A Directional Forgetting Algorithm Based on the Decomposition of the Information Matrix

Liyu Cao*and Howard M. Schwartz†
Department of Systems and Computer Engineering
Carleton University
1125 Colonel By Drive, Ottawa, On. K1S 5B6, Canada

Abstract

A novel directional forgetting algorithm is developed based on a decomposition of the information matrix. This algorithm performs exponential forgetting only to a specified part of the information matrix, thus preventing the problem known as estimator windup which is a characteristic of the standard exponential forgetting algorithm. This algorithm is able to track fast parameter changes and is similar in complexity to the standard least square algorithm. The superior performance of the algorithm is verified via theoretical and simulation studies.

1 Introduction

The exponential forgetting(EF) recursive least squares(RLS) algorithm is a well known method for on-line parameter estimation. Its main drawback is the so-called estimator windup which occurs when the system input is not persistently excited as shown in (Åström and Wittenmark, 1995,Ch.11). A number of modifications to the exponential forgetting algorithm have been proposed in order to overcome this drawback. Among them a notable one is the directional forgetting(DF) strategy suggested by Hägglund(1985) and Kulhavý(1987). The basic idea is that we should forget old data *only in the direction where the new one is coming from.* When the input is not persistently excited, then in some directions no information about the system dynamics is available. Therefore, the forgetting operation should be applied only to the part of the information matrix, where new information is available from the input and output data.

There are many possibilities to implement the directional forgetting strategy. Kulhavý(1987) proposed a directional forgetting algorithm which can prevent estimator windup. However, it has been recognized that in this algorithm all eigenvalues of the information matrix are not bounded from above as shown in (Bittanti *et al.*, 1990). This means that in some directions the algorithm will eventually lose its tracking capability. In particular, this algorithm is not suitable for tracking jump changes in parameters. An alternative approach to directional forgetting is developed by Parkum *et al.*(1992) who called their method *selective forgetting*(SF). Although the name is different, this approach can also be viewed as a kind of the DF method. In the SF method, eigenvectors of the information matrix are used as references for determining the direction of incoming data, and the corresponding eigenvalues are used as indicators of the amount of information coming in the eigenvectors' direction. This consideration provides very

*Email:*cao@sce.carleton.ca*
†Email:*schwartz@sce.carleton.ca*

useful insights into the way in which the directional forgetting strategy may be implemented effectively. A drawback of the SF method is that the eigenvalues and eigenvectors have to be computed at each update, which greatly increases the computational requirement compared to the standard EF algorithm.

In this paper, a new algorithm for directional forgetting is developed based on a decomposition of the information matrix. Through the matrix decomposition, the directional forgetting can be implemented in an effective way. In this approach, only a specified *part* of the information matrix is forgotten at each update. Under the condition of persistent excitation, the algorithm has the same behavior as that of the standard EF algorithm, which means the dynamic tracking performance is good. When excitation is poor, windup does not occur because nothing is forgotten in the part of the information matrix, which is *orthogonal* to the excited space in some sense. Compared with various modifications to the EF algorithm, the new algorithm is simple in the sense that there are only two adjustable constants that need to be prespecified. A detailed analysis of the properties of the resultant algorithm is also given. Finally, a simulation example is given that compares the proposed algorithm with the SF algorithm.

2 Matrix Decomposition and Directional Forgetting

Before describing the new algorithm, let us have a look at the update equation for the information matrix in the EF algorithm, which is given by

$$R(t) = \mu R(t-1) + \varphi(t)\varphi^T(t) \tag{1}$$

where $0 < \mu < 1$ is the forgetting factor. From the above equation, it can be seen that the old data in $R(t-1)$ is forgotten uniformly in all directions and continuously in time regardless of the new information content associated with $\varphi(t)$. This does not cause any problems if the input is persistently excited, because the information forgotten at time t in any direction can be compensated by the data vectors $\varphi(t), \varphi(t+1), \cdots$. However, when the input does not excite the system in all directions, then continuous forgetting in all directions will lead to the loss of some information in certain directions(some eigenvalues of $R(t)$ will tend to zero), because no compensation in these directions is available. This is the reason why estimator windup occurs.

The above observation motivated the proposal of the directional forgetting algorithm. That is, forgetting should be restricted to the directions where the new data is coming from. In order to implement the directional forgetting effectively, one should first connect the information matrix $R(t)$(or its inverse $P(t)$) with the data vector $\varphi(t)$ in terms of some 'spatial' relationship. And then, one performs the forgetting operation on $R(t)$ based on this relationship. Kulhavý used the Bayesian estimation scheme to get a directional forgetting algorithm without giving an explicit relationship between $R(t)$ and the data vector. In the SF approach, the eigenvectors of $P(t)$ are used to determine the directions of the data vectors with the added cost of computing the eigenvectors and the eigenvalues at each update. Here, we propose a directional forgetting algorithm based on a decomposition of the information matrix. Our approach is motivated by the observations made in (Åström and Wittenmark,1995, Ch.11). We extend their work by proposing a complete information matrix decomposition theory.

Our approach is started by decomposing the information matrix $R(t-1)$ into two parts before performing forgetting

$$R(t-1) = R_1(t-1) + R_2(t-1) \tag{2}$$

where $R_1(t-1)$ is assumed to satisfy the following equation

$$R_1(t-1)\varphi(t) = 0, \quad \varphi(t) \neq 0 \tag{3}$$

which means that $\varphi(t)$ is in the kernel space of $R_1(t-1)$. The matrix $R_2(t-1)$ represents that part of $R(t-1)$ which should be forgotten. Since $R(t)$ is positive definite, it is required that both $R_1(t-1)$ and $R_2(t-1)$ are nonnegative definite. For $R_1(t-1)$ we have the following lemma.

Lemma 1. Assume that $R_1(t-1)$ satisfies (3). Then the image space of $R_1(t-1)$ is orthogonal to $\varphi(t)$, that is

$$ImR_1(t-1) \subseteq \phi^{\perp}(t) \tag{4}$$

where $\phi^{\perp}(t)$ denotes the orthogonal complement of $\varphi(t)$

$$\phi^{\perp}(t) = \{x : \langle x, \varphi(t)\rangle = 0\} \tag{5}$$

Proof. This lemma is easy to prove by noting that the image space $\mathrm{Im}R_1(t-1)$ is spanned by the columns of $R_1(t-1)$, and each column of $R(t-1)$ is orthogonal to $\varphi(t)$. ■

Obviously, if we let the rank of $R_1(t-1)$ be $n-1$, then we have

$$ImR_1(t-1) = \phi^{\perp}(t) \tag{6}$$

Therefore, by letting $R_1(t-1)$ satisfy (3), we establish an orthogonal relationship between $R_1(t-1)$ and $\varphi(t)$. In fact, if one looks at the image spaces of $R_1(t-1)$ and $\varphi(t)\varphi^T(t)$, one will find they are orthogonal to each other.

Now we turn to $R_2(t-1)$, which is characterized by the following equation

$$R_2(t-1)\varphi(t) = R(t-1)\varphi(t) \tag{7}$$

With (3) and (7), $R_1(t-1)$ or $R_2(t-1)$ may not be determined uniquely. We need to add some restriction on the rank of $R_1(t-1)$ and $R_2(t-1)$. Because the new information is coming in the form $\varphi(t)\varphi^T(t)$ and the rank of $\varphi(t)\varphi^T(t)$ is 1, it is reasonable to require that the rank of $R_2(t-1)$ is 1 and the rank of $R_1(t-1)$ is $n-1$(n is the order of $R(t)$). With these rank conditions, $R_2(t-1)$ is found to be

$$R_2(t-1) = \alpha(t)[R(t-1)\varphi(t)][R(t-1)\varphi(t)]^T \tag{8}$$

where $\alpha(t)$ is a scalar and is given by

$$\alpha(t) = \frac{1}{\varphi^T(t)R(t-1)\varphi(t)} \tag{9}$$

In order to ensure that $\alpha(t)$ is well-defined, $\varphi(t)$ must be a nonzero vector. In fact, to ensure the algorithm is well-behaved, we need to set a dead zone for $\varphi(t)$. The algorithm is not activated unless $\varphi(t)$ is outside the dead zone. Therefore, when $\varphi(t)$ is within the dead zone, we have

$$\alpha(t) = 0, \quad \text{if } |\varphi(t)| < \epsilon \tag{10}$$

where ϵ can be determined based on the noise level in data. In the following, it is assumed that $|\varphi| \geq \epsilon$ unless the opposite situation is mentioned.

From equation (8), one can get $R_1(t-1)$ as

$$R_1(t-1) = R(t-1) - \alpha(t)[R(t-1)\varphi(t)][R(t-1)\varphi(t)]^T \tag{11}$$

The matrix $R_1(t-1)$ satisfies (3). Furthermore, $R_1(t-1)$ has the following properties.

Lemma 2. Assume that $R(t-1)$ is positive definite. Then the matrix $R_1(t-1)$ given by (11) is nonnegative definite and its rank is $n-1$.

Proof. Consider the following spectral radius

$$\rho[R_2(t-1)R^{-1}(t-1)] = \rho[\alpha(t)R(t-1)\varphi(t)\varphi^T(t)] \tag{12}$$

The rank of the matrix $\alpha(t)R(t-1)\varphi(t)\varphi^T(t)$ is 1, and hence the matrix has only a single nonzero eigenvalue. From the fact that the sum of eigenvalues of a matrix is equal to its trace, we can see that the nonzero eigenvalue is equal to 1. Therefore, we get $\rho[R_2(t-1)R^{-1}(t-1)] = 1$. Then by using Theorem 7.7.3 in (Horn and Johnson, 1985,Ch.7), one can get

$$R_1(t-1) = R(t-1) - R_2(t-1) \geq 0 \tag{13}$$

which[1] shows that $R_1(t-1)$ is nonnegative definite.

One can show that the rank of $R_1(t-1)$ is $n-1$ by proving that the dimension of its kernel space is 1. For this propose, assume that there is a nonzero vector x which belongs to the kernel space. That is, x satisfies

$$R_1(t-1)x = 0 \tag{14}$$

This is equivalent to the following equation

$$\begin{aligned} & R(t-1)x - \alpha(t)R(t-1)\varphi(t)\varphi^T(t)R(t-1)x \\ = \;& R(t-1)(x - \alpha(t)[\varphi^T(t)R(t-1)x]\varphi(t)) = 0 \end{aligned}$$

The matrix $R(t-1)$ is positive definite, therefore the above equation has only a zero solution given by

$$x = \alpha(t)[\varphi^T(t)R(t-1)x]\varphi(t) \tag{15}$$

Equation (15) indicates that the kernel space of $R_1(t-1)$ is spanned by $\varphi(t)$, and hence its dimension is 1. ■

Thus, the matrix $R(t-1)$ has been decomposed into a matrix of rank $n-1(R_1(t-1))$ and a matrix of rank of $1(R_2(t-1))$. Applying exponential forgetting only to the small rank matrix $R_2(t-1)$(refer to equation (1)), the update equation for the information matrix becomes

$$\begin{aligned} R(t) &= R_1(t-1) + \mu R_2(t-1) + \varphi(t)\varphi^T(t) \\ &= \bar{R}(t-1) + \varphi(t)\varphi^T(t) \end{aligned} \tag{16}$$

where $\bar{R}(t-1)$ denotes the modified information matrix,

$$\bar{R}(t-1) = (I - M(t))R(t-1) \tag{17}$$

$$M(t) = (1-\mu)\alpha(t)R(t-1)\varphi(t)\varphi^T(t) \tag{18}$$

Equation (17) also represents the time update equation(refer to (Parkum *et al.*,1992)).

Let

$$N(t) = I - M(t) \tag{19}$$

where $N(t)$ is called *forgetting matrix*. One can observe that $N(t)$ has the same function as that of the forgetting factor μ in the EF method. We have the following lemma for $N(t)$.

Lemma 3. $N(t)$ is nonsingular. Furthermore, there is an eigenvalue of $N(t)$ which is equal to μ, and the other eigenvalues are equal to 1.

[1] in this paper, for symmetric matrices A and B, $A \geq B (A > B)$ means $A - B$ is positive semidefinite(definite).

Proof. First note that the rank of $M(t)$ is 1 and it has only a single nonzero eigenvalue. Let this eigenvalue be λ_1. By using the same arguements as used in the proof for Lemma 2, one can get

$$\lambda_1 = trace[M(t)] = (1-\mu)\alpha(t)\varphi^T(t)R(t-1)\varphi(t) = 1-\mu \tag{20}$$

By using the fact that $\lambda[N(t)] = 1-\lambda[M(t)]$, where $\lambda[\cdot]$ denotes eigenvalue, the lemma is proven. ■

For the modified information matrix $\bar{R}(t-1)$ we have the following lemma.

Lemma 4. If $R(t)$ is positive definite and $0 < \mu < 1$, then $\bar{R}(t)$ is also positive definite.

Proof. Equation (17) can be rewritten as

$$\bar{R}(t-1) = R(t-1) - M(t)R(t-1) \tag{21}$$

If $R(t-1)$ is nonsingular, then according to Theorem 7.7.3 in (Horn and Johnson, 1985,Ch.7), $\bar{R}(t-1)$ is positive definite if and only if

$$\rho([M(t)R(t-1)][R(t-1)]^{-1}) = \rho(M(t)) < 1 \tag{22}$$

where $\rho(\cdot)$ denotes the spectral radius of a matrix. From the proof for Lemma 3, we know that $\rho(M(t)) = 1-\mu < 1$, therefore $\bar{R}(t-1)$ is positive definite. ■

Noting that $M(t)R(t-1)$ is nonnegative definite, then from (21) we get

$$\bar{R}(t-1) \leq R(t-1) \tag{23}$$

Inequality (23) means that the old information is forgotten(discounted) at each update. The forgetting operation is performed through the forgetting matrix $N(t)$, and therefore is not uniform in all directions.

Using the following matrix inversion lemma

$$(A+BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1}+DA^{-1}B)^{-1}DA^{-1} \tag{24}$$

and letting $A = \bar{R}(t-1)$, $B = \varphi(t)$, $C = 1$ and $D = \varphi^T(t)$ in (16), one can obtain the update equation for the covariance matrix

$$P(t) = \bar{P}(t-1) - \frac{\bar{P}(t-1)\varphi(t)\varphi^T(t)\bar{P}(t-1)}{1+\varphi^T(t)\bar{P}(t-1)\varphi(t)} \tag{25}$$

where $\bar{P}(t-1)$ is the modified covariance matrix and is defined by the following equation for $|\varphi(t)| > \epsilon$

$$\begin{aligned} \bar{P}(t-1) &= P(t-1)N^{-1}(t) \\ &= P(t-1) + \frac{1-\mu}{\mu}\frac{\varphi(t)\varphi^T(t)}{\varphi^T(t)P^{-1}(t-1)\varphi(t)} \end{aligned} \tag{26}$$

and for $|\varphi(t)| \leq \epsilon$, $\bar{P}(t-1)$ is defined by

$$\bar{P}(t-1) = P(t-1) \tag{27}$$

In fact, equations (26) and (27) are the time update equations for the covariance matrix. One can find that when the data vector $\varphi(t)$ is very small($|\varphi(t)| < \epsilon$), the update equation (25) is

exactly the same as the standard least squares method. This is reasonable because when the data carries little new information($|\varphi(t)| \leq \epsilon$), there is no need to forget the old data.

As a summary, the proposed algorithm can be represented by the following equations

$$\hat{\theta}(t) = \hat{\theta}(t-1) + K(t)[y(t) - \varphi^T(t)\hat{\theta}(t-1)] \tag{28}$$

$$K(t) = P(t)\varphi(t) = \frac{\bar{P}(t-1)\varphi(t)}{1 + \varphi^T(t)\bar{P}(t-1)\varphi(t)} \tag{29}$$

$$\bar{P}(t-1) = P(t-1) + \frac{1-\mu}{\mu}\frac{\varphi(t)\varphi^T(t)}{\varphi^T(t)R(t-1)\varphi(t)}, \quad |\varphi(t)| > \epsilon \tag{30}$$

$$\bar{P}(t-1) = P(t-1), \quad |\varphi(t)| \leq \epsilon \tag{31}$$

$$P(t) = \bar{P}(t-1) - \frac{\bar{P}(t-1)\varphi(t)\varphi^T(t)\bar{P}(t-1)}{1 + \varphi^T(t)\bar{P}(t-1)\varphi(t)} \tag{32}$$

$$R(t) = [I - M(t)]R(t-1) + \varphi(t)\varphi^T(t) \tag{33}$$

$$M(t) = (1-\mu)\frac{R(t-1)\varphi(t)\varphi^T(t)}{\varphi^T(t)R(t-1)\varphi(t)}, \quad |\varphi(t)| > \epsilon \tag{34}$$

$$M(t) = 0, \quad |\varphi(t)| \leq \epsilon \tag{35}$$

$$\tag{36}$$

The update equation for the information matrix is included in the algorithm because $R(t-1)$ is needed in updating $\bar{P}(t-1)$.

Finally, let us have a look on how the eigenvalues of $P(t)$ is updated during the modification. Consider the modified covariance matrix when the data vector $|\varphi(t)| \geq \epsilon$, given by (26). Noting $\bar{P}(t) = \bar{R}^{-1}(t)$ and inequality (23), one gets

$$\bar{P}(t-1) \geq P(t-1) \tag{37}$$

Based on this inequality one can further get(Horn and Johnson, 1985, Ch. 7)

$$\lambda_i[\bar{P}(t-1)] \geq \lambda_i[P(t-1)], \quad \text{for i=1,2,}\cdots\text{,n} \tag{38}$$

where the eigenvalues of $\bar{P}(t-1)$ and $P(t-1)$ are arranged in the same order. Inequality (38) indicates that all eigenvalues are non-decreasing when modifying $P(t-1)$ according to (26). In fact, some of the eigenvalues are increasing if $\varphi(t) \neq 0$. In particular, we have the following lemma.

Lemma 5. Consider the modified covariance matrix (26). If $\varphi(t)$ is in the direction of one eigenvector of $P(t-1)$, then one of the eigenvalues of $\bar{P}(t-1)$ is given by

$$\lambda_p[\bar{P}(t-1)] = \frac{1}{\mu}\lambda_q[P(t-1)], \quad p, q \in [1, 2, \cdots, n] \tag{39}$$

and all other eigenvalues of $\bar{P}(t-1)$ are given by

$$\lambda_i[\bar{P}(t-1)] = \lambda_i[P(t-1)], \quad i \neq p, i \neq q \tag{40}$$

Proof. Because $P(t-1)$ is positive definite and hence is diagonalizable, $P(t-1)$ can be written as

$$P(t-1) = \sum_{i=1}^{n} \lambda_i(t-1)v_i(t-1)v_i^T(t-1) \tag{41}$$

where $v_i(t-1)$ and $\lambda_i(t-1)$ are the orthonormal eigenvectors and eigenvalues of $P(t-1)$ respectively. If $\varphi(t)$ is in the direction of $v_p(t-1)$, that is, $\varphi(t) = |\varphi(t)|v_p(t-1)$, then from (26) we get

$$\bar{P}(t-1) = \sum_{i=1}^{n} \lambda_i(t-1)v_i(t-1)v_i^T(t-1) + \frac{1-\mu}{\mu}\lambda_p(t-1)v_p(t-1)v_p^T(t-1) \tag{42}$$

From equation (42), one can see that $\bar{P}(t-1)$ has the same eigenvectors as $P(t-1)$. The eigenvalue of $\bar{P}(t-1)$ associated with $v_p(t-1)$ is $\lambda_p(t-1)/\mu$, and the other eigenvalues are the same as those of $P(t-1)$. This proves (39) and (40). ■

Lemma 5 indicates that only when the data vector $\varphi(t)$ is in the direction of one eigenvector of $P(t-1)$, the eigenvalue associated with this eigenvector increases by the factor $1/\mu$ after the modification. This is quite different from the modification in the EF algorithm, where all eigenvalues increase by the factor $1/\mu$ regardless of the direction of $\varphi(t)$. The proposed algorithm has the ability to modify the information matrix according to its eigenvectors' directions. In this sense, the algorithm presented here works in the same way as that of the selective forgetting method of Parkum *et al.*(1992), where updates of the eigenvalues are performed based on direct calculations of the eigenvalues and eigenvectors.

3 Simulation Example

In the previous section, it has been shown that the proposed algorithm has desirable theoretical properties. In this section, its practical properties are examined via a simulation example. The algorithm is compared with the selective forgetting(SF) method to show that the proposed algorithm has almost the same property as that of the SF algorithm, but is easier to implement in that it does not require explicit computation of the eigenvalues.

The modified covariance matrix in SF method is given by Parkum *et al.*(1992) as

$$\bar{P}(t-1) = \sum_{i=1}^{n} \frac{\alpha_i(t-1)}{\mu_i} v_i(t-1)v_i^T(t-1) \tag{43}$$

where $\alpha_i(t-1)$ is an eigenvalue of $P(t-1)$, and μ_i is the forgetting factor corresponding to α_i. It is argued that μ_i can be chosen as an increasing function of α_i in (Parkum,1992). Then the update equation for the ith eigenvalue of $\bar{P}(t-1)$ is given by

$$\bar{\alpha}_i(t-1) = f(\alpha_i(t-1)) \tag{44}$$

Although there are many possible choices for $f(\cdot)$, here we take the following one suggested by Parkum *et al.*(1992).

$$f(x) = \begin{cases} x, & x > \alpha_{max} \\ \alpha_{min} + (1 - \alpha_{min}/\alpha_{max})x, & x \le \alpha_{max} \end{cases} \tag{45}$$

The example is taken from (Parkum *et al.*,1992). The system to be estimated is given by

$$y(t) + ay(t-1) = bu(t-1) + e(t) \tag{46}$$

where $\{e(t)\}$ is a white noise sequence with variance 0.01. The parameters are given by

$$a = \begin{cases} -0.8 & 0 \le t \le 100 \\ -0.4 & t > 100 \end{cases} \tag{47}$$

$$b = 1.0, \quad t \geq 0 \tag{48}$$

The input to the system and output from the system are shown in figure 1. The input is persistently exciting for $0 \leq t \leq 350$.

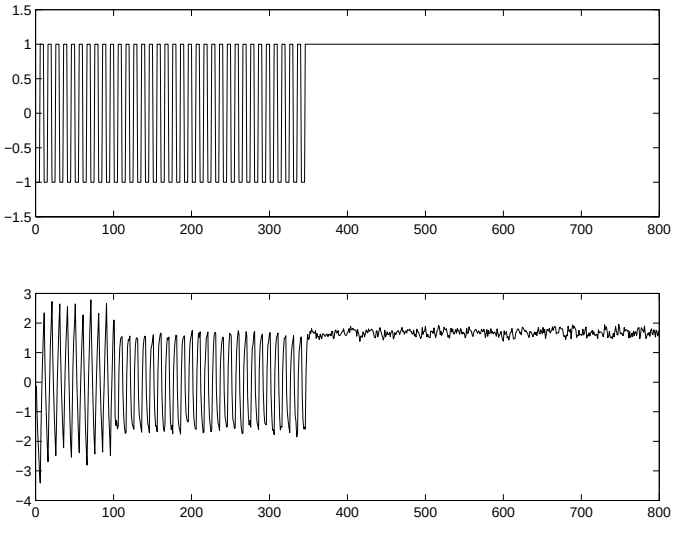

Figure 1: Input(up) and output(down) of the system

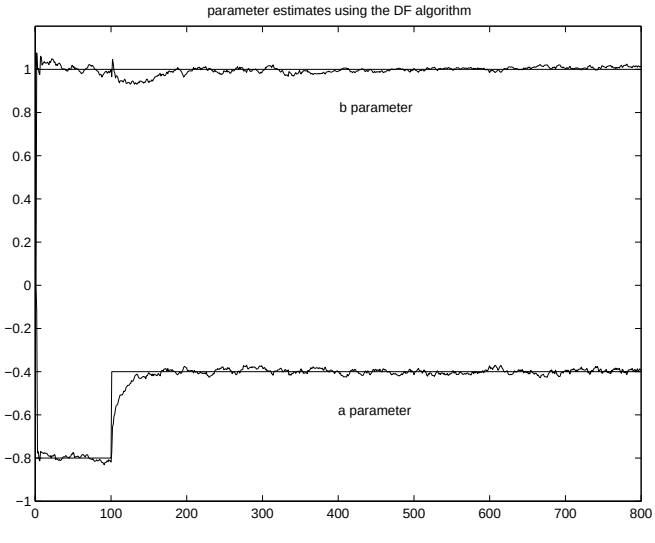

Figure 2: Parameter estimates using the proposed DF method

In order to give a fair comparison in terms of noise sensitivity between these algorithms, we choose the $\mu = 0.87$ for the proposed method, and $\alpha_{min} = 0.01, \alpha_{max} = 0.1$ for the SF method. Both algorithms are started with $P_0 = 1000I$.

The results of the estimation using both methods are presented below. Figure 2 and figure 3 show the parameter estimates. From these figures we see that the tracking ability of both algorithms is basically the same. Figure 2 shows that our algorithm is able to track fast parameter changes, which is in sharp contrast to the DF method proposed in (Kulhavý,1987). Like the SF method, our algorithm also behaves well during the period when the input is a constant. This

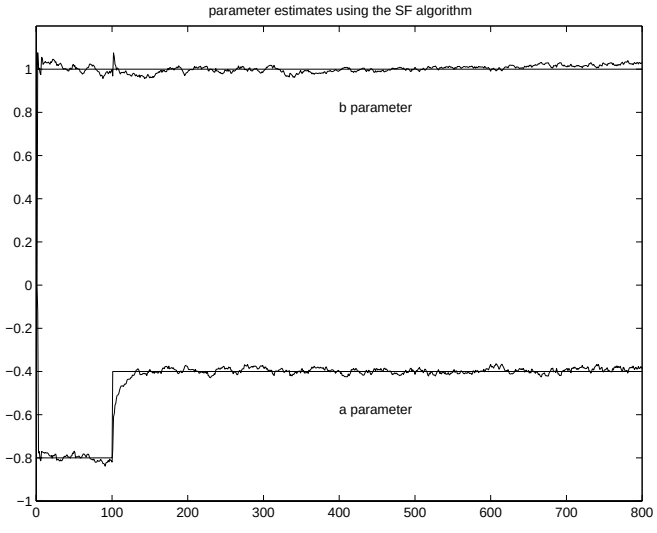

Figure 3: Parameter estimates using the SF method

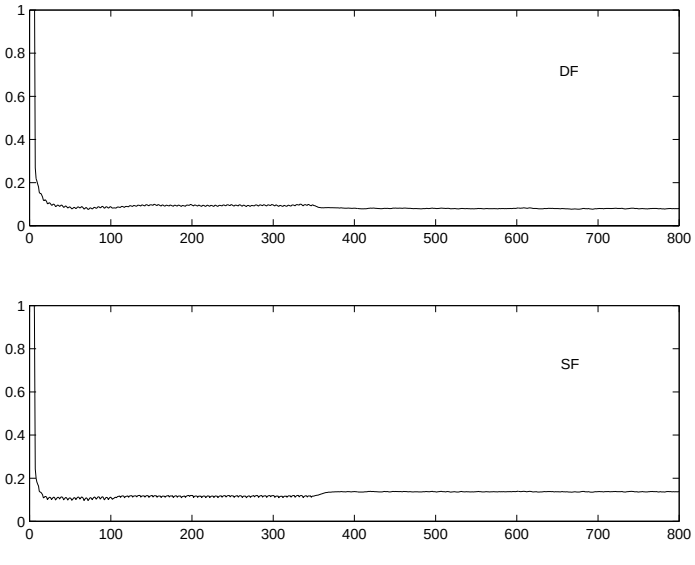

Figure 4: Trace of the covariance matrix

is illustrated in figure 4, where it can be seen that the trace of the covariance matrix tends to a constant for both algorithms, and no windup occurs.

4 Conclusion

In this paper, a new directional forgetting algorithm based on the decomposition of the information matrix has been developed. Theoretical and simulation studies have shown that this algorithm has desirable properties, such as to forget old data according to incoming information in various directions, the ability to track fast parameter changes, satisfactory dynamic behavior and prevention of the windup phenomenon. The algorithm has almost the same order complexity as the standard exponential forgetting algorithm, and the computational requirement is also low. These properties make this algorithm a very attractive selection for on-line identification.

References

Åström,K.J and B.Wittenmark(1995).*Adaptive Control*, second edition, Addison-Wesley, Reading.

Hägglund T.(1985)."Recursive Estimation of Slowly Time-varying Parameters," in *The 7th IFAC Symposium on Identification and System Parameter Estimation*, pp.1137-1142.

Kulhavý,R.(1987)."Restricted Exponential Forgetting in Real-time Identification," *Automatica*, **23**, no. 5, pp.589-600.

Parkum,J.E., N.K. Poulsen and J. Holst(1992)."Recursive Forgetting Algorithms," *International Journal of Control*, **55**, no. 1,pp.109-128.

Bittanti,S., P. Bolzern and M. Campi(1990)."Convergence and Exponential Convergence of Identification Algorithms with Directional Forgetting Factor,"*Automatica*, **26**, no. 5,pp.929-932.

Horn,R.A and C.R.Johnson(1985).*Matrix Analysis*, Cambridge University Press, Cambridge.