Simulated Pseudo Maximum Likelihood Identification of Nonlinear Models

Mohamed Rasheed Abdalmoaty and Håkan Hjalmarsson *

* *Automatic Control Department and ACCESS Linnaeus Center, KTH Royal institute of Technology, Stockholm, Sweden (e-mail: mohamed.abdalmoaty, hakan.hjalmarsson@ee.kth.se).*

Abstract: Nonlinear stochastic parametric models are widely used in various fields. However, for these models, the problem of maximum likelihood identification is very challenging due to the intractability of the likelihood function. Recently, several methods have been developed to approximate the analytically intractable likelihood function and compute either the maximum likelihood or a Bayesian estimator. These methods, albeit asymptotically optimal, are computationally expensive. In this contribution, we present a simulation-based pseudo likelihood estimator for nonlinear stochastic models. It relies only on the first two moments of the model, which are easy to approximate using Monte-Carlo simulations on the model. The resulting estimator is consistent and asymptotically normal. We show that the pseudo maximum likelihood estimator, based on a multivariate normal family, solves a prediction error minimization problem using a parameterized norm and an implicit linear predictor. In the light of this interpretation, we compare with the predictor defined by an ensemble Kalman filter. Although not identical, simulations indicate a close relationship. The performance of the simulated pseudo maximum likelihood method is illustrated in three examples. They include a challenging state-space model of dimension 100 with one output and 2 unknown parameters, as well as an application-motivated model with 5 states, 2 outputs and 5 unknown parameters.

Keywords: System identification, Nonlinear systems, Stochastic systems, Monte Carlo method.

1. INTRODUCTION

In this paper, we consider the problem of parameter identification for nonlinear models. Its major difficulty stems from the intractability of the likelihood function. The likelihood function is crucial for many favored statistical inference methods. For example, the Maximum Likelihood Estimator (MLE) is defined as its global maximizer and Bayesian estimators use it to construct the posterior distribution. On the other hand, the classical Prediction Error Methods (PEM) [Ljung 1999] rely indirectly on the likelihood function for the construction of optimal predictors.

Several numerical approximation methods have been recently suggested in the literature. A class of algorithms that attracted some attention relies on sequential Monte-Carlo algorithms, see for example Hürzeler and Künsch [2001], Ionides et al. [2006], Olsson et al. [2008], Malik and Pitt [2011], Schön et al. [2011], Poyiadjis et al. [2011], DeJong et al. [2012], Lindsten et al. [2014], and Doucet et al. [2015]. These algorithms employ particle filters and/or smoothers to approximate the involved high-dimensional integrals. However, they come with major computational demand and several numerical issues, see for example Snyder et al. [2008] and Kantas et al. [2015].

On the other hand, the Pseudo Maximum Likelihood (PML) method relies on a misspecified likelihood function. This is a rather old method which is often used to construct a simpler estimation problem or closed-form estimators. It can be traced back to Besag [1975] and its properties have been analyzed for parametric conditional models in Gouriéroux et al. [1984]. The pseudo likelihood is usually constructed using the first moment(s) of the model. Under some conditions, it can provide consistent and asymptotically normal estimators. Unfortunately, for models with intractable likelihood functions, those first few moments happen to be intractable. However, Monte-Carlo approximations of the moments can be used as suggested in Gouriéroux and Monfort [1990] without changing the asymptotic properties of the resulting estimators. In this case, the resulting estimator is known as the Simulated Pseudo Maximum Likelihood (SPML) estimator.

In this contribution, we show that the SPML method, when based on a multivariate normal family, is equivalent to a PEM with a parameterized norm and an implicitly defined predictor that is linear in the observations. Interestingly, we observe that this method behaves similarly to the use of an Ensemble Kalman Filter (EnKF) when solving a similar PEM problem. The EnKF is a Monte-Carlo implementation of the Kalman recursions. It is well suited for high-dimensional state-space applications, see Gillijns et al. [2006], Le Gland et al. [2009] and Roth et al. [2015]. It has also been used for combined state and parameter estimation [Evensen 2009], but was not considered in the PEM framework before. Its sequential nature can be advantageous from a computational point of view. We also note that, even though the SPML estimator is not asymptotically efficient, it can provide computationally cheap

⋆ This work was supported by the Swedish Research Council under contracts 2015-05285, 2016-06079 and by the European Research Council under the advanced grant LEARN, contract 267381.

consistent estimators. It can be further improved by one Gauss-Newton step to match the asymptotic properties of an efficient estimator, like the MLE for example. For this, we only need one evaluation of the score function.

The outline of the paper goes as follows. In Section 2, we fix the notation, set the assumptions and define the estimation problem. In Section 3, the simulated pseudo maximum likelihood method is introduced. Section 4 analyses the connection to the prediction error methods and compares the predictor defined by the SPML to the predictor defined using an EnKF. Section 5 investigates the behavior of the method on three numerical examples. Finally, the paper is concluded in Section 6.

2. PROBLEM DEFINITION

In this paper, we work in discrete-time. The inputs and outputs of the system at time k will be denoted by u_k and y_k respectively. Both are allowed to be vectors of arbitrary finite dimension. To simplify the notations, we stack both the known inputs u_k and the observed outputs y_k for $k = 1, \dots, N$ in column vectors as follows

$$Y := \begin{bmatrix} y_1^T & y_2^T & \dots & y_N^T \end{bmatrix}^T, \quad U := \begin{bmatrix} u_1^T & u_2^T & \dots & u_N^T \end{bmatrix}^T.$$

We consider a general stochastic nonlinear dynamical model for which y_k depends nonlinearly on the past known inputs and the history of an unobserved (latent) stochastic process $\{z_k\}$. The unobserved stochastic process models all sources of randomness affecting the observations. The dependence is assumed to be parameterized by a finite dimensional vector θ such that

$$y_k = f_k\left(\{u_i\}, \{z_i\}; \theta\right), \qquad i = 1, \dots, k$$

for some possibly time-dependent functions $f_k(\cdot)$. These are general nonlinear maps and we only assume the parameterization by θ. Furthermore, we define $Z := \begin{bmatrix} z_1^T & z_2^T & \dots & z_N^T \end{bmatrix}^T$. The observations vector Y can then be written as

$$Y = \mathcal{M}(U, Z; \theta), \tag{1}$$

for some nonlinear map $\mathcal{M}$ parameterized with θ. We will be referring to $\mathcal{M}$ as the model. This definition is fairly general and includes well-known models as special cases. For example, consider a nonlinear state-space model

$$\begin{aligned} x_{k+1} &= h(x_k, u_k, w_k; \theta), \quad x_1 \sim p_{x_1}, \;\; w_k \sim p_w, \\ y_k &= g(x_k, u_k, e_k; \theta), \quad e_k \sim p_e. \end{aligned} \tag{2}$$

If we define $z_k = \begin{bmatrix} x_1^T & w_k^T & e_k^T \end{bmatrix}^T$, then we have

$$\begin{aligned} f_1(u_1, z_1; \theta) &= g(x_1, u_1, e_1; \theta), \\ f_2(\{u_i\}_1^2, \{z_i\}_1^2; \theta) &= g(h(x_1, u_1, w_1; \theta), u_2, e_2; \theta), \\ &\vdots \\ f_N(\{u_i\}_1^N, \{z_i\}_1^N; \theta) & \\ = g(h(\dots h(x_1, u_1, w_1; \theta) \dots, u_{N-1}, w_{N-1}; \theta), u_N, e_N; \theta) & \end{aligned}$$

which in turns completely defines the model $\mathcal{M}$.

In what follows, we will assume that we have the correct structure (parameterization) so that there exists a 'true' parameter vector $\theta_\circ$. We also require Z to follow a known probability law, p_Z, which might be parameterized by θ. It will be assumed that it is easy to generate independent samples according to p_Z. This assumption restricts the models to those that can be simulated once an input U and a parameter θ are fixed.

For a given pair (Y, U), our goal is to construct an estimator $\hat{\theta}(Y, U)$ of $\theta_\circ$. The MLE is an attractive candidate that makes an efficient use of the observations. It is well known to possess desirable optimal asymptotic properties; however, it requires the evaluation of the likelihood function, which we denote by $p(Y|U;\theta)$ with fixed Y. The MLE is defined by

$$\hat{\theta}_{ML} := \arg\max_{\theta \in \Theta} p(Y|U;\theta).$$

Unfortunately, due to the involved nonlinear transformation of random variables, the evaluation of the likelihood function is very challenging. For a general model (1), the likelihood function is defined by marginalization, that is

$$p(Y|U;\theta) = \int p(Y, Z|U;\theta) dZ. \tag{3}$$

This multidimensional integral is in general analytically intractable. One might consider using deterministic numerical integration; nevertheless, it is based on deterministic griding which is usually extremely inefficient and is practically unfeasible for large values of N. For example, applying the Simpson's rule with only 10 points per dimension requires a grid of a size 10^N. In such cases, Monte-Carlo integration is the only numerical integration tool that can be used to obtain any acceptable approximations.

So far, the available literature on the MLE follows two main approaches. The first seeks a direct approximation of the integral (3) either by using sequential Monte-Carlo samplers [Hürzeler and Künsch 2001, Malik and Pitt 2011, DeJong et al. 2012] or by importance sampling [Durbin and Koopman 1997, Abdalmoaty and Hjalmarsson 2016]. The other approach uses iterative algorithms like the Monte-Carlo Expectation-Maximization algorithm, which relies on a particle smoother [Olsson et al. 2008, Schön et al. 2011, Lindsten et al. 2014]. In spite of their optimal asymptotic properties, a major difficulty of such approaches -beside others- is the computational demand.

One of the main messages of this paper is that, by relaxing the requirement of asymptotic efficiency and only requiring the consistency of the estimator, the computational costs may be reduced. In the following section, we present a simulation-based technique based on a simulated pseudo likelihood. We stress here that this method does not require any sequential filtering or smoothing to get a consistent estimator.

3. SIMULATED PSEUDO MAXIMUM LIKELIHOOD

As the name suggests, the pseudo likelihood method relies on a misspecified likelihood function. Here, we will use the pseudo family of distributions,

$$\check{p}(Y|U;\theta) = \mathcal{N}\left(Y; \mu(U;\theta), \Sigma(U;\theta)\right),$$

in which the right hand side denotes a multivariate normal density. The mean $\mu(U;\theta)$ and the covariance $\Sigma(U;\theta)$ are defined by the first two moments of the model

$$\begin{aligned} \mu(U;\theta) &:= \mathbf{E}_Z\left[\mathcal{M}(U, Z;\theta)\right], \text{ and} \\ \nu(U;\theta) &:= \mathbf{E}_Z[\mathcal{M}(U, Z;\theta)\mathcal{M}(U, Z;\theta)^T], \\ \Sigma(U;\theta) &:= \nu(U;\theta) - \mu(U;\theta)\mu(U;\theta)^T. \end{aligned} \tag{4}$$

The Pseudo Maximum Likelihood (PML) estimator is defined as the global maximizer of $\check{p}(Y|U;\theta)$. Under mild conditions on the parameterization and the data, it has been shown in Gouriéroux et al. [1984] that the PML is consistent and asymptotically normal. However, due to the use of a misspecified likelihood function, the estimator is not guaranteed to be asymptotically efficient.

As can be seen in (4), the computation of the first two moments requires the evaluation of the multidimensional integrals

$$\begin{aligned}\mu(U;\theta) &= \int \mathcal{M}(U,Z;\theta)p_z(Z;\theta)dZ,\\ \nu(U;\theta) &= \int \mathcal{M}(U,Z;\theta)\mathcal{M}^T(U,Z;\theta)p_z(Z;\theta)dZ.\end{aligned} \tag{5}$$

At first glance, it might seem that there is no gain in considering such a pseudo family of likelihoods because the integrals in (5) are also analytical intractable. However, it is found that approximating these two integrals using Monte-Carlo simulations is indeed computationally easier than approximating the true likelihood in (3). The difference is due to the integrands. The approximation of (3) requires an efficient sampler in the high-dimensional space of Z. Such a sampler has to make use of the data to generate the samples. Naïve samplers will waste most of the samples in regions where the integrand $p(Y|Z,U;\theta)$ is practically equal to zero and only very few samples will contribute to the value of the integral. On the other hand, when approximating the moments in (5), we only rely on the model and every sample contributes equally to the value of the integral as shown below.

The Simulated Pseudo Maximum Likelihood (SPML) is defined using the simulated pseudo family of distributions

$$\tilde{p}(Y|U;\theta) = \mathcal{N}\left(Y; \widehat{\mu(U;\theta)}, \widehat{\Sigma(U;\theta)}\right),$$

in which the mean $\widehat{\mu(U;\theta)}$ and the covariance $\widehat{\Sigma(U;\theta)}$ are Monte-Carlo approximations of those of $\mathcal{M}(U,Z;\theta)$. These approximations are calculated as follows. For some given candidate value θ, we assume that we are able to generate M independent samples (trajectories)

$$Z_m(\theta) \sim p_z(Z;\theta), \text{ i.i.d. over } m = 1,\dots,M$$

of the unobserved process Z. Observe that p_z is parameterized by θ. Then, we can use the model (1) to simulate pseudo observations $Y_m = \mathcal{M}(U, Z_m(\theta);\theta)$. These are M independent samples $Y_m \sim p(Y|U;\theta)$ i.i.d. over m distributed according to the true distribution of $\mathcal{M}(U,Z;\theta)$. An important observation to be made here is that, given the input and a parameter θ, these samples rely only on the model $\mathcal{M}$. Then we define the estimates

$$\widehat{\mu(U;\theta)} = \frac{1}{M}\sum_{m=1}^{M} Y_m,$$

$$\widehat{\Sigma(U;\theta)} := \frac{1}{M-1}\sum_{m=1}^{M}(Y_m - \widehat{\mu(U;\theta)})(Y_m - \widehat{\mu(U;\theta)})^T.$$

Observe that each sample contributes equally to the final estimates. By direct application of the strong law of large numbers, these estimates converge almost surely to the true values in (4). An important feature of the estimates is that the convergence rate does not depend on the dimension of Y.

The SPML estimator is then defined as

$$\tilde{\theta} = \arg\max_{\theta\in\Theta} \ \tilde{p}(Y|U;\theta). \tag{6}$$

By running simulations for a sufficient amount of time, we can get arbitrarily close to the PML estimate. Under similar conditions as those used for the PML, it is possible to show that $\tilde{\theta} \overset{\text{a.s.}}{\to} \theta_0$ as $M, N \to \infty$ [Gouriéroux and Monfort 1990].

Notice that due to the way the parameter θ appears in $\tilde{p}(Y|U;\theta)$, it is not possible to find a closed form expression for the SPML estimator $\tilde{\theta}$. Numerical optimization algorithms, like the quasi-Newton algorithm for example, are therefore necessary to solve (6). Even though the function $\tilde{p}(Y|U;\cdot)$ is defined using a Monte-Carlo method, it is possible to guarantee its continuity by using common random numbers for the different evaluations of the function or its gradient. This is an advantage of the SPML method.

4. RELATION TO PREDICTION ERROR METHODS

In this section, we will describe the relationship between the PML estimator and the prediction error method. Prediction Error Methods (PEM) is a broad family of parameter estimation methods that can be applied to fairly general statistical models [Ljung 1999]. The idea behind the PEMs relies on the definition of the model (1) in terms of a predictor of future outputs. The predictor of the output at time k given the inputs and the observations up to and including time l is denoted by $\hat{y}_{k|1:l}(\theta)$. In what follows, we will consider the case when $l = k-1$. The predictor is then called one-step ahead predictor and is simply denoted by $\hat{y}_k(\theta)$. The prediction error process is defined by the difference

$$\varepsilon_k(\theta) := y_k - \hat{y}_k(\theta)$$

for which we define

$$\lambda_k(\theta) := \mathbf{Cov}(\varepsilon_k(\theta)) = \mathbf{Cov}(y_k - \hat{y}_k(\theta)).$$

The covariance operator is applied with respect to the distribution of possible output vectors of $\mathcal{M}$. PEMs construct an estimator by minimizing the size of the predictor errors using some norm. The estimate is given by

$$\hat{\theta}_{PE} := \arg\min_{\theta\in\Theta} \sum_{k=1}^{N} \ell(\varepsilon_k(\theta);\theta),$$

where $\ell(\cdot)$ is a scalar valued function that can be parameterized by θ. By a particular choice of ℓ and a corresponding definition of $\hat{y}_k$, the prediction error estimator boils down to the MLE. However, for the general model (1), such choices lead to analytically intractable predictors.

4.1 Implicit linear predictor

We will show that the PML estimator solves a prediction error minimization problem. For brevity, we will suppress the arguments U and θ for all notations in the following part. We start by an LDL decomposition of the covariance matrix $\Sigma = L\Lambda L^T$, in which L is a lower unitriangular matrix (with 1's on its main diagonal), and Λ is a diagonal matrix with positive entries. This decomposition is unique and always exists for symmetric positive definite matrices. The PML estimator is then defined by

$$\arg\min_{\theta} \ (Y-\mu)^T L^{-T}\Lambda^{-1}L^{-1}(Y-\mu) + \log\det\Lambda$$

where we ignored all constants and used the fact that $\det(L\Lambda L^T) = \det\Lambda$. Let us define the prediction errors vector $\mathcal{E} = \begin{bmatrix}\varepsilon_1^T & \varepsilon_2^T & \dots & \varepsilon_N^T\end{bmatrix}^T := L^{-1}(Y-\mu)$, and denote the vector of predictors by $\hat{Y} = \begin{bmatrix}\hat{y}_1^T & \hat{y}_2^T & \dots & \hat{y}_N^T\end{bmatrix}^T$. Because the error $\mathcal{E} = Y - \hat{Y} = L^{-1}(Y-\mu)$, we get an implicitly defined predictor that is linear in Y;

$$\hat{Y} := (I - L^{-1})Y + L^{-1}\mu. \tag{7}$$

Observe that if we specialize the model to a linear state-space model, this predictor is optimal, and Λ will contain the variance of the innovations [Kailath et al. 2000].

For the general nonlinear model (1), the predictor in (7) depends only on the second order statistics of the model. To clarify this point, let us assume that $N = 3$ and that the output is scalar, and let the mean and covariance (which are nonlinear functions of U) be denoted by

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix}.$$

The predictor (7) is then

$$\hat{y}_1 = \mu_1, \quad \hat{y}_2 = \mu_2 + \frac{\sigma_{21}}{\sigma_{11}}(y_1 - \mu_1), \text{ and}$$
$$\hat{y}_3 = \mu_3 + \frac{\sigma_{32} - \sigma_{31}}{\sigma_{22} - \sigma_{21}^2/\sigma_{11}}(y_2 - \mu_2) + \frac{\sigma_{31}(\sigma_{22} - \sigma_{21}^2) - \sigma_{21}(\sigma_{32} - \sigma_{31})}{\sigma_{11}(\sigma_{22} - \sigma_{21}^2)}(y_1 - \mu_1),$$

with variances

$$\lambda_1 = \sigma_{11}, \quad \lambda_2 = \sigma_{22} - \frac{\sigma_{21}^2}{\sigma_{11}}, \text{ and}$$
$$\lambda_3 = \sigma_{33} - \frac{\sigma_{31}^2\sigma_{22}}{\sigma_{11}^2} - \frac{\sigma_{22}^2}{\sigma_{11}}.$$

The PML minimization problem can then be equivalently written as a prediction error minimization problem:

$$\check{\theta} = \arg\min_{\theta} \sum_{k=1}^{N} \left(\frac{\varepsilon_k^T(\theta)\varepsilon_k(\theta)}{\lambda_k(\theta)} + \log \lambda_k(\theta) \right) \tag{8}$$

in which $\lambda_k(\theta)$ is the k^{th} entry of Λ. Therefore, the PML estimator uses an implicitly defined linear predictor and assumes that the prediction errors are independent and Gaussian with parameter-dependent variances.

The above defined predictor depends on the moments (5) of the model and is not available in closed form. The SPML estimator uses Monte-Carlo simulations to approximate both the predictor and the variances by using $\hat{\mu}$ and $\hat{\Sigma}$.

For the case of state-space models, a related linear predictor can be defined using the EnKF as we describe next.

4.2 The ensemble Kalman filter

The Ensemble Kalman Filter (EnKF) uses sequential Monte-Carlo simulations to approximately compute and propagate the moments of the filtering and predictive densities for the state-space models (2). It can be seen as a Monte-Carlo implementation of the Kalman recursions. We refer the reader to Le Gland et al. [2009] and Roth et al. [2015] for details on properties and implementation. Because the EnKF recursions are sample-based, the empirical covariance matrix of the state does not need to be stored. This allows the filter to work with very high-dimensional state-space models. In cases of linear Gaussian models, the EnKF is known to converge to the Kalman filter and therefore to the optimal Bayesian solution.

Let us denote the EnKF one-step ahead predictions by $\hat{y}_{k|k-1}(\theta)$. This is the mean of the propagated output ensemble at time k. Furthermore, denote the output ensemble covariance by $\lambda_k(\theta)$. Then we can define the prediction error estimate

$$\hat{\theta} = \arg\min_{\theta} \sum_{k=1}^{N} \left(\varepsilon_k^T(\theta)\lambda_k^{-1}(\theta)\varepsilon_k(\theta) + \log\det \lambda_k(\theta) \right) \tag{9}$$

where the prediction error $\varepsilon_k(\theta) = y_k - \hat{y}_{k|k-1}(\theta)$.

We see that this estimator looks very similar to the SPML estimator as defined in (8). Both solve a prediction error problem with parameterized norm using a linear predictor. However, they differ in the way they define the predictor. The predictor (7) of the SPML is implicitly defined by first simulating the output vector of $\mathcal{M}$, while the EnKF constructs the predictor sequentially.

4.3 Numerical comparison

To check the relation between the linear predictor used by the SPML estimator and the one used by the EnKF, we perform a numerical experiment. First, we consider a model for which the relationship between x_k and y_k is linear

$$\begin{aligned} x_{k+1} &= \theta \frac{x_k}{x_k^2 + 1} + w_k, & w_k &\sim \mathcal{N}(0, 0.1) \\ y_k &= x_k + e_k, & e_k &\sim \mathcal{N}(0, 0.1), \end{aligned} \tag{10}$$

The state equation is nonlinear and is a variant of that of the standard benchmark of particle filters in Gordon et al. [1993]. We assume that $x_0 = 0$ and $\theta = 0.7$ and generate a realization Y for $N = 100$. We then run both algorithms with the true θ, $M = 10^5$ and plot the prediction errors, prediction errors variance, and the one-step ahead predictions. Finally, we calculate the cost for both estimators (8) and (9) for values of θ between 0.1 and 0.9 with a step 0.05. To further control the comparison, we use the same random numbers in both cases. The results are presented in Fig. 1. Although not identical, it is clear that both predictors give very close values. More interestingly, the cost functions have the same shape and seem to have very close minimizers.

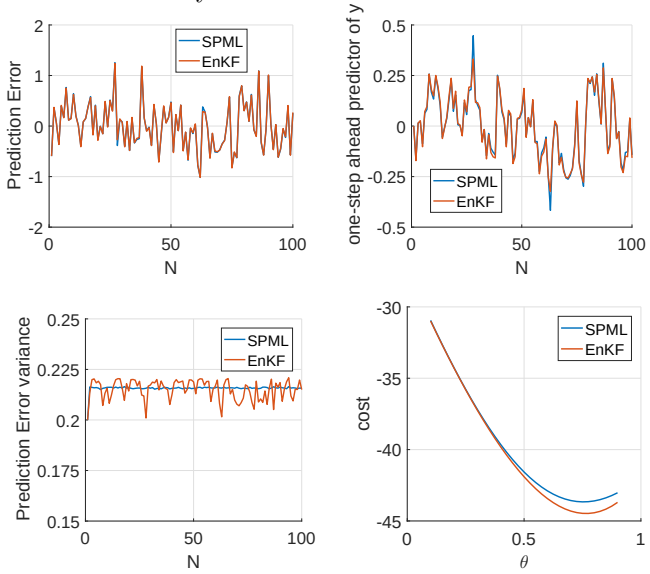

Fig. 1. Simulation results for the model in (10)

Next, we repeat the same experiment with the same setting but using a quadratic observation model

$$x_{k+1} = \theta \frac{x_k}{x_k^2 + 1} + w_k, \qquad y_k = x_k^2 + e_k \tag{11}$$

This gives a bi-modal case. Fig. 2 shows similar conclusions to those found in the case of a linear observation model. This experiment seems to highlight a fairly close relationship between the two estimators. In the following section, we extend the study by evaluating the performance of both estimators on a parameter identification example.

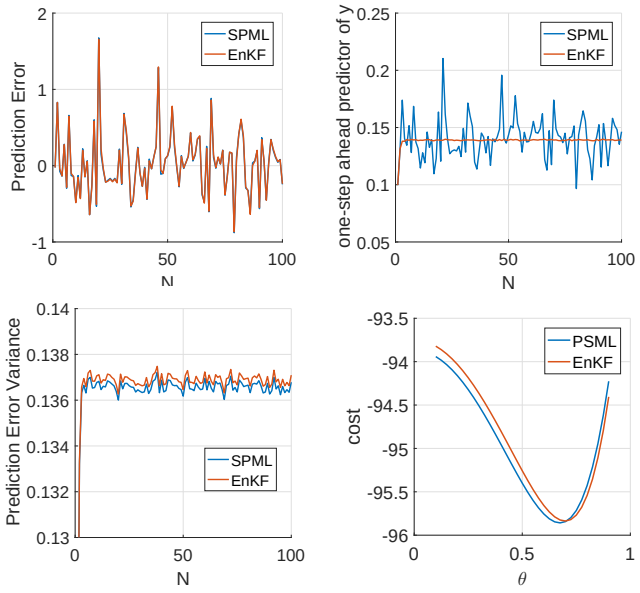

Fig. 2. Simulation results for the model in (11)

5. SIMULATION EXAMPLES

In this section, we report the results of Monte-Carlo studies for three examples. Observe that the models in the first two examples are bi-modal. All the algorithms are vectorized and implemented in MATLAB where the `fminunc` function (implementing a quasi-newton algorithm) is used to solve the optimization problems. Forward differences are used to approximate the gradients and the solver is initialized at the true parameters. All the statements regarding computational times are for an Intel-based laptop with a 2.7 GHz processor and 8 Gbyte RAM.

5.1 First order Wiener model

$$\begin{aligned} x_{k+1} &= \theta x_k + w_k, & \theta &= 0.7, \\ y_k &= x_k^2 + e_k, & x(0) &= 0, \\ e_k &\sim \mathcal{N}(0, 0.1), & w_k &\sim \mathcal{N}(0, 0.1) \end{aligned}$$

In this first example, we examine the performance of the SPML and the EnKF for parameter estimation. To do so, we use the same random numbers for both estimators and 10^5 samples. The result of the Monte-Carlo study is shown in Fig. 3. Both estimators give comparable performances. The required time to compute an estimate is a matter of minutes.

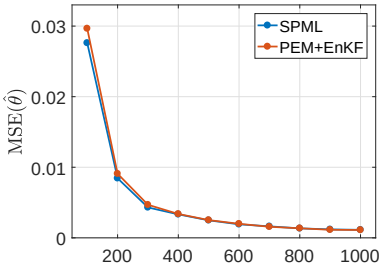

Fig. 3. MSE of SPML and EnKF estimators averaged over 1000 realizations for each N, fixed $M = 10^5$.

5.2 High-dimensional state-space example

In this example, we consider a nonlinear system with 100 states and one output,

$$\begin{aligned} x_1(k+1) &= \theta_1 \frac{x_1(k)}{x_{100}^2(k)+1} + w_1(k), \qquad x_0 = 0, \\ x_i(k+1) &= \theta_1 \frac{x_i(k)}{x_{i-1}^2(k)+1} + w_i(k), \quad i = 2, \ldots, 100 \\ y_k &= \left(\sum_{i=1}^{100} x_i(k) \right)^2 + e_k, \quad e_k \sim \mathcal{N}(0, 0.1), \\ w_i(k) &\sim \mathcal{N}(0, \theta_2), \quad M = 10^4, \quad \theta_1 = 0.7, \quad \text{and } \theta_2 = 0.1. \end{aligned}$$

Due to the high-dimensional state-space, this example is quite challenging for estimators based on optimal filtering methods because of the cyclic dependence between all the states. It also poses a challenge for any estimation method that relies on approximations of the true likelihood. To the best of authors' knowledge, the parameter identification algorithms targeting the MLE which are based on sequential Monte-Carlo algorithms have been applied only to problems with small dimensions.

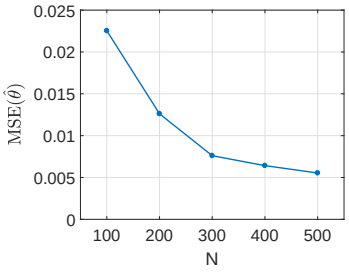

Fig. 4. MSE of SPML averaged over 250 realizations.

Applying the SPML estimator is straightforward, as in the previous first order example, and no special attention is required. Fig. 4 presents the result which indicates the consistency of the estimator. Here, we estimated two parameters, one of which is the variance of the process disturbance. We also observed comparable behavior for the EnKF estimator (we only tested the case with known θ_2).

5.3 Cascaded anaerobic digestion process

We conclude this section with an application-motivated example. We consider a dynamical model of a continuous anaerobic digestion process. This is a process used for the treatment of organic waste in which the microorganism is broken into a mixture of methane and carbon dioxide. The details of such a process and possible models can be found in Bastin and Dochain [1990] for example. We consider a stochastic two-stage bioreactor model

$$\begin{aligned} dS_1 &= [-k_1\mu_1(S_1, X_1) + D_1(S_{1,\text{in}} - S_1)]dt + dW_1 \\ dX_1 &= [\mu_1(S_1, X_1) - D_1 X_1]dt + dW_2 \\ dS_{21} &= [k_3\mu_1(S_1, X_1) - D_1 S_{21}]dt + dW_3 \\ dX_2 &= [\mu_2(S_{22}, X_2) - D_2 X_2]dt + dW_4 \\ dS_{22} &= [-k_2\mu_2(S_{22}, X_2) + D_2(S_{21} - S_{22})]dt + dW_5 \end{aligned}$$

in which the dilution rates $D_1 = 0.04$ and $D_2 = 0.01$/day. S_1 represents the substrate concentration in tank 1, and $S_{1,\text{in}}$ is the substrate concentration in the influent. S_{21} and S_{22} represent the product substrate concentration in tank 1 and 2 respectively. X_1 and X_2 are the concentrations of the biomass in tank 1 and 2 respectively. The parameters k_1, k_2, and k_3 are yield coefficients. The growth rates are modeled using a Monod-law

$$\mu_1(S_1, X_1) = \frac{\mu_1^* S_1 X_1}{S_1 + K_{m1}}, \quad \mu_2(S_{22}, X_2) = \frac{\mu_2^* S_{22} X_2}{S_{22} + K_{m2}},$$

with the assumption that $\mu_1^* = \mu_2^* = 1/\text{hr}$. The goal is to identify the parameters vector $\theta = [k_1\ k_2\ k_3\ K_{m1}\ K_{m2}]$ using the two measurements

$$y_1(k) = S_1(kT) + e_1(k), \text{ and } y_2(k) \quad = S_{22}(kT) + e_2(k),$$

for $k = 1. \ldots, N$. To generate the estimation data, we discretize the model using Euler's method with time interval $T = 7/24$ day. The process is then simulated in discrete-time with the initial values $S_1 = X_1 = S_{22} = 0.5$, and $X_2 = S_{21} = 1$, and $\theta = [5\ 10\ 6\ 10\ 20]$. The input $S_{1,\text{in}}$ is a square wave with levels 20 and 40 g/L. The initial level is 20 and it is changed every 7 days. The variance of the discrete-time process disturbance is $0.001T$, and the measurement noise is independent and normal with variance 0.005. We performed a Monte-Carlo experiment of 1000 realizations for three experiment durations: $60, 80,$ and 100 days. They correspond to $N = 205, 274,$ and 342 samples respectively. The number of simulations M is fixed to 10^4. The results are summarized in the following table

		k_1	k_2	k_3	K_{m1}	K_{m2}	MSE
N=205	mean	4.99	9.83	6.01	9.99	19.6	8.18
	std	0.28	1.41	0.76	0.15	2.31	
N=274	mean	5	9.88	5.99	9.99	19.7	5.14
	std	0.27	1.16	0.72	0.13	1.76	
N=342	mean	4.98	9.93	6.02	9.99	19.8	3.81
	std	0.27	1.06	0.72	0.12	1.43	

The results show that the SPML gives a consistent estimator. The time required to estimate the parameters is, once more, in the order of few minutes.

6. CONCLUSIONS

In this paper, we studied the simulated pseudo maximum likelihood (SPML) method for parameter identification of nonlinear models. It is used to construct a computationally inexpensive consistent estimator, which can be made efficient if required by an additional Gauss-Newton step. We showed that solving the SPML problem is equivalent to solving a prediction error problem with a parameterized norm and a linear predictor. We also highlighted the relation between such a predictor and the predictor defined by the ensemble Kalman filter; the simulation study suggests a close relationship between the two. The numerical examples illustrate the good performance of the suggested estimator using models which are considered challenging for the current mainstream approaches.

REFERENCES

Abdalmoaty, M.R. and Hjalmarsson, H. (2016). A simulated maximum likelihood method for estimation of stochastic Wiener systems. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, 3060–3065.

Bastin, G. and Dochain, D. (1990). *On-line Estimation and Adaptive Control of Bioreactors*. Elsevier.

Besag, J. (1975). Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 24(3), 179–195.

DeJong, D.N., Liesenfeld, R., Moura, G.V., Richard, J.F., and Dharmarajan, H. (2012). Efficient likelihood evaluation of state-space representations. *The Review of Economic Studies*.

Doucet, A., Pitt, M.K., Deligiannidis, G., and Kohn, R. (2015). Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika*.

Durbin, J. and Koopman, S.J. (1997). Monte Carlo maximum likelihood estimation for non-gaussian state space models. *Biometrika*, 84(3), 669–684.

Evensen, G. (2009). The ensemble Kalman filter for combined state and parameter estimation. *IEEE Control Systems*, 29(3), 83–104.

Gillijns, S., Mendoza, O.B., Chandrasekar, J., Moor, B.L.R.D., Bernstein, D.S., and Ridley, A. (2006). What is the ensemble Kalman filter and how well does it work? In *2006 American Control Conference*, 6–pp.

Gordon, N.J., Salmond, D.J., and Smith, A.F.M. (1993). Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEE Proceedings F - Radar and Signal Processing*, 140(2), 107–113.

Gouriéroux, C., Monfort, A., and Trognon, A. (1984). Pseudo maximum likelihood methods: Theory. *Econometrica*, 52(3), 681–700.

Gouriéroux, C. and Monfort, A. (1990). Simulation based inference in models with heterogeneity. *Annales d'Économie et de Statistique*, (20/21), 69–107.

Hürzeler, M. and Künsch, H.R. (2001). *Approximating and Maximising the Likelihood for a General State-Space Model*, 159–175. Springer New York.

Ionides, E.L., Bret, C., and King, A.A. (2006). Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 103(49), 18438–18443.

Kailath, T., Sayed, A., and Hassibi, B. (2000). *Linear Estimation*. Prentice-Hall information and system sciences series. Prentice Hall.

Kantas, N., Doucet, A., Singh, S.S., Maciejowski, J., and Chopin, N. (2015). On particle methods for parameter estimation in state-space models. *Statist. Sci.*, 30(3), 328–351.

Le Gland, F., Monbet, V., and Tran, V.D. (2009). Large sample asymptotics for the ensemble Kalman filter. Research Report RR-7014, INRIA.

Lindsten, F., Jordan, M.I., and Schön, T.B. (2014). Particle Gibbs with ancestor sampling. *J. Mach. Learn. Res.*, 15(1), 2145–2184.

Ljung, L. (1999). *System Identification (2nd Ed.): Theory for the User*. Prentice Hall PTR.

Malik, S. and Pitt, M.K. (2011). Particle filters for continuous likelihood evaluation and maximisation. *Journal of Econometrics*, 165(2), 190 – 209.

Olsson, J., Cappé, O., Douc, R., and Moulines, E. (2008). Sequential Monte Carlo smoothing with application to parameter estimation in nonlinear state space models. *Bernoulli*, 14(1), 155–179. doi:10.3150/07-BEJ6150.

Poyiadjis, G., Doucet, A., and Singh, S.S. (2011). Particle approximations of the score and observed information matrix in state space models with application to parameter estimation. *Biometrika*, 98(1), 65–80.

Roth, M., Fritsche, C., Hendeby, G., and Gustafsson, F. (2015). The ensemble Kalman filter and its relations to other nonlinear filters. In *Signal Processing Conference (EUSIPCO), 2015 23rd European*, 1236–1240.

Schön, T.B., Wills, A., and Ninness, B. (2011). System identification of nonlinear state-space models. *Automatica*, 47(1), 39 – 49.

Snyder, C., Bengtsson, T., Bickel, P., and Anderson, J. (2008). Obstacles to high-dimensional particle filtering. *Monthly Weather Review*, 136(12), 4629–4640.