \frac{D(q)}{C(q)} x^{k-1}(\theta_n, t) : \{k = 1, \cdots, N_t\}$$

For updating the nonlinear parameters in the separated problem, $\theta_n^{(j+1)}$ can be rewritten as

$$\theta_n^{(j+1)} = \theta_n^{(j)} + d^{(j)}$$

The dependence of the linear parameters, $\theta_t$, on the nonlinear parameters, $\theta_n$, must be taken into account when computing the search direction in (20). To do this, the Jacobian for the unseparated problem is computed, and then split into its linear and nonlinear parts, as $J = [J_n \mid J_t]$, where the Jacobians $J_n$ and $J_t$ are obtained by taking the derivative of the predictor output with respect to $\theta_n$ and $\theta_t$ respectively while all other parameters remain fixed:

$$J_n(t,:) = \frac{\partial y(t|t-1, \theta_n)}{\partial \theta_n}$$

$$J_t(t,:) = \frac{\partial y(t|t-1, \theta_t)}{\partial \theta_t}$$

Next, define the projection onto the linear part

$$P_t = J_t(J_t^T J_t)^{-1} J_t^T$$

and the complementary projection

$$Q_t = I - P_t$$

The search step, $d^{(j)}$, of the separated problem is computed by projecting $J_n$ onto $Q_t$ (Sjöberg and Viberg, 1997):

$$J_n = Q_t J_n$$

To compute the Levenberg Marquardt step robustly (Ljung, 1999), first define

$$J_{0t} = \left[ J_t \right]$$

and

$$\epsilon_{\delta} = \begin{bmatrix} \epsilon \\ 0 \end{bmatrix}$$

and then solve the linear regression:

$$d^{(i)} = (J_{0t}^T J_{0t})^{-1} J_{0t}^T \epsilon_{\delta}$$

The Jacobians in (21) and (22) can be computed for the described Box-Jenkins model given in (1). Starting from Eq. (7), and substituting the elements of $G(q)$ and $H(q)$ will result in an expression of the predictor like

$$y(t|t-1, \theta) = \frac{C(q)}{D(q)} y(t) + \frac{D(q)}{C(q)} \sum_{i=0}^{M} \gamma^{(i)} \left( \frac{B(q)}{F(q)} u(t) \right)^i$$

Now, we take the derivative of the predictor with respect to each parameter of the polynomials $B(q)$, $F(q)$, $D(q)$ and $C(q)$ to form the nonlinear Jacobian matrix $J_n$. For the polynomial $B(q)$ the derivative of the first term is zero, since $y(t)$ contains measured data, and is therefore independent of the parameters, and the second is differentiated with respect to $b_k$. This gives
\[
\frac{\partial \hat{y}(t|t-1, \theta)}{\partial b_k} = \frac{D(q)}{C(q)} \left[ \sum_{i=1}^{M} i \gamma(i) \left( \frac{B(q)}{F(q)} u(t) \right)^{i-1} \right] \times q^{-k} \frac{B(q)}{F(q)} F(q) u(t) \quad (28)
\]

for \( k = 0, \ldots, n_b - 1 \).

Here, \( k \) is the order of the coefficient in the polynomial.

Similarly, taking the derivative of (27) with respect to \( F(q) \), \( D(q) \) and \( C(q) \) respectively, produce:

\[
\frac{\partial \hat{y}(t|t-1, \theta)}{\partial d_k} = \frac{D(q)}{C(q)} (y_d(t, \theta) - y(t)) \quad (30)
\]

for \( k = 0, 1, \ldots, n_d \)

\[
\frac{\partial \hat{y}(t|t-1, \theta)}{\partial c_k} = \frac{q^{-k} D(q)}{C(q)C(q)} (g(t) - y_d(t, \theta)) \quad (31)
\]

for \( k = 0, 1, \ldots, n_c \).

Therefore, the nonlinear Jacobian, \( J_n \), will have the form

\[
J_n(t,:) = \begin{bmatrix}
\frac{\partial \hat{y}(t|t-1, \theta)}{\partial b_0} & \ldots & \frac{\partial \hat{y}(t|t-1, \theta)}{\partial b_n} \\
\frac{\partial \hat{y}(t|t-1, \theta)}{\partial d_0} & \ldots & \frac{\partial \hat{y}(t|t-1, \theta)}{\partial d_n} \\
\frac{\partial \hat{y}(t|t-1, \theta)}{\partial c_0} & \ldots & \frac{\partial \hat{y}(t|t-1, \theta)}{\partial c_n}
\end{bmatrix}^T
\]

Now, to compute the Jacobian \( J_t \) take the derivative of Eq. (27) with respect to the polynomial coefficients \( \gamma^{(i)} \).

\[
\frac{\partial \hat{y}(t|t-1, \theta)}{\partial \gamma^{(i)}} = \frac{D(q)}{C(q)} \left( \frac{B(q)}{F(q)} u(t) \right)^i X_f(t,i) \quad (33)
\]

where \( X_f(t,i) \) is defined in (19), and \( i = 0, 1, \ldots, M \), where \( M \) is the order of the nonlinear polynomial. Thus,

\[
J_t = X_f \quad (34)
\]

**Algorithm 1:**

1. Initialize the nonlinear parameters \( \theta_n^{(0)} \).
2. Solve for \( \dot{x}(t) \) using (2) and then use (17) to estimate \( \theta_t^{(0)} \) by fitting a linear regression.
3. Set \( j = 0 \) and initialize the Levenberg-Marquardt ridge parameter \( d^{(0)} = 1 \), and a parameter that controls the convergence rate: \( \mu > 1 \).
4. Compute the Jacobians \( J_n \) and \( J_t \) using (21) and (22).
5. Compute \( d^{(j)} \) using (26), and update the nonlinear parameter vector by (20).
6. Solve for \( \dot{x}(t) \) using the new updates of \( \theta_n \), then use (17) to fit \( \theta_t^{(j+1)} \).
7. Estimate the error to compute the cost function using (8) and (9).
8. If \( V_N(\theta^{(j+1)}) < V_N(\theta^{(j)}) \), the step resulted in an reduction in the cost function. Therefore, keep the new parameter vector and increase the convergence speed by updating \( \delta = \frac{1}{2} \), otherwise put \( \delta = \delta_\mu \), to slow the convergence, and keep the previous parameter vector \( \theta^{(j+1)} = \theta^{(j)} \).
9. Set \( j = j+1 \) and go back to step 4 until the algorithm converges.

### 3.4 Initialization

Convergence to the global minimum is not guaranteed by any iterative optimization method which results in the importance of having a good initial estimate of the parameter vector. One of the major factors to have a successful identification is having a reasonable initial estimate of \( \theta \). If the input to the system is Gaussian then correlation-based methods, which are based on Bussgang’s theorem, can be applied to initialize the algorithm (Westwick and Kearney, 2003). Westwick and Verhaegen (1996) describe an Instrumental Variable (IV) based method for identifying Wiener systems that are excited by Gaussian inputs. These results may also hold for some other classes of inputs (Enqvist, 2003).

### 4. SIMULATION

To implement the proposed algorithm, a simulation has been performed for the model in Figure 1 using \( N=1000 \) data. The dynamic linear element is selected to be a second-order digital Butterworth filter with a normalized cutoff frequency of 0.25, so that \( G(q) \) will be given as

\[
G(q) = \frac{0.09763 + 0.1953q^{-1} + 0.09763q^{-2}}{1 - 0.9428q^{-1} + 0.3333q^{-2}}
\]

and the noise filter is

\[
H(q) = \frac{1 + 0.6q^{-1}}{1 - 0.9q^{-1}}
\]

A third-order power series was used for the nonlinearity:

\[
m(x(t)) = 1 + x(t) + x^2(t) + x^3(t)
\]

So, the parameter vector is given as in (14), where \( \theta_n = [0.09763 \ 0.1953 \ 0.09763 \ -0.9428 \ 0.3333 \ 0.6 \ -0.9] \) and

\[
\theta_t = [1 \ 1 \ 1 \ 1]
\]

When the linear element of the Wiener model is scaled by a constant, \( k \), scaling the \( i \)th degree nonlinear coefficient by \( k^i \), for \( i = 0 \ldots M \) results in no change to the system’s
output, or to its Volterra kernels (Westwick and Kearney, 2003). To avoid this ambiguity, the coefficients of \( B(q) \) were normalized by multiplying them by \( k = \frac{1}{\|B(q)\|_2} \), with 
\[
\theta_{B(q)} = [b_0 \ b_1 \ b_2] = [0.09763 \ 0.1953 \ 0.09763], \]
and every \( i^{th} \) coefficient of the polynomial describing the nonlinearity is multiplied by \( k \). Therefore the normalized parameters will be

\[
\theta_{n} = [0.4082 \ 0.8165 \ 0.4082 - 0.9428 \ 0.3333 \ 0.6 - 0.9]
\]
and

\[
\theta_{e} = [1 \ 0.2391 \ 0.0572 \ 0.0137]
\]

For comparison purposes, the same normalization was applied to all identified models. The input, \( u(t) \), and \( e(t) \) are white Gaussian signals. The signal to noise ratio (SNR) was 20 dB.

The identification was implemented using Algorithm 1, assuming that the length of the linear and nonlinear parameter vectors is known. The experiment was repeated 100 times and for each run the Levenberg-Marquardt optimization ran for 50 iterations. The nonlinear parameter vector, \( \theta_{e} \), was initialized each time using a simple Instrumental variable method where past inputs were used as instruments. An initial noise model was obtained by fitting an ARMA model to the residuals from the initial estimate of the Wiener system. The means of the estimated parameters and their standard deviations are shown in Table 1. The predictor output, computed using these parameters, is plotted in Figure 2. Figure 3 shows that the cost function converges in an average of 5 iterations. Comparison of the true and estimated nonlinearities is shown in Figure 4. For each parameter a histogram was plotted for the 100 runs in Figure 5 and 6.

Table 1. The normalized true parameters and the estimated parameters (mean & std) of the model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True</th>
<th>Estimated with std</th>
</tr>
</thead>
<tbody>
<tr>
<td>( b_0 )</td>
<td>0.4082</td>
<td>0.4082 ± 0.0046</td>
</tr>
<tr>
<td>( b_1 )</td>
<td>0.8165</td>
<td>0.8165 ± 0.0031</td>
</tr>
<tr>
<td>( b_2 )</td>
<td>0.4082</td>
<td>0.4081 ± 0.0069</td>
</tr>
<tr>
<td>( f_1 )</td>
<td>-0.9428</td>
<td>-0.9435 ± 0.0077</td>
</tr>
<tr>
<td>( f_2 )</td>
<td>0.3333</td>
<td>0.3336 ± 0.0063</td>
</tr>
<tr>
<td>( \gamma )</td>
<td>0.5</td>
<td>0.5028 ± 0.0296</td>
</tr>
<tr>
<td>( d_1 )</td>
<td>-0.45</td>
<td>-0.8471 ± 0.0183</td>
</tr>
<tr>
<td>( \gamma^{(0)} )</td>
<td>1</td>
<td>1.0033 ± 0.0316</td>
</tr>
<tr>
<td>( \gamma^{(2)} )</td>
<td>0.2391</td>
<td>0.2390 ± 0.0041</td>
</tr>
<tr>
<td>( \gamma^{(3)} )</td>
<td>0.0572</td>
<td>0.0572 ± 0.0009</td>
</tr>
<tr>
<td>( \gamma^{(4)} )</td>
<td>0.0137</td>
<td>0.0137 ± 0.0003</td>
</tr>
</tbody>
</table>

5. CONCLUSION

A Box-Jenkins Wiener model operating in open-loop was identified using Separable Least Squares, where the nonlinear parameters vector is separated from the linear one and it is updated using the Levenberg Marquardt optimization method. The parameters of the linear dynamic, \( B(q) \) and the coefficients, \( \gamma^{(i)} \), were normalized to compare them with the estimated parameters. It is clear from the results how efficient the method is. Using the separable least squares method reduces the dimension of the search space; as a result, the optimization problem becomes easier to solve. In our example we used 7 nonlinear parameters rather than using the whole set of 11 linear and nonlinear parameters. Also, it is observed that the time consumed to make 100 runs with 50 iterations was very short, around 16 seconds.

In addition to the Wiener open-loop model, this method could be extended and applied to Wiener Box-Jenkins model structure operating in closed-loop using the direct approach as reviewed by Forssell and Ljung (1999).
The histogram of estimated $b_0$

The histogram of estimated $b_1$

The histogram of estimated $b_2$

The histogram of estimated $f_1$

The histogram of estimated $f_2$

The histogram of estimated $c_1$

The histogram of estimated $d_1$

The histogram of estimated $\gamma_0$

The histogram of estimated $\gamma_1$

The histogram of estimated $\gamma_2$

The histogram of estimated $\gamma_3$

Fig. 5. Histogram of the nonlinear parameters $\theta_n = [b_0 \ b_1 \ b_2 \ f_1 \ f_2 \ c_1 \ d_1].$

Fig. 6. Histogram of the polynomial coefficients $\theta_l = \gamma^{(i)}, \{i = 0, 1, 2, 3\}.$

REFERENCES


