Proceedings of the 17th World Congress
The International Federation of Automatic Control
Seoul, Korea, July 6-11, 2008

Maximum Likelihood Identification of Wiener Models ⋆

Anna Hagenblad * **Lennart Ljung** ** **Adrian Wills** ***

* *Division of Automatic Control, Linköpings Universitet, SE-581 80 Linköping, Sweden (e-mail:* annah@isy.liu.se*)*
** *Division of Automatic Control, Linköpings Universitet, SE-581 80 Linköping, Sweden (e-mail:* ljung@isy.liu.se*)*
*** *School of Electrical Engineering and Computer Science, University of Newcastle, Callaghan, NSW, 2308, Australia (e-mail:* adrian.wills@newcastle.edu.au*)*

Abstract: The Wiener model is a block oriented model having a linear dynamic system followed by a static nonlinearity. The dominating approach to estimate the components of this model has been to minimize the error between the simulated and the measured outputs. We show that this will in general lead to biased estimates if there is other disturbances present than measurement noise. The implications of Bussgang's theorem in this context are also discussed. For the case with general disturbances we derive the Maximum Likelihood method and show how it can be efficiently implemented. Comparisons between this new algorithm and the traditional approach confirm that the new method is unbiased and also has superior accuracy.

1. INTRODUCTION

So called *block-oriented models* have turned out to be very useful for the estimation of non-linear systems. Such models are built up from linear dynamic systems and nonlinear static mappings in various forms of interconnection. These models are of interest both as reflecting physical realities and as approximations of more general systems. See, e.g. Schoukens et al. (2003) or Hsu et al. (2006) for some general aspects on block-oriented models.

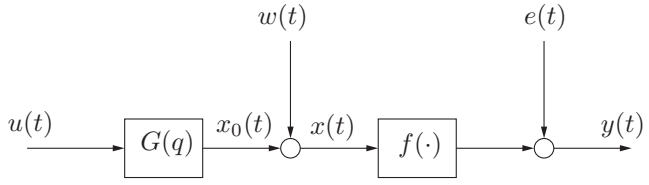

Fig. 1. The Wiener model. The input $u(t)$ and the output $y(t)$ are measurable, but not the intermediate signal $x(t)$. $w(t)$ and $e(t)$ are noise sources. $x_0(t)$ denotes the output of the linear dynamic system G. f is nonlinear and static (memoryless).

The Wiener model, Figure 1, is one such block oriented model. It describes a system where the first part is linear and dynamic, and the second part, in series with the first, is static and nonlinear. This is a reasonable model for, e.g., a distillation column (Zhu, 1999*a*) a pH control process (Kalafatis et al., 1995), biological examples (Hunter & Korenberg, 1986), or a linear system with a nonlinear measurement device. If the blocks are multi-variable, it can be shown (Boyd & Chua, 1985) that almost any nonlinear system can be approximated arbitrarily well by a Wiener model. In this paper, however, we focus on single input - single output systems.

We will use the notation defined in Figure 1. The input signal is denoted by $u(t)$, the output signal by $y(t)$ and $x(t)$ denotes the intermediate, unmeasurable signal. We will call $w(t)$ process noise and $e(t)$ measurement noise, and assume that they are independent. Note that since G is a linear system, the process noise can equally well be applied anywhere before the nonlinearity with an additional filter.

The Wiener system can be described by the following equations:

$$\begin{aligned} x_0(t) &= G(q,\theta)u(t) \\ x(t) &= x_0(t) + w(t) \\ y(t) &= f\big(x(t),\eta\big) + e(t) \end{aligned} \tag{1}$$

This paper will focus on parametric models. We will assume f and G each belongs to a parameterized model class. Examples of such a model class may be polynomials, splines, or neural networks for the nonlinear function f – in general a basis function expansion. The nonlinearity f may also be a piecewise linear function, like a saturation or a dead-zone. Common model classes for G are FIR filters, rational transfer functions (OE models) or state space models, but also for example Laguerre filters may be used.

If the intermediate signal x is unknown, then the parameterization of the Wiener model is not unique. A linear block G and a nonlinear block f gives the same complete system as a linear block KG in series with a nonlinear block $f(\frac{1}{K}\cdot)$. (We may also need to scale the process noise variance with a factor K.)

Given input and output data, and model classes for G and f, we want to find (estimate) the parameters θ and η that best match the data.

2. A STANDARD METHOD AND POSSIBLE BIAS PROBLEMS

Several different methods to identify Wiener models have been suggested in the literature. A common approach is to parameterize the linear and the nonlinear block, and to estimate the parameters from data, by minimizing an error criterion.

⋆ This work was supported by the Swedish Research Council and the Australian Research Council.

If the process noise $w(t)$ in Figure 1 is disregarded, or zero, a natural criterion is to minimize

$$V_N(\theta,\eta) = \frac{1}{N}\sum_{t=1}^{N}\Big(y(t) - f\big(G(q,\theta)u(t),\eta\big)\Big)^2 \qquad (2)$$

This is a standard approach, and has been used in several papers, i.e., Bai (2003), Westwick & Verhaegen (1996), Wigren (1993). It is also the method for Wiener models, used in available software packages like Ninness & Wills (2006) and Ljung (2007). If the process noise is indeed zero, this is the prediction error criterion. If the measurement noise is white and Gaussian, (2) is also the Maximum Likelihood criterion, see Ljung (1999), and the estimate is thus consistent.

While measurement noise e is discussed in several papers, few consider process noise w. Hunter & Korenberg (1986) is one exception where both the input and the output are subject to noise. Consistency of the estimation method is however not discussed in that paper. It may seem reasonable to use an error criterion like (2) even in the case where there is process noise. However, $f\big(G(q,\theta)u(t),\eta\big)$ is not the true predictor in this case. We will name this method the *approximative Prediction Error Method*, and we will show that the estimate obtained this way is not necessarily consistent.

2.1 Conditions for Consistency

Suppose that the true system can be described within the model class (cf. Figure 1), i.e., there exist parameters (θ_0,η_0) such that (c.f. Equation (1))

$$y(t) = f\big(G(q,\theta_0)u(t) + w(t),\eta_0\big) + e(t) \qquad (3)$$

An estimate from a certain estimation method is said to be *consistent* if the parameters converge to their true values when the number of data, N, tends to infinity.

To investigate the minimum of the approximative PEM criterion (2) we write the true system as

$$y(t) = f\big(G(q,\theta_0)u(t),\eta_0\big) + \tilde{w}(t) + e(t) \qquad (4)$$

where

$$\tilde{w}(t) = f\big(G(q,\theta_0)u(t)+w(t),\eta_0\big) - f\big(G(q,\theta_0)u(t),\eta_0\big) \quad (5)$$

We may regard $\tilde{w}(t)$ as a (input-dependent) transformation of the process noise to the output. The stochastic properties such as mean and variance of the process noise will typically not be preserved in the transformation from w to $\tilde{w}$.

Now insert the expression for y in Equation (4) into the criterion (2):

$$\begin{aligned} V_N(\theta,\eta) &= \frac{1}{N}\sum_{t=1}^{N}\Big(f_0 - f + \tilde{w}(t) + e(t)\Big)^2 \qquad (6)\\ &= \frac{1}{N}\sum_{t=1}^{N}\Big(f_0 - f\Big)^2 + \frac{1}{N}\sum_{t=1}^{N}\big(\tilde{w}(t)+e(t)\big)^2 \\ &\quad + \frac{2}{N}\sum_{t=1}^{N}\Big(f_0 - f\Big)\big(\tilde{w}(t)+e(t)\big) \end{aligned}$$

where

$$f_0 \triangleq f\big(G(q,\theta_0)u(t),\eta_0\big), \qquad f \triangleq f\big(G(q,\theta)u(t),\eta\big). \quad (7)$$

Now, assume all signals are ergodic, so that ensemble averages tend to their mathematical expectations as N tends to infinity. Assume also that u is a (quasi)-stationary sequence, so that is also has well defined sample averages. Let, E denote both mathematical expectation and averaging over time signals (cf. $\bar{E}$ in Ljung (1999)). Using the fact that the measurement noise e is zero mean, and independent of the input u and the process noise w means that several cross terms will disappear. The criterion then tends to

$$\begin{aligned} \bar{V}(\theta,\eta) = E\Big(f_0 - f\Big)^2 + E\tilde{w}^2(t) + Ee^2(t) \\ + 2E\Big(f_0 - f\Big)\tilde{w}(t) \qquad (8) \end{aligned}$$

The transformed process noise $\tilde{w}$, however, need not be independent of u, so the last term will not disappear.

Note that the criterion (8) has a quadratic form, and the true values (θ_0,η_0) will minimize the criterion if (and essentially only if)

$$E\Big(f_0 - f\Big)\tilde{w}(t) = 0 \qquad (9)$$

This condition typically does not need to hold, due to the possible dependence between u and $\tilde{w}$. The parameter estimates will thus be biased in general. This is illustrated by the simulation in Section 5.

2.2 Bussgang's Theorem and its Implication for Wiener Models

Bussgang's theorem (Bussgang, 1952) says the following:

Theorem 1. [**Bussgang**] Let $m(t)$ and $n(t)$ be two real-valued, jointly Gaussian stationary processes. Let $f(\cdot)$ be a nonlinear function and let the stochastic process $g(t)$ be defined by

$$g(t) = f(n(t))$$

Then the cross spectrum between m and n, $\Phi_{mn}(\omega)$, is proportional to the cross spectrum between m and g:

$$\Phi_{mg}(\omega) = \kappa\Phi_{mn}(\omega) \qquad (10)$$

where κ is a real-valued constant (that may be zero.)

This theorem has been applied to the estimation of Wiener model by many authors, e.g. (Westwick & Verhaegen, 1996), (Greblicki, 1994). It can be used to obtain a good estimate of the linear part of the model. It is interesting to note that the result applies also to our more general situation with process noise w as in Figure 1. In fact, we have the following Lemma:

Lemma 2. Consider the model structure defined by Figure 1. Assume that the the input $u(t)$ and the process noise $w(t)$ are independent, Gaussian, stationary processes (not necessarily white). Assume that the measurement noise $e(t)$ is a stationary stochastic process, independent of u and w (but not necessarily white nor Gaussian). Let $G(q,\theta)$ be an arbitrary transfer function parameterization with freely adjustable gain, such that $G(q,\theta_0) = G_0(q)$ (the true linear part of the system) for some parameter value θ_0. Let θ be estimated from u and y using an output error method, neglecting any possible presence of a nonlinearity:

$$\hat{\theta}_N = \arg\min_{\theta}\sum_{t=1}^{N}(y(t) - G(q,\theta)u(t))^2 \qquad (11)$$

Then

$$G(q,\hat{\theta}_N) \to \kappa G_0(q) \quad \text{as} \quad N\to\infty \qquad (12)$$

for some real constant κ (that may be zero).

Proof: (see Hagenblad et. el. (2007)).

The theorem is a consequence of the fact that the best linear system approximation (cf Ljung, 2001) that relates u to y is proportional to the linear part G_0 of the true system.

Basically this means that an estimate of the linear system $G(q)$ will be consistent (up to the gain) for many other common linear identification methods. Note that the gain of G cannot be estimated anyway, since a gain factor can be moved between G and f without affecting the input-output behavior.

Remark: It is essential that no noise model is built simultaneously with estimating G. The best linear description of the noise will be pretty complicated, since all the nonlinear effects are pushed to the residuals. This means that ARMAX, ARX and state-space models in innovations form that have common dynamics between noise and input, will not give an input dynamics model subject to (12). Also subspace methods, like N4SID (e.g. Van Overschee & DeMoor, 1996) will give biased results if they employ prediction horizons with past outputs. (More precisely, in the language of the system identification toolbox, (Ljung, 2007), the property N4HORIZON must be of the form [r,0,su].) This is in line with the use of MOESP as a subspace method in (Westwick & Verhaegen, 1996).

Having found G means that we know x_0 (up to scaling). If there is no process noise w so $x(t) = x_0(t)$, it is a simple problem to estimate the static nonlinearity $y(t) = f(x(t)) + e(t)$ from y and x.

However, if w is non-zero, the remaining problem to estimate f is still non-trivial: Find f from x_0 and y, where

$$y(t) = f(x_0(t) + w(t)) + e(t) \tag{13}$$

This is a nonlinear regression problem with disturbances affecting the regressors. The estimate of the parameters of f need not be consistent if simple methods are applied, as illustrated in Section 5.

3. MAXIMUM LIKELIHOOD ESTIMATION

3.1 Derivation of the Likelihood Function for White Disturbances

The likelihood function is the probability density function (PDF) of the outputs $y^N = \{y(1), y(2), ..., y(N)\}$ for given parameters θ and η. We shall also assume that the input sequence $u^N = \{u(1), u(2), ...u(N)\}$ is a given, deterministic sequence. (Alternatively, we condition the PDF wrt to this sequence, if it is described as a stochastic process.) Let $p_{y^N}(\theta, \eta; u^N)$ denote this PDF. For an observed data set y_*^N, the ML estimate is the one maximizing the likelihood function:

$$(\hat{\theta}, \hat{\eta}) = \arg\max_{\theta,\eta} p_{y^N}(\theta, \eta; Z_*^N) \tag{14}$$

where $Z_*^N = \{u^N, y_*^N\}$.

For the Wiener model (Figure 1) we first assume that the disturbance sequences $e(t)$ and $w(t)$ are white noises. This means that for given u^N, $y(t)$ will also be a sequence of independent variables. This in turn implies that the PDF of y^N will be the product of the PDFs of $y(t), t = 1, \dots, N$. It is thus sufficient to derive the PDF of $y(t)$. To simplify notation we shall use $y(t) = y, x(t) = x$ for short.

To find the PDF, we introduce the intermediate signal x as a nuisance parameter. The PDF of y given x is basically a reflection of the PDF of e, since $y(t) = f\big(x(t)\big) + e(t)$ It is easy to find if e is white noise:

$$p_y(y|x) = p_e\big(y - f(x, \eta)\big) \tag{15}$$

where p_e is the PDF of e.

The same is true for the PDF of x given u^N if w is white noise:

$$x(t) = G(q, \theta)u(t) + w(t) = x_0(t, \theta) + w(t) \tag{16}$$

With given u^N and θ, x_0 is a known, deterministic variable, so

$$p_x(x|u^N, \theta) = p_w\big(x - x_0(\theta)\big) = p_w\big(x - G(q, \theta)u(t)\big) \tag{17}$$

where p_w is the PDF of w.

Now by integrating over all $x \in \mathbf{R}$, we then eliminate this unmeasurable signal from our equations:

$$\begin{aligned} p_y(\theta, \eta; Z_*^N) &= \int_{x\in\mathbf{R}} p_{x,y}(x, y|\theta, \eta; u^N)dx \\ &= \int_{x\in\mathbf{R}} p_{y|x}(y|\theta, \eta, x; u^N)\, p_x(x|\theta, \eta; u^N)dx \\ &= \int_{x\in\mathbf{R}} p_e\big(y - f(x, \eta)\big)\, p_w\big(x - G(q, \theta)u(t)\big)dx \end{aligned} \tag{18}$$

We now assume that the process noise $w(t)$ and the measurement noise $e(t)$ are Gaussian, with zero means and variances λ_w and λ_e respectively, i.e.

$$p_e\big(\epsilon(t)\big) = \frac{1}{\sqrt{2\pi\lambda_e}} e^{-\frac{1}{2\lambda_e}\epsilon^2(t)}, \quad p_w\big(v(t)\big) = \frac{1}{\sqrt{2\pi\lambda_w}} e^{-\frac{1}{2\lambda_w}v^2(t)} \tag{19}$$

for each time instant t. Since the noise is white, the joint likelihood is the product over all time instants, and thus

$$\begin{aligned} p_y(y^N|\theta, \eta; u^N) &= \Big(\frac{1}{2\pi\sqrt{\lambda_e\lambda_w}}\Big)^N \prod_{t=1}^N \int_{-\infty}^{\infty} e^{-\frac{1}{2}E(t,\theta,\eta)}\, dx(t) \\ &= \Big(\frac{1}{2\pi\sqrt{\lambda_e\lambda_w}}\Big)^N \int_{x(1)=-\infty}^{\infty} \cdots \int_{x(N)=-\infty}^{\infty} e^{-\frac{1}{2}\sum_{t=1}^N E(t,\theta,\eta)}\, dx^N \end{aligned} \tag{20}$$

where

$$E(t, \theta, \eta) = \frac{1}{\lambda_e}\Big(y(t) - f\big(x(t), \eta\big)\Big)^2 + \frac{1}{\lambda_w}\big(x(t) - G(q, \theta)u(t)\big)^2 \tag{21}$$

Given data $Z_*^N = \{u_*^N, y_*^N\}$, we can calculate p_y and its gradients for each θ and η. This means that the ML criterion (14) can be maximized numerically.

We may also note that each integral in (20) depends on $x(t)$ for only one time instant t, so they can be computed in parallel.

If the noise covariances λ_w and λ_e are unknown, they can just be included among the parameters θ and η and their ML estimates are still obtained by (14). The derivation of the Likelihood function appeared in Hagenblad & Ljung (2000).

Note that if there is no process noise, then the above criterion reduces to (2).

3.2 Colored Noise

The following equations give the output:

$$\begin{aligned} x(t) &= G(q, \theta)u(t) + H_w(q, \theta)w(t) \\ y(t) &= f\big(x(t), \eta\big) + H_e(q, \eta)e(t) \end{aligned} \tag{22}$$

By using predictor form, see (Ljung, 1999), we may write this as

$$\begin{aligned} x(t,u^N,\theta) &= \hat{x}(t|x^{t-1},u^N,\theta) + w(t) \\ \hat{x}(t|x^{t-1},u^N,\theta) &\triangleq x(t) + H_w^{-1}(q,\theta)\big(G(q,\theta)u(t) - x(t)\big) \\ y(t) &= \hat{y}(t|y^{t-1},x^N,\eta) + e(t) \\ \hat{y}(t|y^{t-1},x^N,\eta) &= y(t) + H_e^{-1}(q,\eta)\big(f\big(x(t),\eta\big) - y(t)\big). \end{aligned} \tag{23}$$

The only stochastic parts are e and w. For a given sequence x^N, the joint PDF of y^N is obtained in the standard way, cf eq (5.74), Lemma 5.1, in Ljung (1999):

$$p_{y^N}(y^N|x^N) = \prod_{t=1}^{N} p_e(y(t) - \hat{y}(t|y^{t-1},x^N(u^N,\theta),\eta)) \tag{24}$$

By the same calculations, the joint PDF for x^N is

$$p_{x^N}(x^N) = \prod_{t=1}^{N} p_w(x(t) - \hat{x}(t|y^{t-1},u^N,\theta)) \tag{25}$$

The likelihood function for y^N is thus obtained from (24) by integrating out the nuisance parameter x^N using its PDF (25):

$$\begin{aligned} &p_{y^N}(y^N,\theta,\eta;u^N) = \\ &\int_{x^N\in\mathbf{R}^N} \prod_{t=1}^{N} p_w\big(H_w^{-1}(q,\theta)\Big(G(q,\theta)u(t) - x(t)\big)\Big) \times \\ &\qquad p_e\Big(H_e^{-1}(q,\eta)\big(f\big(x(t),\eta\big) - y(t)\big)\Big)\, dx^N \end{aligned} \tag{26}$$

In the case e and w are Gaussian, we obtain

$$\begin{aligned} &\prod_{t=1}^{N} p_e\big(H_w^{-1}(q,\theta)\left[x(t) - G(q,\theta)u(t)\right]\big) \\ &\times p_w\Big(H_e^{-1}(q,\eta)\left[y(t) - f\big(x(t),\eta\big)\right]\Big) = e^{-\frac{1}{2}\sum_{t=1}^{N} E(t,\theta,\eta)} \end{aligned} \tag{27}$$

similar to (20), where, this time,

$$\begin{aligned} E(t,\theta,\eta) = &\frac{1}{\lambda_w}\big(H_w^{-1}(q,\theta)\left[x(t) - G(q,\theta)u(t)\right]\big)^2 \\ &+ \frac{1}{\lambda_e}\Big(H_e^{-1}(q,\eta)\left[y(t) - f\big(x(t),\eta\big)\right]\Big)^2 \end{aligned} \tag{28}$$

Notice that this time filtered versions of $x(t)$ enter the integral, so the integration is a true multi-integral over all sequences x^N. This is hardly doable by direct integration in practice. It would then be interesting to evaluate the integration over x^N by probabilistic techniques.

4. IMPLEMENTATION

It was mentioned in Section 3 that numerical methods could be used to evaluate the likelihood integral in Equation (20), which could in turn be used as part of an iterative search procedure to find for the maximum likelihood estimate. While this may appear intractable, this section describes a practical algorithm for achieving the above.

In particular, we use a gradient-based iterative search method combined with numerical integration to form the ML estimate. The algorithm is profiled against the approximative PEM in Section 5 where the results from several Monte-Carlo simulations are discussed. It should be noted that the computation time for this algorithm is relatively modest and can easily be carried out on standard desktop computers.

In order to avoid numerical conditioning problems, we consider the equivalent problem of minimizing the negative log-likelihood function provided below.

$$(\hat{\theta},\hat{\eta},\hat{\lambda}_w,\hat{\lambda}_e) = \arg\min_{\theta,\eta,\lambda_w,\lambda_e} L(\theta,\eta,\lambda_w,\lambda_e) \tag{29}$$

where

$$\begin{aligned} &L(\theta,\eta,\lambda_w,\lambda_e) \triangleq -\log\big(p_y(\theta,\eta,\lambda_w,\lambda_e;Z_*^N)\big) \\ &= N\log(2\pi) + \frac{N}{2}\log(\lambda_w\lambda_e) - \sum_{t=1}^{N}\log\left(\int_{-\infty}^{\infty} e^{-\frac{1}{2}E(t,\theta,\eta)}dx\right) \end{aligned} \tag{30}$$

and $E(t)$ is given by Equation (21).

An essential element of solving (29) via gradient-based search is to have access to the gradient vector for a given value of the parameters

$$\vartheta \triangleq \begin{bmatrix}\theta^T & \eta^T & \lambda_w & \lambda_e\end{bmatrix}^T. \tag{31}$$

If we denote the gradient vector at iteration k of the search procedure as g_k, then the i'th element of g_k, denoted $g_k(i)$, is given by

$$\begin{aligned} g_k(i) = \Bigg[&\frac{N}{2}\frac{\partial\log(\lambda_w)}{\partial\vartheta(i)} + \frac{N}{2}\frac{\partial\log(\lambda_w)}{\partial\vartheta(i)} \\ &+\frac{1}{2}\sum_{t=1}^{N}\frac{\int_{-\infty}^{\infty}\frac{\partial E(t,\theta,\eta)}{\partial\vartheta(i)}e^{-\frac{1}{2}E(t,\theta,\eta)}dx}{\int_{-\infty}^{\infty}e^{-\frac{1}{2}E(t,\theta,\eta)}dx}\Bigg]\Bigg|_{\vartheta=\vartheta_k} \end{aligned} \tag{32}$$

This in turn requires that we compute the integrals in (30) and (32). Note, that the exponential term $\exp(-\frac{1}{2}E(t,\theta,\eta))$ appears in both these integrals, and the derivatives of $E(t,\theta,\eta)$ with respect to θ and η can be computed prior to evaluating the integral.

In general, evaluating these integrals will amount to approximating them via numerical integration methods, which is the approach used in this paper. In particular, we employ a fixed-interval grid over x and use the composite Simpson's rule to obtain the approximation (Press et al., 1992, Chapter 4). More generally however, the reason for using a fixed grid (not necessarily of fixed-interval as used here) is that it allows straightforward computation of $L(\vartheta_k)$ and its derivative g_k at the same grid point. Hence, a more elaborate approach might employ an adaptive numerical integration method that ensures the same grid points in calculating $L(\vartheta_k)$ and g_k.

Algorithm 2 details this computation and generates a number $\bar{L}$ and a vector $\bar{g}$ such that $L(\vartheta)\approx\bar{L}$ and $g(\vartheta)\approx\bar{g}$. For clarity, the algorithm is written as iterations over t and j, but these steps are not interdependent, and can be computed in parallel. The algorithm can also be extended to compute the Hessian if desired.

Algorithm 2:Numerical computation of likelihood and derivatives

Given and odd number of grid points M, the parameter vector ϑ and the data Z_*^N, perform the following steps.

NOTE: After the algorithm terminates, $L(\vartheta)\approx\bar{L}$ and $g(\vartheta)\approx\bar{g}$.

(1) Simulate the system $x_0(t) = G(\theta,q)u(t)$.
(2) Specify grid vector $\Delta\in\mathbb{R}^M$ as M equidistant points between the limits $[a\ b]$, so that $\Delta(1) = a$ and $\Delta(i{+}1) = \Delta(i){+}(b{-}a)/M$ for all $i = 1,\ldots,M-1$.
(3) Set $\bar{L} = N\log(2\pi) + \frac{N}{2}\log(\lambda_w\lambda_e)$, and $\bar{g}(i) = 0$ for $i = 1,\ldots,n_\vartheta$.

(4) **FOR t=1:N,**
(a) **FOR j=1:M,** compute

$$x = x_0(t) + \Delta(j), \tag{33}$$
$$\alpha = x - x_0(t), \tag{34}$$
$$\beta = y(t) - f(x, \eta), \tag{35}$$
$$\gamma_j = e^{-\frac{1}{2}(\alpha^2/\lambda_w + \beta^2/\lambda_e)}, \tag{36}$$
$$\delta_j(i) = \gamma_j \frac{\partial E(t)}{\partial \vartheta(i)}, \qquad i = 1, \ldots, n_\vartheta, \tag{37}$$

ENDFOR
(b) Compute (for each $i = 1, \ldots, n_\vartheta$ where necessary)

$$\kappa = \frac{(b-a)}{3M}\left(\gamma_1 + 4\sum_{j=1}^{\frac{M-1}{2}}\gamma_{2j} + 2\sum_{j=1}^{\frac{M-3}{2}}\gamma_{2j+1} + \gamma_M\right),$$

$$\pi(i) = \frac{(b-a)}{3M}\left(\delta_1(i) + 4\sum_{j=1}^{\frac{M-1}{2}}\delta_{2j}(i) + 2\sum_{j=1}^{\frac{M-3}{2}}\delta_{2j+1}(i) + \delta_M(i)\right),$$

$$\bar{L} = \bar{L} - \log(\kappa),$$
$$\bar{g}(i) = \bar{g}(i) + \frac{1}{2}\left(\frac{\partial \log(\lambda_w \lambda_e)}{\partial \vartheta(i)} + \frac{\pi(i)}{\kappa}\right),$$

ENDFOR

5. SIMULATION STUDY

In this study we considered a Wiener system in the form of Figure 1, where the static nonlinearity $f(\cdot)$ was chosen as a saturation. More precisely,

$$x_0(t) = G(q, \theta)u(t) \tag{38}$$
$$G(q, \theta) = \frac{1 + b_1 q^{-1} + b_2 q^{-2}}{1 + a_1 q^{-1} + a_2 q^{-2}} \tag{39}$$
$$x(t) = x_0(t) + w(t) \tag{40}$$
$$y(t) = f\big(x(t)\big) + e(t) \tag{41}$$
$$f\big(x(t)\big) = \begin{cases} c_1 & \text{for } x(t) \le c_1 \\ x(t) & \text{for } c_1 < x(t) \le c_2 \\ c_2 & \text{for } c_2 < x(t) \end{cases} \tag{42}$$

We conducted a Monte-Carlo simulation with 1000 data sets, and each set contained 1000 data points. In each case the input u and process noise w were sampled from a Gaussian distribution, with zero mean and unity variance, while the measurement noise e is Gaussian with zero mean and variance 0.1.

The parameter values $\{a_1, a_2, b_1, b_2, c_1, c_2\}$ were estimated using the two different methods described in the paper, namely the approximative PEM described in Section 2, and the ML method described in Section 3.

The prediction error criterion was minimized using the UNIT toolbox, (Ninness & Wills, 2006). To help avoid possible local minima, the search was initialized using the true values.

The ML implementation is described in Section 4. In addition to the system parameters, the noise covariances λ_w and λ_e were estimated. The parameter search was initialized with the results from the approximative PEM. The limits for the integration $[a, b]$ (see Algorithm 2) were selected as $\pm 6\sqrt{\lambda_w}$, which corresponds to a confidence interval of 99.9999 % for the signal $x(t)$ (at least when λ_w is estimated correctly). The number of grid points was chosen to be 1001.

The true values of the parameters, and the results of the approximative PEM and ML estimation are summarized in Table 1. The estimates of the nonlinear saturation function $f\big(x(t)\big)$ from Equation (42) are plotted in Figure 2.

Parameters	True	Approx. PEM	ML
a_1	0.3000	0.3007 ± 0.2059	0.3091 ± 0.1735
a_2	-0.3000	-0.2805 ± 0.2193	-0.2922 ± 0.1846
b_1	-0.3000	-0.2889 ± 0.1600	-0.2932 ± 0.1339
b_2	0.3000	0.3031 ± 0.1109	0.3034 ± 0.0947
c_1	-0.4000	-0.2932 ± 0.0212	-0.4005 ± 0.0206
c_2	0.2000	0.0997 ± 0.0198	0.2004 ± 0.0205
λ_w	1.0000	n.e.	0.9734 ± 0.2020
λ_e	0.1000	n.e.	0.1000 ± 0.0074

Table 1. Parameter estimates with standard deviations for Example 1, using approximative PEM and ML. The mean and standard deviations are computed over 1000 runs. The notation n.e. stands for "not estimated" as the noise variances are not estimated with the approximate PEM.

Parameters	True	Approx. PEM	ML
a_1	0.3000	0.2873 ± 0.1181	0.2980 ± 0.1000
a_2	-0.3000	-0.2852 ± 0.1224	-0.2980 ± 0.1058
b_1	-0.3000	-0.3005 ± 0.0886	-0.3009 ± 0.0752
b_2	0.3000	0.3046 ± 0.0672	0.3025 ± 0.0560
c_1	-0.4000	-0.3686 ± 0.0198	-0.4011 ± 0.0191
c_2	0.2000	0.1712 ± 0.0171	0.2011 ± 0.0166
λ_w	0.1765	n.e.	0.1724 ± 0.0540
λ_e	0.1000	n.e.	0.0995 ± 0.0057

Table 2. Parameter estimates with standard deviations for Example 1 with colored noise, using approximative PEM and ML. See Table 1 for details.

This simulation confirms that while a straightforward, approximative PEM gives biased estimates, the Maximum Likelihood method derived in this paper gives a consistent estimate of the system parameters, including noise variances, even when starting the numerical search in the biased estimate obtained from the approximative PEM. In these examples, also the variance of the approximative PEM estimates is larger than the estimates from the ML method.

As a further test, we were interested to know if the estimates were consistent in the ML case even if the process noise was colored. Therefore, we repeated the above simulation, but replaced the process noise with

$$w(t) = \frac{0.3q^{-1}}{1 - 0.7q^{-1}}\bar{w}(t), \tag{43}$$

where $\bar{w}(t)$ was sampled from a Gaussian distribution with zero mean and unity variance. The results are collected in Table 2 and show that the ML method generates consistent estimates while the approximate PEM method does not.

6. SUMMARY AND CONCLUSIONS

In the quite extensive literature on Wiener model estimation, the most studied method has been to minimize the criterion (2). We have called that approach the *Approximative Prediction Error Method* in this contribution. This method apparently is also the dominating approach for Wiener models in available software packages, like Ljung (2007) and Ninness & Wills (2006). We have in this contribution shown that this approach may lead to biased estimates in common situations. If disturbances

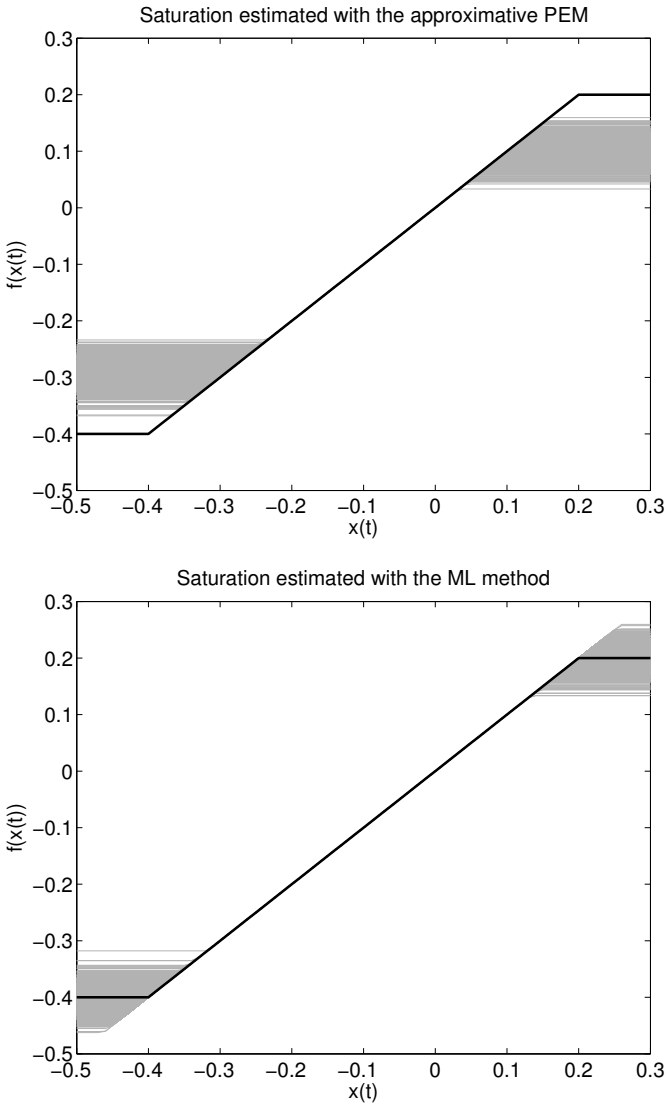

Fig. 2. Example 1: The true saturation curve as a thick black line and the 1000 estimated saturations, appearing as a grey zone. Top: approximative PEM. Bottom: ML.

are present in the system before the nonlinearity at the output, the estimates of the linear part and the nonlinearity will typically be biased, even when true descriptions are available in the model parameterizations. For example, Figure 2 clearly shows the bias in the estimate of an output saturation, in an otherwise ideal situation: Gaussian input, unbiased estimate of the linear part of the model. The reason for the bias is, in short, that the disturbances when transformed to the output error are no longer zero mean and independent of the input.

These deficiencies of the Approximative Prediction Error Method led us to a more serious statistical study of the Wiener model problem in the realistic case of both disturbances at the output measurements and process disturbances inside the dynamic part. We formulated the Likelihood function for the full problem. Although the maximization of this function at first sight may appear forbidding, an algorithm was developed that is not considerably more time-consuming than the Approximative Prediction Error Method. This ML method has the general property of consistency, which was also illustrated in the simulations.

In the general case of colored process noise, the Likelihood function is more complex to evaluate. However, in tests it has been found that the ML-method based on an assumption of white process noise produce consistent results also in the colored noise case. No proof of this observation has been established, though. A further challenge is to find efficient methods to evaluate the true likelihood function for this situation.

REFERENCES

Bai, E.-W. (2003), 'Frequency domain identification of Wiener models', *Automatica* **39**(9), 1521–1530.

Boyd, S. & Chua, L. O. (1985), 'Fading memory and the problem of approximating nonlinear operators with Volterra series', *IEEE Transactions on Circuits and Systems* **CAS-32**(11), 1150–1161.

Bussgang, J. J. (1952), Crosscorrelation functions of amplitude-distorted Gaussian signals, Technical Report 216, MIT Research Laboratory of Electronics.

Greblicki, W. (1994), 'Nonparametric identification of Wiener systems by orthogonal series', *IEEE Transactions on Automatic Control* **39**(10), 2077–2086.

Hagenblad, A. & Ljung, L. (2000), Maximum likelihood estimation of wiener models, *in* 'Proc. 39:th IEEE Conf. on Decision and Control', Sydney, Australia, pp. 2417–2418.

Hagenblad, A., Ljung, L. & Wills, A. (2008), Maximum Likelihood Identification of Wiener Models, *Automatica*, To appear.

Hsu, K., Vincent, T. & Poolla, K. (2006), A kernel based approach to structured nonlinear system identification part i: Algorithms, part ii: Convergence and consistency, *in* 'Proc. IFAC Symposium on System Identification', Newcastle.

Hunter, I. W. & Korenberg, M. J. (1986), 'The identification of nonlinear biological systems: Wiener and Hammerstein cascade models', *Biological Cybernetics* **55**, 135–144.

Kalafatis, A., Arifin, N., Wang, L. & Cluett, W. R. (1995), 'A new approach to the identification of pH processes based on the Wiener model', *Chemical Engineering Science* **50**(23), 3693–3701.

Ljung, L. (1999), *System Identification, Theory for the User*, second edn, Prentice Hall, Englewood Cliffs, New Jersey, USA.

Ljung, L. (2001), 'Estimating linear time invariant models of nonlinear time-varying systems', *European Journal of Control* **7**(2-3), 203–219. Semi-plenary presentation at the European Control Conference, Sept 2001.

Ljung, L. (2007), *The System Identification Toolbox: The Manual*, The MathWorks Inc. 1st edition 1986, 7th edition 2007, Natick, MA, USA.

Ninness, B. & Wills, A. (2006), An identification toolbox for profiling novel techniques, *in* '16th IFAC symposium on system identification'. `http://sigpromu.org/idtoolbox/`.

Nocedal, J. & Wright, S. J. (2006), *Numerical Optimization, Second Edition*, Springer-Verlag, New York.

Press, W. H., Teukolsky, S. A., Vetterling, W. A. & Fannery, B. P. (1992), *Numerical Recipes in C, the Art of Scientific Computing, Second Edition*, Cambridge University Press, Cambridge.

Schoukens, J., Nemeth, J. G., Crama, P., Rolain, Y. & Pintelon, R. (2003), 'Fast approximate identification of nonlinear systems', *Automatica* **39**(7), 1267–1274. July.

Van Overschee, P. & DeMoor, B. (1996), *Subspace Identification of Linear Systems: Theory, Implementation, Applications*, Kluwer Academic Publishers.

Westwick, D. & Verhaegen, M. (1996), 'Identifying MIMO Wiener systems using subspace model identification methods', *Signal Processing* **52**, 235–258.

Wigren, T. (1993), 'Recursive prediction error identification using the nonlinear Wiener model', *Automatica* **29**(4), 1011–1025.

Zhu, Y. (1999*a*), Distillation column identification for control using Wiener model, *in* '1999 American Control Conference', Hyatt Regency San Diego, California, USA.

Zhu, Y. (1999*b*), Parametric Wiener model identification for control, *in* '14th World Congress of IFAC', Beijing, China, pp. 37–42.