Application of Unscented Transformation in Nonlinear System Identification

Matej Gašperin ∗, Dani Juričić ∗

∗ Jožef Stefan Institute, Jamova 39, Ljubljana, Slovenia
(e-mail: matej.gasperin@ijs.si).

Abstract:
This article addresses the problem of system identification of nonlinear dynamical state-space models from input and output data. The problem is tackled by using Expectation Maximization (EM) algorithm for calculating the Maximum Likelihood (ML) estimates of the model parameters. The novelty of the presented algorithm is related to an efficient employment of the Unscented Transformation (UT) in the expectation step, which lowers the number of computations required. This property enables the algorithm to cope with high-dimensional models without a significant increase in computational load. The overall performance of the algorithm is demonstrated using both numerical and real data examples.

Keywords: Nonlinear System Identification; Maximum Likelihood Estimate; Unscented Transformation; Expectation-Maximization;

1. INTRODUCTION

System identification of nonlinear dynamic systems is recognized as a challenging problem (Ninness [2009]). The attention that the field has received in the past decades resulted in a broad spectrum of approaches (Ljung [2008]). In this paper, we follow the ideas laid down by Gibson and Ninness [2005], Schön et al. [2011], which is to employ the Expectation-Maximization (EM) algorithm for obtaining the Maximum-Likelihood (ML) estimate of the model parameters. The novelty in our work comes from the use of approximation techniques, namely Unscented Transformation, to compute the required conditional expectations.

The method presented in this paper can be applied to the identification of a very general class of systems, which can be represented in state-space form. The advantage of this formulation is that almost any system, usually described by a set of differential equations, can be transformed to the state-space representation.

In mathematical terms, the state-space model representation is of the following form:

\[ x_{t+1} = f(x_t, u_t, \theta) + w_t \]
\[ y_t = g(x_t, u_t, \theta) + v_t \]  

(1)

Here \( x_t \in \mathbb{R}^n \) is the state vector, \( y_t \in \mathbb{R}^m \) is the system output or measurement vector, \( u_t \in \mathbb{R}^p \) denotes the measured system input, \( f(\cdot) \) and \( g(\cdot) \) are nonlinear functions parameterized by parameter vector \( \theta \in \mathbb{R}^k \). Finally \( w_t \) and \( v_t \) are zero-mean, mutually independent stochastic processes, represented by probability density functions (pdf’s) \( p_w(\cdot) \) and \( p_v(\cdot) \). The pdf’s are assumed to be normal with unknown covariance matrices that can be included into \( \theta \) if necessary.

Different estimators for unknown model parameters can be constructed. This paper focuses on the Maximum-Likelihood (ML) estimators, due to their attractive properties, e.g. asymptotic unbiasedness, asymptotic minimum variance (Dekking et al. [2005]) as well as their general acceptance. This is a well established approach, especially in the case of linear dynamical systems, where different gradient based search techniques are commonly used to calculate the maximum of the data likelihood function (Ljung [1999]). An interesting alternative that has been presented in Gibson and Ninness [2005], Shumway and Stoffer [2005] is to adopt the Expectation-Maximization (EM) algorithm to compute the ML estimate of the model parameters.

The EM algorithm has so far been successfully applied to system identification of different system setups (McLachlan and Krishnan [2008]). In the case of linear models with additive Gaussian noise, it has been shown that a closed form expression for ML estimate of the model parameters can be formed (Shumway and Stoffer [2005]). This result has also been extended to bilinear systems by Gibson and Ninness [2005]. Furthermore, Andrieu et al. [2004] have considered the on-line estimation, using EM algorithm and split data.

On the other hand, the nonlinear models pose a far greater estimation problem. The problem arises from the fact, that a nonlinear transformation of a random variable may result in a posterior probability density function, that can not be described by a finite number of moments. Exact solutions of such problems only exist for a limited class of nonlinearities, such as Benes type (Elliott and Haykin [2010]). Therefore, any identification algorithm, for general nonlinear models, has to rely on approximations.

A versatile algorithm, capable to solve the problem, has been proposed recently by Wills et al. [2008], Schön.
et al. [2011]. The approach is based on so-called particle smoother and sequential Monte Carlo method, and therefore takes no prior constraints on model structure and noise distributions. Although versatile, the particle based approach may suffer from relatively large computational complexity as relatively large amount of particles might need to be propagated through the nonlinear functions in order to achieve sufficient accuracy.

To alleviate the problem of computational load, we introduce an algorithm based on Unscented Transformation (Julier [2002], van der Merwe [2004]). The approximations are used to solve the smoothing problem and also to approximate the conditional expectations of a nonlinear function. However, the introduction of approximation methods, such as unscented transformation, introduces an error in the estimate, which depends on the properties of the nonlinearities in the model (van der Merwe [2004]). The developed algorithm will therefore be tested under different setups, which demonstrate its efficiency and also highlight its limitations.

2. STATE-SPACE SYSTEM IDENTIFICATION

In this section we will address the following problem: given realization $Y_T = \{y_1, y_2, \ldots, y_T\}$ of the stochastic process, represented by a state-space model (1), with known structure of the nonlinear functions $f(\cdot)$ and $g(\cdot)$, our aim is to find the ML estimate of the unknown parameter vector $\theta$. Without the loss of generality, the algorithm will be derived under the assumption that the model has no measurable inputs.

2.1 Maximum likelihood estimate

Optimal values of the model parameters can be obtained in different manners, depending on the definition of optimality. One of them is the Maximum Likelihood (ML) criterion, which can be understood as the selection of parameter values $\hat{\theta} = \theta^*$, where the value of data likelihood function $p_\theta(Y)$ is a maximum. This estimate of $\theta$ is known as the maximum likelihood (ML) estimate. However, in most cases it proves to be more numerically efficient to search for the optimum of the log-likelihood function of the parameters,

$$L_\theta(Y_T) \triangleq \log p_\theta(Y_T).$$

Since log is a strictly increasing function of the argument, the value of $\theta$ that maximizes $L_\theta(Y_T)$, also maximizes $p_\theta(Y_T)$.

However, when the system output also depends on the vector of system states ($X_T = \{x_1, x_2, \ldots, x_T\}$), the direct maximization of log-likelihood function is not possible if the information about the state vector is not available. Classical maximum likelihood methods do not provide the solution to joint state and parameter estimation problem, as considered in this paper.

To include the state vector into the likelihood function we can use the multiplication rule (Dekking et al. [2005]) to derive the relation between classical likelihood function and joint likelihood function of the measurement data and the system state:

$$p_\theta(Y_T, X_T) = p_\theta(X_T|Y_T)p_\theta(Y_T)$$

In order to evaluate the measured data likelihood function $L_\theta(Y_T)$, it is clear that besides maximization of the function with respect to $\theta$, the estimation of state vector ($X_T$) is also required. The solution for this problem was introduced by Dempster et al. [1977]. The Expectation-Maximization algorithm was derived to solve the ML estimation problem using incomplete or missing data. If the states $X_T$ are considered as missing data, this algorithm can be successfully deployed to solve the described system identification problem.

2.2 Expectation-Maximization algorithm

EM algorithm solves the problem of simultaneously estimating system states and model parameters by alternating between two steps. Firstly, it approximates the likelihood function with its expected value over the missing data (E-step), and secondly, maximizes the likelihood function w.r.t. $\theta$ (M-step). A short overview of the algorithm will be presented, while a more detailed explanation can be found in Haykin [2001], Gibson and Ninness [2005].

Taking logarithms of (3) and rearranging it, we can write the following relation between classical and complete data log-likelihood function:

$$\log p_\theta(Y_T) = \log p_\theta(X_T, Y_T) - \log p_\theta(X_T|Y_T)$$

Taking the expectation of (4) with respect to the probability density function $p_{\theta_k}(X_T|Y_T)$, where $\theta_k$ is any fixed value of the parameter vector, we obtain

$$E_{\theta_k}\{\log p_\theta(Y_T)\} = \log p_\theta(Y_T) = \int \log p_\theta(X_T, Y_T)p_{\theta_k}(X_T|Y_T)dX_T - \int \log p_\theta(X_T|Y_T)p_{\theta_k}(X_T|Y_T)dX_T$$

or

$$L_\theta(Y_T) \triangleq Q(\theta, \theta_k) - \int \log p_\theta(X_T|Y_T)p_{\theta_k}(X_T|Y_T)dX_T$$

It has been shown by Gibson and Ninness [2005], Shumway and Stoffer [2005], that increasing the value of $Q(\theta, \theta_k)$ also increases the value of the likelihood function $L_\theta(Y_T)$. Indeed, using definition from (6) we write

$$L_\theta(Y_T) - L_{\theta_k}(Y_T) = Q(\theta, \theta_k) - Q(\theta_k, \theta_k) + \int \log \frac{p_\theta(X_T|Y_T)}{p_\theta(X_T|Y_T)}p_{\theta_k}(X_T|Y_T)dX_T$$

where the last term in (7) is the Kullback-Leibler divergence term and is by definition non-negative. Therefore,

$$L_\theta(Y_T) - L_{\theta_k}(Y_T) \geq Q(\theta, \theta_k) - Q(\theta_k, \theta_k)$$

from which it follows that increasing the value of $Q(\theta, \theta_k)$ will in fact increase the value of $L_\theta(Y_T)$.

The EM algorithm procedure can be summarized as follows:

**EM Algorithm**

1. Begin with an initial parameter vector estimate $\hat{\theta}_0$.
2. (E-step) Form the expected value of $p_\theta(X_T, Y_T)$ over the unknown system states $X_T$, based on a current parameter estimate $\hat{\theta}_k$ and system output $Y_T$.

$$Q(\theta, \theta_k) = \int \log p_\theta(X_T, Y_T)p_{\theta_k}(X_T|Y_T)dX_T$$

4429
(3) (M-step) Find a new estimate \( \theta_{k+1} \) by maximizing 
\[ Q(\theta, \theta_k) \] with respect to \( \theta \).
\[ \theta_{k+1} = \arg \max_\theta Q(\theta, \theta_k) \] (10)

(4) Repeat the procedure until convergence criteria are achieved.

3. THE E STEP: COMPUTING \( Q(\theta, \theta_k) \)

The first part of the algorithm is the Expectation step (E-step), that calculates the expected value of complete data log-likelihood function, or
\[
Q(\theta, \theta_k) = \int \log p_\theta(X_T, Y_T)p_{\theta_k}(X_T|Y_T)dX_T
= E_{\theta_k}\{p_\theta(X_T, Y_T)\}.
\] (11)
The problem can be divided into two sub-problems. The first one is to find the probability density function \( p_{\theta_k}(X_T|Y_T) \), where \( \theta_k \) is a fixed value parameter vector. This problem corresponds directly to solving the well-known smoothing problem (Sarkka [2008]). The second part concerns calculating the expected value of the function, over the distribution of the system states.

3.1 Discrete-time optimal smoothing

The problem of optimal discrete-time filtering and smoothing is well elaborated in the literature (Haykin [2001], Pillonetto and Bell [2008]). The algorithms are based either on Kalman filtering framework (Extended Kalman Filter, Unscented Kalman filter) or Monte-Carlo methods (particle filtering). The setup described in this work does not pose any explicit prior requirements about the algorithm selection.

In our work, we consider the forward-backward or Rauch-Tung-Striebel (RTS) type implementation of the smoother algorithms. The basic characteristic of the RTS type algorithms is, that it is composed of a forward filtering and backward smoothing pass. From Bayesian viewpoint, the filtering and smoothing pass can be described as follows:

1) Forward filtering equations:
\[
\text{Prediction step} \quad p_{\theta_k}(x_i|Y_{t-1}) = \int p_{\theta_k}(x_i|x_{t-1})p_{\theta_k}(x_{t-1}|Y_{t-1})dx_{t-1}
\] (12)
\[
\text{Update step} \quad p_{\theta_k}(x_i|Y_t) = \frac{p_{\theta_k}(y_t|x_i)p_{\theta_k}(x_t|Y_{t-1})}{p_{\theta_k}(y_t|Y_{t-1})}
\] (13)

2) Backward smoothing equations:
\[
\text{Joint smoothed density} \quad p_{\theta_k}(x_{t+1}, x_i|Y_T) = p_{\theta_k}(x_i|Y_T) \times \frac{p_{\theta_k}(x_{t+1}|x_i)p_{\theta_k}(x_t|Y_T)}{p_{\theta_k}(x_{t+1}|Y_T)}
\] (14)
\[
\text{Smoothed density} \quad p_{\theta_k}(x_i|Y_T) = \int p_{\theta_k}(x_{t+1}, x_i|Y_T)dx_{t+1}
\] (15)

In the case of nonlinear systems, closed form expressions of the described procedure cannot be derived. Therefore, smoother algorithms rely on various approximation methods. In the presented work, we use the Unscented RTS Smoother (Sarkka [2008]), which uses the Unscented transformation to approximate the statistics of a random variable, that undergoes a nonlinear transformation using a set of deterministically selected points (called sigma points). The points are selected in such a way, that they correctly capture the posterior distribution up to the second term in the Taylor series expansion of the nonlinear function.

After forward and backward run of the smoother algorithm we obtain the approximations to the state distributions \( p_{\theta_k}(x_i|Y_T) \) and joint state distributions \( p_{\theta_k}(x_{t+1}, x_i|Y_T) \), at some value of the parameter vector \( \theta_k \) for time index \( t \in \{1, 2, \ldots, T\} \).

3.2 Novel result: Approximating \( Q(\theta, \theta_k) \) function by UT

Given the distribution of the state vector with respect to a certain parameter estimate \( \theta_k \), next step is to calculate the expected value of the expression (11). However, for most nonlinear systems, an analytical expression for the integral can not be derived. Thus, most approaches rely either on the gradient based approximations or on Monte-Carlo techniques. The main idea of the Monte-Carlo based methods can be presented by the following equation for a nonlinear function \( f(\cdot) \):
\[
E\{f(x)\} = \int f(x)p(x)dx \approx \frac{1}{N} \sum_{i=1}^{N} f(x^i)
\] (16)

where \( x^i \) are i.i.d. samples form the probability distribution \( p(x) \). In our case, this relates to sampling from the state probability density function \( p_{\theta_k}(X_T|Y_T) \). Unfortunately, this type of sampling is very inefficient for large data (Gopaluni [2008]), as the number of samples \( N \) needs to be relatively large, which in consequence drastically affects the computational load of the algorithm. To overcome this issue, we propose an approximation based on the Unscented Transformation (Julier [2002], van der Merwe [2004]). Unscented Transformation is an approximate method of calculating the statistics of a random variable that undergoes a nonlinear transformation.

3.3 Unscented Transformation

The main idea behind UT is, that it is easier to approximate the distribution, rather than a nonlinear function. Let \( x \) be an \( n \) dimensional random variable with probability density function \( p(x) \). First, select a set of weighted samples \( \{x_i, w_i\} \) that correctly capture the first two moments (mean \( \bar{x} \) and covariance \( P_x \)) of the random variable \( x \). The selection scheme, proposed by Julier [2002], that satisfies this requirement, is as follows:
\[
X_0 = \bar{x} \quad \text{for } i = 0
\]
\[
X_i = \bar{x} + \left( \sqrt{(n + \kappa)P_x} \right)_i \quad \text{for } i = 1, \ldots, n
\]
\[
X_i = \bar{x} - \left( \sqrt{(n + \kappa)P_x} \right)_i \quad \text{for } i = n + 1, \ldots, 2n
\] (17)
and the corresponding weights as
\begin{align*}
  w_0 &= \frac{\kappa}{n + \kappa} & i = 0 \\
  w_i &= \frac{\kappa}{2(n + \kappa)} & i = 1, \ldots, n \\
  w_i &= \frac{\kappa}{2(n + \kappa)} & i = n + 1, \ldots, 2n
\end{align*}

where \( w_i \) is the weight associated with \( i \)-th sigma point and satisfies the condition \( \sum_{i=0}^{2n} w_i = 1 \). \( \sqrt{(n + \kappa)} \) is the \( i \)-th column of the matrix square root of the weighted covariance matrix and \( \kappa \) is the scaling parameter. Each sigma point \( X_i \) is then propagated through the nonlinear function

\[ Y_i = f(X_i) \quad i = 0, \ldots, 2n \]  

and the approximated mean and covariance of \( y \) are obtained as follows:

\[ \hat{y} \approx \sum_{i=0}^{2n} w_i Y_i \]

\[ \Sigma_p \approx \sum_{i=0}^{2n} w_i (Y_i - \hat{y})(Y_i - \hat{y})' \]

The posterior mean value and covariance, obtained using this approximation, are accurate to the second order of the Taylor series expansion of \( f(x) \) for any nonlinear function (van der Merwe [2004]). However, depending on the type of nonlinearities, these errors can still prevail and in some cases, the approximation errors can be relatively large (Hendeby and Gustafsson [2008]).

In our identification problem, the nonlinear function reads as follows:

\[
\log p_{\theta}(X_T, Y_T) = \log p_{\theta}(X_1) + \sum_{t=1}^{T-1} \log p_{\theta}(X_{t+1}|X_t) \\
+ \sum_{t=1}^{T} \log p_{\theta}(y_t|X_t),
\]

which is expanded using Markov property of the model and Bayes’ rule.

The function for E-step of the EM procedure \( Q(\theta, \theta_k) \) (11) can be viewed as the expected value, or first moment of logarithm of the posterior distribution \( p_{\theta}(X_T, Y_T) \). To approximate the value of \( Q(\theta, \theta_k) \) for some prior parameter estimate \( \theta_k \) we select a set of points \( X_i^k \) along with corresponding weights \( w_i^k \), based on the estimated distribution \( p_{\theta_k}(X_T|Y_T) \). Taking expectations of (21) yields

\[ Q(\theta, \theta_k) \approx \int \log p_{\theta}(X_1) p_{\theta_k}(X_1|Y_T) \left\{ \log p_{\theta}(X_T, Y_T) \right\} dx_1 \]

\[ = \int \log p_{\theta}(x_1)p_{\theta_k}(x_1|Y_T) dx_1 \\
+ \sum_{t=1}^{T-1} \int \log p_{\theta}(x_{t+1}|x_t)p_{\theta_k}(x_{t+1}, x_t|Y_T) dx_{t+1} dx_t \\
+ \sum_{t=1}^{T} \int \log p_{\theta}(y_t|x_t)p_{\theta_k}(x_t|Y_T) dx_t
\]

which can be approximated using unscented transformation based on (20) as

\[ Q(\theta, \theta_k) \approx \sum_{i=0}^{2n} w_i^k \log p_{\theta}(X_i^k) \\
+ \sum_{t=1}^{T-1} \sum_{j=0}^{4n} w_i^k \log p_{\theta}(X_i^k|X_{i-1}) \\
+ \sum_{t=1}^{T} \sum_{j=0}^{2n} w_i^k \log p_{\theta}(y_t|X_i^k)
\]

Where \( X_i^k \) are the sigma points representing the distribution \( p_{\theta_k}(x_1|Y_T) \), similarly \( [X_{i+1}^k, X_i^k] \) represent the joint distribution \( p_{\theta_k}(x_{t+1}, x_t|Y_T) \) and \( X_i^k \) represent \( p_{\theta_k}(x_t|Y_T) \).

The resulting expression (23) presents an approximation to the function \( Q(\theta, \theta_k) \) (22). The number or calculations required for approximating \( Q(\theta, \theta_k) \) depends only on the dimension of the state vector \( N \). This means that at every time step \( t \) only a fixed number of computations is required. The computational complexity of the step is thus \( O(4N + 1) \). Compared to the computational complexity of Monte Carlo based approximation, which is \( O(M) \), where \( M \) is the number of samples used and \( M \propto N^2 \), this represents a significant decrease of complexity, especially as the dimension of the problem \( N \) increases.

In this section, the performance of the algorithm will be presented using a simple numerical example and a more demanding high-order model.

4. EXAMPLES

In this section, the performance of the algorithm will be presented using a simple numerical example and a more demanding high-order model.

4.1 Nonlinear System

The model, taken here as a general nonlinear example, is adopted from previous studies of similar algorithms (Schönh et al. [2011]). It is described by the following nonlinear system:

\[
\begin{bmatrix}
x_{t+1} \\
y_t
\end{bmatrix} =
\begin{bmatrix}
a x_t + b \frac{x_t}{1 + x_t^2} + c \cos(1.2t) + w_t \\
d x_t^2 + v_t
\end{bmatrix}
\]

\[
\begin{bmatrix}
w_t \\
v_t
\end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ R \end{bmatrix}, \begin{bmatrix} Q & 0 \\ 0 & 0 \end{bmatrix} \right)
\]

(24)

It has been chosen for being acknowledged as a challenging estimation problem (Doucet et al. [2000]). The true parameter values are chosen as follows

\[
\theta = [a, b, c, d, Q, R] = [0.5, 2, 8, 0.05, 0, 0.001]
\]

(25)

To analyze the performance of the algorithm, a Monte-Carlo analysis has been performed, using 100 different data realizations \( Y_T \) with the length \( N = 100 \) samples. At the beginning of every estimation procedure, the initial values of the parameter vector \( \theta \) has been selected randomly form uniform distribution on interval equal to \( \pm 50\% \) of the corresponding true parameter values.
The resulting estimates of parameters $a$ and $b$ as a function of iteration number are shown in figure 1 and final estimates of all the parameters are given in table 4.1.

The model output, using the estimated parameter values, is shown in Figure 4, along with the 95% confidence interval. The values of other parameters are known process constants and can be found in the work by Stake et al. [2006].

The estimation procedure consisted of 20 iterations and has been repeated 25 times with different initial parameter values selected from the uniform distributions on the intervals:

$\gamma_{NH0} \in [200 \, 400]; \quad K_{NH0} \in [4 \, 6]; \quad k_{10} \in [0.5 \, 1]; \quad k_{20} \in [2.5 \, 4];$

Figure 3 shows parameter estimates as a function of iteration number. Table 2 summarizes the results in terms of mean parameter estimate value.

The vector of the unknown model parameters is $\theta = [\gamma_{NH} K_{NH} k_1 k_2]$, while the process and measurement covariance values have been set to $Q = 0.1$ and $R = 0.01$. The values of other parameters are known process constants and can be found in the work by Stake et al. [2006].

The estimation procedure consisted of 20 iterations and has been repeated 25 times with different initial parameter values selected from the uniform distributions on the intervals:

$\gamma_{NH0} \in [200 \, 400]; \quad K_{NH0} \in [4 \, 6]; \quad k_{10} \in [0.5 \, 1]; \quad k_{20} \in [2.5 \, 4];$

Figure 3 shows parameter estimates as a function of iteration number. Table 2 summarizes the results in terms of mean parameter estimate value.
Table 2. Mean values and variance of the final parameter estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>( r_{NH} )</th>
<th>( K_{NH} )</th>
<th>( k_1 )</th>
<th>( k_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML estimate mean</td>
<td>386</td>
<td>1.52</td>
<td>0.36</td>
<td>5.33</td>
</tr>
<tr>
<td>ML estimate variance</td>
<td>106.3</td>
<td>0.29</td>
<td>0.04</td>
<td>0.43</td>
</tr>
</tbody>
</table>

Fig. 4. Model output mean and confidence interval along with the measured output.

interval. The results suggest that the model using the estimated parameter values is able to correctly capture the process dynamics. These errors are partially a consequence of the approximations used in the estimation procedure. However, different estimation procedures provide similar results [Stare et al. (2006), Azman and Kocijan (2007)], which suggest that the real plant process includes additional dynamics, that are not modeled in the simplified nonlinear model, as considered in this example.

Finally, it is important to note that, in this case, only 17 sigma points \( (x_i) \) have been used to estimate the statistics of the posterior distributions, involved in the evaluation of \( (23) \). As the value of \( (23) \) has to be evaluated many times during the maximization step, this feature can be a crucial advantage of the presented algorithm, especially when considering real-life applications.

5. CONCLUSION

The contribution of this paper is a novel procedure for approximation based EM approach to nonlinear system identification. The main advantage of the algorithm is that it can successfully deal with high-order systems due to the use of computationally efficient approximations.

The overall robustness and accuracy of the estimates of this type of approximation is generally lower than of the approaches based on particle filters, e.g. in Schönh et al. [2011]. However, when nonlinearities in the system are not extremely severe, this method is a valuable tool, especially in applications, where computational load needs to be kept low. Further investigation in the tuning of the UT design parameters would hopefully contribute to the reduction of the estimation error.

REFERENCES