ON THE CONVERGENCE OF BAYESIAN ADAPTIVE FILTERING

Tayeb Sadiki, Dirk T.M. Slock

Eurecom Institute
2229 route des Crêtes, B.P. 193, 06904 Sophia Antipolis Cedex, FRANCE
Tel: +33 4 9300 2656/2606; Fax: +33 4 9300 2627
e-mail:{sadiki,slock}@eurecom.fr

ABSTRACT

Standard adaptive filtering algorithms, including the popular LMS and RLS algorithms, possess only one parameter (step-size, forgetting factor) to adjust the tracking speed in a non-stationary environment. Furthermore, existing techniques for the automatic adjustment of this parameter are not totally satisfactory and are rarely used. In this paper we pursue the concept of Bayesian Adaptive Filtering (BAF) that we introduced earlier, based on modeling the optimal adptive filter coefficients as a stationary vector process, in particular a diagonal AR(1) model. Optimal adaptive filtering with such a state model becomes Kalman filtering. The AR(1) model parameters are determined with an adaptive version of the EM algorithm, which leads to linear prediction on reconstructed optimal filter correlations, and hence a meaningful approximation/estimation compromise. In this paper we will introduce the convergence behavior of the adaptive part.

1. INTRODUCTION

In Bayesian Adaptive Filtering (BAF) [1], the evolution of filter coefficients is modeled as a stationary process. A simple choice for search process is first-order autoregressive process (AR(1)). This AR(1) model can be considered a state model. Hence Bayesian Adaptive Filtering leads to Kalman filtering. This Kalman filtering needs to be adaptive because the model parameters are unkown. Even though adaptive Kalman filtering is a difficult problem, a surprisingly large number of solutions exist. The following approaches can be identified:

1. Recursive Prediction-Error Method (RPEM)
2. Extended Kalman Filter (EKF)
3. Best Quadratic Unbiased Estimator (BQUE)
4. Expectation-Maximization (EM)
5. Second-Order Statistics (SOS)
6. Subspace-Based Estimation Method (SBEM)

A common approach is the well-known Recursive Prediction-Error Method (RPEM), which provides an estimator that minimizes a prediction error criterion function $V_N(\theta)$, of the form

$$\hat{\theta} = arg\min_{\theta} V_N(\theta) \tag{1}$$

where θ is the set of parameters to be estimated. However, for many scenarios (1) has no closed-form solution, due to non convexity of $V_N(\Theta)$ in Θ. A popular choice for solving the optimization problem are gradient-based search techniques, and therefore implementation complexity becomes similar to the ML approach. A standard state estimation method used for polynomial systems is the Extended Kalman Filter (EKF), which allows simultaneous estimation of states and parameters through a Recursive prediction-correction model [2]. As an approximate conditional mean filter, the EKF is suboptimal. A popular and robust alternative to these algorithms is provided by the subspace-based estimation methods [3]. These algorithms extract the estimates of system state-space matrices directly from data by first dividing that data into past and future data and then projecting the future data onto the space spanned by the past data. A bank of Kalman filters is employed to compute the estimation of the state sequence, which results in an approximation of Kalman filter estimate of the state. A good alternative to the described schemes is given by the EM algorithm, where the estimate of the state sequence is found by a single Kalman smoothed estimation instead. In this case, the smoothed state estimates are calculated under the assumption that the parameters of the true system are the same as the current estimate. Other approaches like Second-Order Statistics (SOS) methods and Best Quadratic Unbiased Estimator (BQUE) can be found in [4]. In our work we focus on EM parameter estimation techniques.

Some general references on the tracking behavior of adaptive filtering algorithms are [5], [6], [7], [8], [9], [10], [11], [12], [10]. The KF'ing framework can be straightforwardly extended to incorporate time-varying optimal parameters. The simplest way is probably through the following $AR(1)$ model state equation for optimal filter variation

$$y_k = X_k^H H_{k-1}^0 + v_k \tag{2}$$

where X_k is a $M \times 1$ input compelx vector and M is the length of the filter. The error or noise term v_k is assumed to be zero-mean uncorrelated normally distributed noise vector with common covariance matrix R. The BF series H_k^0 is assumed to be of primary interest [1]; it is modelled as a first order multivariate process of the form

$$H_k^0 = AH_{k-1}^0 + w_k \tag{3}$$

where $E[W_k W_k^H] = Q$ is a $M \times M$ transition matrix describing the way the underlying series move through successive time periods. The BF H_k^0 may be non-stationary since we do not make special assumptions about the roots of the characteristic equation A. The $M \times 1$ noise terms W_k, are zero-mean uncorrelated normal vectors with common covariance matrix Q.

Eurécom's research is partially supported by its industrial partners: Hasler Foundation, Swisscom, Thales Communications, ST Microelectronics, CEGETEL, France Télécom, Bouygues Telecom, Hitachi Europe Ltd. and Texas Instruments. The work reported herein was also partially supported by a PACA regional scholarship.

The motivation for the model defined by (2) and (3) originates from a desire to account separately for uncertainties in the model as defined by model error W_k and uncertainties in measurements made on the model as expressed by the measurement noise process v_k. It might be helpful to envision (2) as kind of random effects model for time varying, where the effect vector H_k^0 has a correlation structure over time imposed by the multivariate autoregressive model (3). In this context, it is a generalization of the ordinary autoregressive *AR* model which accounts for observation noise as well as model induced noise. One may regard the X_k^H as fixed design input vector which define the way we observe the components of the BF H_k^0. In this paper, we provide a convenient method for dealing with the incomplete data problem introduced by missing observations.

The primary aim of a smoothing procedure is to estimate the unobserved time-varying H_k^0. If one knows the values for the parameters Q and A the conventional Kalman smoothing estimators can be calculated as conditional expectations and will have MMSE .

Since the smoothed values in a Kalman filter estimator will depend on the initial values assumed for the above parameters, it is of interest to consider various ways in which they might be estimated. In most cases this has been accomplished by Maximum likelihood techniques involving the use of scoring or Newton-Raphson technique to solve the nonlinear equations wich result from differentiating the log-likelihood function. In this paper, we introduce an EM approach for iteratively update the parameter model. Experimental results will be shown for the proposed algorithm, comparing to KF filtering.

2. PARAMETER ESTIMATION VIA THE EM ALGORITHM

In this section we develop the EM algorithm for estimating the parameters of (3)-(4) [1]. Perhaps the most important step in applying the EM algorithm to a particular problem is that of choosing the missing data. The missing data should be chosen so that the task of maximizing $U(\theta,\theta^k)$ for any value of $\theta=(A,Q)$ is easy and so that it is possible to perform the expectation step.

Fortunately, in this case, the choice of missing data is not too difficult. Let us imagine for a moment that, in addition to the system inputs and outputs, X_k and Y_k respectively, the state H_k^0 was avaible then ML estimation of A reduces to applying to (3). The covariance elements, Q, of W_k could then be calculated from the residuals. Moreover, the conditional expectation of state sequence may be calculated using a (slightly augmented) Kalman Smoother. All of this suggests that the state sequence is a desirable conditionate for the missing data. We therefore designate Y as the incomplete data so that the complete data set is $Z=(H_k^0,Y_k)$.

In order to develop a procedure for estimating the parameters in the state-space model defined by (5) and (6), we note first that the joint log-likelihood of the complet data Z can be written in the form

First, by repeated application of Bayes Rule

$$f_Z(z,\theta)=f(H^0|Y=y).f_Y(y;\theta) \tag{4}$$

where $f_Z(z,\theta)$ is the probability vdensity associated with Z and $f_{Z|Y=y}(z,\theta).f_Y(y;\theta)$ is the conditional probability density of Z given $Y=y$. Taking the logarithm on both sides of (4),

$$\log f_Y(y,\theta)=\log f_Z(z,\theta)-\log f(H^0|Y=y) \tag{5}$$

Note that the logarithm function is monotonic in its semi-positive argument and any probability density function (p.d.f.) is semi-positive, it follows that the maximising argument of any p.d.f. will be the same as for the logarithm of that function.

Of course equation (5) requires knowledge of the complete data set and therfore cannot be calculated. Suppose that, instead of calculating equations (5), we caclulate an approximation of (5) derived as an expectation over the space of H_N^0, and conditioned upon the actual observations, as well as some estimate of the vector θ say $\hat{\theta}$ then we obtain

$$E_{\hat{\theta}}\{log f_Y(y,\theta)\}=E_{\hat{\theta}}\{\log f_Z(z,\theta)\}-E_{\hat{\theta}}\{\log f(H^0|Y=y)\} \tag{6}$$

or alternatively,

$$L(\theta)=U(\theta,\hat{\theta}_k)-V(\theta,\hat{\theta}_k)$$

where the following definitions have been used.

$$\begin{aligned} L(\theta) &= \log f_Y(y,\theta) \\ U(\theta,\hat{\theta}_k) &= E_{\hat{\theta}}\{\log f_Z(z,\theta)\} \\ V(\theta,\hat{\theta}_k) &= E_{\hat{\theta}}\{\log f(H^0|Y=y)\} \end{aligned}$$

We can interpret the function $U(\theta,\hat{\theta}_k)$ as the projection of the likelihood function that we want to solve onto the space spanned by Z and in directions informed by $\hat{\theta}$. In other words, it is our estimate of the log-likelihood function associated with the complet data.

With this definition we can writ

$$\begin{aligned} L &= \log f_\theta(H_k^0,Y_M,\theta|Y_M) \\ &= M\log detQ+M\log detR \\ &\quad+\sum_{k=1}^{M} tr(H_k^0-AH_{k-1}^0)Q^{-1}(H_k^0-AH_{k-1}^0)^H \\ &\quad+\sum_{k=1}^{M} tr(y_k-X_k^HH_{k-1}^0)R^{-1}(y_k-X_k^HH_{k-1}^0)^H \end{aligned} \tag{7}$$

where $\log L$ is to be maximized with respect to parameters A and Q . Since the log-likelihood given above depends on the unobserved data H_k^0, we consider applying the EM algorithm conditionally with respect to the observed Y. That is, the estimated parameters at the $(k+1)-th$ iterate as the values A and Q which maximize

$$U(\theta,\hat{\theta}_k)=E_{\hat{\theta}_k}\{\log f_\theta(H_k^0,Y_M,\theta|Y_M)\} \tag{8}$$

where $E_{\hat{\theta}_k}$ denotes the conditional expectaion relative to a density containing the kth iterate values.

In order to calculate the conditional expectation defined in (7), it is convenient to define the conditional mean

$$\begin{aligned} \hat{H}_k^0 &= E_{\hat{\theta}_k}\{H_k^0|Y_M\} \\ \hat{P}_k &= E[\tilde{H}_k^0\tilde{H}_k^{0H}|Y_M] \\ \hat{P}_{k-1} &= E[\tilde{H}_{k-1}^0\tilde{H}_{k-1}^{0H}|Y_M] \end{aligned} \tag{9}$$

we suppose the following definitions

$$\begin{aligned} B1 &= \sum_{k=1}^{M}(E_{\hat{\theta}_k}\{H_{k-1}^0(H_{k-1}^0)^H|Y_M\}+\hat{P}_{k-1}) \\ B2 &= \sum_{k=1}^{M}(E_{\hat{\theta}_k}\{H_k^0(H_k^0)^H|Y_M\}+\hat{P}_k) \\ B12 &= \sum_{k=1}^{M}(E_{\hat{\theta}_k}\{H_k^0(H_{k-1}^0)^H|Y_M\}+\hat{P}_{k,k-1}) \end{aligned} \quad (10)$$

The Kalman filter terms $\hat{H}_k^0$, $\hat{P}_k$ and $\hat{P}_{k,k-1}$ are computed under the parameter values $A^{(k)}$ and $Q^{(k)}$ using the recursions in (7).Furthermore, it is easy to see that the choices

$$\hat{\mathbf{Q}} = \frac{1}{M}(B2_k - B12_k B1_k^{-1} B12_k^H) \quad (11)$$

$$\hat{\mathbf{A}} = B12_k B1_k^{-1} \quad (12)$$

maximize the last two lines in the likelihood function (7).

3. ADAPTIVE KALMAN ALGORITHM

In our study, the tasks of smoothing in a missing data context are interpreted as basically the problem of estimating the BAF H_k^0 in the state-space model (2)-(3). The conditional means provide a minimum MSE solution based on the observed data. The parameters Q and A are estimated by ML using the EM algorithm. We simplify the estimation problem by considering A and Q diagonal matrices. The filter parameters are iteratively computed through M iterations. The estimation of the optimal filter variation is carried out by KF'ing and one step smoothing and we introduce an EM approach for iteratively update the parameter model.

4. CONVERGENCE PROPERTIES OF THE EM BASED ALGORITHM

For any estimation algorithm, convergence properties are of major importance [13], [14], [15]. Withim this paper are a collection of some of the more useful results pertaining to the EM algorithm. The properties of the EM algorithm it will frequently be convenient to think of it as merely a way of generating mappings from one parameter vector to another. To see how multiple instances of the EM algorithm can be generated, notice that once the observed data set and the initial parameter vector is specified the EM algorithm is entirely autonomous, generating a whole sequence of parameter vectors and requiring no further user interaction. On the other hand, if a different set of data is observed and used in the EM calculations, then the resulting sequence of estimates will be different to the first, even if the initialisation is unchanged. We shall describe these obsrvation-dependent mappings as being instances of an EM algorithm. In order to clarify the notion of the EM algorithm as an iterative mapping we provide the following definition.

Defenition- *An iterative algorithm with mapping* $M(.):\Theta \mapsto \Theta$ *is an EM algorithm if*

$$U(M(\theta),\theta_k) \geq U(\theta,\hat{\theta}_k)$$

. *for every pair* $(\theta,\hat{\theta}_k) \in \Theta$.

Let consider that an EM algorithm does converge to a sngle optimal element $\theta^o \in \Theta$.

Suppose that $\hat{\theta}_k$ is an instance of an EM algorithm such that

1. $\hat{\theta}$ converge to θ^o
2. $\frac{\partial}{\partial\theta}U(\theta,\hat{\theta}_k)=0$
3. $\frac{\partial^2}{\partial\theta\partial\theta^T}U(\theta,\hat{\theta}_k)$ is negative definite with eigenvalues bounded away from zero.

Then

$$\frac{\partial}{\partial\theta}L(\theta)=0,$$

$$\frac{\partial^2}{\partial\theta\partial\theta^T}U(\theta,\hat{\theta}_k) \text{ is negative definite}$$

and

$$\frac{\partial}{\partial\theta}M(\theta)=[\frac{\partial^2}{\partial\theta\partial\theta^T}U(\theta,\theta^o)]^{-1}\frac{\partial^2}{\partial\theta\partial\theta^T}V(\theta,\theta^o) \quad (13)$$

In order to see the utility of this items, note that if we linearise the EM algorithm about the point to wich it is converging, by finding its first-order Taylor expansion then we obtain

$$\begin{aligned} \hat{\theta}_{k+1} &= M(\hat{\theta}_k) \\ &\approx \theta^o + \frac{\partial}{\partial\theta}M(\theta)|_{\theta=\theta^o}(\hat{\theta}_k-\theta^o) \end{aligned}$$

then

$$\tilde{\theta}_{k+1} \approx (\frac{\partial}{\partial\theta}M(\theta)|_{\theta=\theta^o})^{N-1}\tilde{\theta}_o \quad (14)$$

Equation (14) formulate the EM algorithm as an autonomouslinear time-invariant. Under such conditions it is well known that θ_k will converge to an optimal value at an exponential rate determined by the largest eigenvalue of $\frac{\partial}{\partial\theta}M(\theta)$.

We shall discuss in greater depth the rate of convergence of the EM algorithm in the light of equation (14).
Note that eqaution (13) may be re-expressed as

$$\begin{aligned} \frac{\partial}{\partial\theta}M(\theta) &= [\frac{\partial^2}{\partial\theta\partial\theta^T}U(\theta,\theta^o)]^{-1}\frac{\partial^2}{\partial\theta\partial\theta^T}V(\theta,\theta^o) \\ &= [\frac{\partial^2}{\partial\theta\partial\theta^T}U(\theta,\theta^o)]^{-1} \\ &\quad \times[\frac{\partial^2}{\partial\theta\partial\theta^T}U(\theta,\theta^o)-\frac{\partial^2}{\partial\theta\partial\theta^T}L(\theta)] \\ &= I-[\frac{\partial^2}{\partial\theta\partial\theta^T}U(\theta,\theta^o)]^{-1}\frac{\partial^2}{\partial\theta\partial\theta^T}L(\theta)] \\ &= I-\Gamma_{aug}^{-1}\Gamma_{obs} \end{aligned} \quad (15)$$

by using equation (7)

$$\Gamma_{aug}=\frac{\partial^2}{\partial\theta\partial\theta^T}E_{\hat{\theta}}\{\log f_Z(z,\theta)\}|_{\theta=\theta^o} \quad (16)$$

is the expected information matrix of the complete data set and

$$\Gamma_{obs}=\frac{\partial^2}{\partial\theta\partial\theta^T}E_{\hat{\theta}}\{\log f_Y(Y,\theta)\}|_{\theta=\theta^o} \quad (17)$$

is the observed information matrix.
Note that the rate of convergence of the EM algorithm as shown by equation (14)is dictated by the largest eigenvalue of $\frac{\partial}{\partial\theta}M(\theta)$. If this eigenvalue has a magnitude close to unity then the algorithm will be slow to converge. Conversely, fast convergence correspond to this eigenvalue being close to zero. under this scenario it follows from eqation (15) that it is desirable to choose the missing data, filter coefficient sequence, so that the smallest eigenvalue of $\Gamma_{aug}^{-1}\Gamma_{obs}$ is as large as possible. Clearly, Γ_{obs} is independent of the missing data so therfore the key to ensuring fast convergence is to find filter coefficient sequence so that Γ_{aug} is small.
Let consider the system defined by (2)-(3), to simpilify we make a Component-Wise system.
The system (2)-(3) becomes $for\ n = 1 \dots M$, where M is the length of the filter

$$h_{k,n}^o = a_n h_{k-1,n}^o + w_{k,n} \quad (18)$$

$$y_k = h_{k-1,n}^o x_{k,n} + \sum_{j\neq n}^{M} h_{k-1,n}^o x_{k,n} + v_k \quad (19)$$

we can write

$$y_k - \sum_{j\neq n}^{N} \hat{h}_{k-1,n}^o x_{k,n} = h_{k-1,n}^o x_{k,n} + \sum_{j\neq n}^{M} \tilde{h}_{k-1,n}^o x_{k,n} + v_k$$

In each iteration y_k and v_k are updated as follows

$$y_k' = y_k - \sum_{j\neq n}^{M} \hat{h}_{k-1,n}^o x_{k,n}$$

and

$$v_k' = \sum_{j\neq n}^{M} \tilde{h}_{k-1,n}^o x_{k,n} + v_k$$

where w_k and v_k' are sequences of scalar-valued i.i.d. random variables distribued as $E[w_k w_k^T] = q$ and $E[v_k' v_k'^T] = r$.
for convenience, the parameters of this system shall be collected into the optimal vector $\theta^o = [a^o \ \ q^o]^T$
In order to avoid problems with the information matrices becoming unbounded as we allow the number of data to tend to infinity, we shall, entirely equivalently, employ the average value of this information matrix per sample is defined as

$$\begin{aligned}\Gamma_{aug}^{-} &= \lim_{N\to\infty} \frac{1}{N}\Gamma_{aug} \\ &= \lim_{N\to\infty} \frac{-1}{N}\left[\frac{\partial^2}{\partial\theta\partial\theta^T}U(\theta,\theta^o)\right]|_{\theta=\theta^o} \\ &= \begin{pmatrix} \frac{1}{1-(a^o)^2} & 0 \\ 0 & \frac{q^o}{r(1-(a^o)^2)} \end{pmatrix} \end{aligned} \quad (20)$$

5. CONCLUSION

The global rate of convergence of the EM algorithm is detemined by the eigenvalue of Γ_{aug} small eigenvalues imply fast convergence. Since the eigenvalues of a diagonal matrix are its diagonal elements it's quite clear, from equation (20), how the rate convergence of the algorithm is affected by the system paramters, as the number of data tends to infinity.
The first diagonal element of the matrix in equation (20) will be small if $a^o << 1$, that is, if the underlying system has fast dynamics.

REFERENCES

[1] T. Sadiki and D. T. M. Slock, "Bayesian adaptive feltering: Principles and practical approaches," in *Eusipco Coference*, 16-17 September 2004.

[2] Torsten Soderstrom, *System Identification*, Prentice Hall Sgnal Prossing Series, 1989.

[3] B. Jansson, M.; Goransson and B. Ottersten, *A subspace method for direction of arrival estimation of uncorrelated emitter signals*, IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 47 , Issue: 4,945 - 956, April 1999.

[4] Jerry M. Mendel, *Lessons In Estimation Thory For Signal Prossing, Communications, And Control*, Prentice Hall Signal Prossing Series, 1995.

[5] S. Haykin, *"Adaptive Filter Theory"*, Prentice Hall, 2001, 4th edition.

[6] L. Ljung and T. Soderstrom, *"Theory and Practice of Recursive Identification"*, MIT Press, Cambrdge, MA, 1983.

[7] V. Solo and X. Kong, *"Adaptive Signal Processing Algorithms: Stability and Performance"*, Prentice-Hall, 1994.

[8] M. Niedzwiecki, *"Identification of Time-Varying Systems"*, Wiley, 2000.

[9] B.D.O. Anderson and J.B. Moore, *"Optimal Filtering"*, Prentice-Hall, Englewood Cliffs, NJ, 1979.

[10] Lei Guo and Lennart Ljung, ""Performance Analysis of General Tracking Algorithms"," *IEEE Trans. Automatic Control*, vol. 40, no. 8, Aug. 1995.

[11] L. Guo, ""Estimation Time-Varying Parameters by Kalman Filter Based Algorims: Stability and Convergence"," *IEEE Trans. Automat. Cont.*, vol. 35, pp. 141–147, 1990.

[12] L. Ljung L. Guo, ""Performance Analysis of the Forgetting Factor RLS Algorithm"," *Int. J. Adaptive Cont. Sig. Proces.*, vol. 7, pp. 141–537, 1993.

[13] D. M. Titterington, "Recursive parameter estimation using incomplet data," in *J. R. Statist. Soc. B 46, No. 2, pp. 257-267*, Scotland, 1984.

[14] C. F. Jeef Wu, "on the convergence properties of the em algorithm," in *The Annals of Statistics, Vol. 11, No. 1, 95-103*, 1983.

[15] Yunxin Zhao, "An em algorithm for linear distortion channel estimation based on observations for a mixture of gaussian sources," in *IEEE Trans. on Speech and Audio Processing, vol. 7, No. 4*, July.