Sequential Bayesian Estimation for Adaptive Classification

1Ji Won Yoon, Stephen J. Roberts, Matt Dyson and John Q. Gan *

Abstract—This paper proposes a robust algorithm to adapt a model for EEG signal classification using a modified Extended Kalman Filter (EKF). By applying Bayesian conjugate priors and marginalising the parameters, we can avoid the needs to estimate the covariances of the observation and hidden state noises. In addition, Laplace approximation is employed in our model to approximate non-Gaussian distributions as Gaussians.

Keywords: Nonlinear dynamics, Extended Kalman filter, Marginalisation, Laplace Approximation

1 Introduction

The goal of Brain Computer Interfacing (BCI) is to enable people with severe neurological disabilities to operate computers by manipulation of the brain’s electrical activity rather than by physical means. It is known that the generation and control of electrical brain activity (the electroencephalogram or EEG) signals for a BCI system often requires extensive subject training before a reliable communication channel can be formed between a human subject and a computer interface [1, 2]. In order to reduce the overheads of training and, importantly, to cope with new subjects, adaptive approaches to the core data modelling have been developed for BCI systems. Such adaptive approaches differ from the typical methodology, in which an algorithm is trained off-line on retrospective data, in so much that the process of ‘learning’ is continuously taking place rather than being confined to a section of ‘training data’. In the signal processing and machine learning communities this is referred to as sequential classification. Previous research in this area applied to BCI data has used state space modelling of the time series [3, 4, 5]. In particular Extended Kalman Filter (EKF) approaches have been effective [3, 5].

In this paper, we propose an adaptive classification algorithm extending the work in [3, 5] such that a more robust algorithm is developed by applying a full Bayesian framework to unknown noise covariances upon which the EKF is very sensitive.

In previous models based upon the KF and EKF the variances of the observation noise and hidden state noise are re-estimated using a maximum-likelihood style framework, ultimately due to the work of Jazwinski [6]. It is notoriously difficult to infer these parameters using only on-line sequential data and hence the resultant outcome variables from the methods are prone to systematic errors. In this paper we offer a simple, computationally practical, modification of the EKF approach in which we marginalise (integrate out) the unknown parameters for variances by applying Bayesian conjugate priors [7]. This enables us to avoid the need for spurious parameter estimation steps. Unfortunately this process leads to non-Gaussian distributions in parts of the prediction and filtering steps. We utilise a Laplace approximation to form a set of conditionally Gaussian update steps which may be efficiently implemented in the overall EKF (or KF) framework.

2 Mathematical model

We consider an observed input stream (i.e. a time series) $x_t$ at time $t = 1, 2, \cdots, T$ which we project into a nonlinear basis space, $x \rightarrow \varphi(x)$. We also (partially) observe $z_t \in \{0, 1\}$ such that

$$Pr(z_t = 1|x_t, w_t) = g(f(x_t, w_t))$$  \hspace{1cm} (1)

where $g(\cdot)$ can be logistic model or probit link function which takes the latent variables to the outcome decision variable. In this paper, we use the logistic function, i.e. $g(a) = 1/(1 + e^{-a})$. The state space model for the BCI system can hence be regarded as a hierarchical model in that the noise of observations influences the model indirectly through the logistic regression model $g(\cdot)$. Such indirect influence makes for a more complicated model, but we can circumvent much of this complexity by forming a three stage state space model and by introducing an auxiliary variable $y_t$. The latter variable acts so as to link the indirect relationships between observations and the logistic regression model given by

$$p(z_t|y_t) = g(y_t)^{z_t}(1 - g(y_t))^{1-z_t}$$

$$y_t = w_t^T \varphi_t(x_t) + v_t$$

$$w_t = w_{t-1} + h_t$$  \hspace{1cm} (2)

where $v_t \sim N(0, R)$ and $h_t \sim N(0, Q) = N(0, qI)$. In this paper, we use a simplified notation $\varphi_t$ instead of $\varphi_t(x_t)$.
\[ \varphi_t = \varphi_t(x_t) = \begin{bmatrix} x_t \\ \phi_t^{(1)}(x_t) \\ \vdots \\ \phi_t^{(N_b)}(x_t) \\ 1 \end{bmatrix} \]

and \( \phi_t^{(i)}(x_t) \) is the \( i \)-th Gaussian basis function for \( i \in \{1, \ldots, N_b\} \) at time \( t \). Here, \( N_b \) is the number of Gaussian basis functions.

In the adaptive classification, there are two steps in our state space model which perform model inference:

- **Prediction:** \( \dot{w}_{t|t-1} = \int w_t p(w_t|z_{1:t-1}, x_{1:t-1}) dw_t \)
- **Filtering:** \( w_{t|t} = \int w_t p(w_t|z_{1:t}, x_{1:t}) dw_t \)

### 2.1 Prediction \( p(w_t|z_{1:t-1}, x_{1:t-1}) \)

To build the prediction equation, we need a posterior distribution, \( p(w_t|z_{1:t-1}, x_{1:t-1}) \) given by

\[
p(w_t|z_{1:t-1}, x_{1:t-1}) = \int_q \int_{w_{t-1}} p(w_t, w_{t-1}, q|z_{1:t-1}, x_{1:t-1}) dw_{t-1} dq \]

\[
= \int_q \int_{w_{t-1}} p(w_t|w_{t-1}, q) p(q) \times p(w_{t-1}|z_{1:t-1}, x_{1:t-1}) dw_{t-1} dq \\
= \int_{w_{t-1}} \int_q \left[ p(w_t|w_{t-1}, q)p(q) dq \right] dw_{t-1} \\
\times \int_q p(w_t|w_{t-1}, q)p(q) dq \\
= \int_{w_{t-1}} \int_q \left[ p(w_t|w_{t-1}, q)p(q) dq \right] dw_{t-1} \\
\times \left[ \int_q (p(w_t|w_{t-1}, q)p(q)) dq \right] dw_{t-1} \\
= \int_{w_{t-1}} \left[ \int_q \left( \frac{(2\pi)^{N_w/2} \beta^\alpha \Gamma(\alpha_q) \Gamma(\alpha_r)}{\beta_q + \frac{1}{2} (w_t - w_{t-1})^T (w_t - w_{t-1})} \right)^{-\alpha_q} dw_{t-1} \right] \\
\times \left( \frac{(2\pi)^{N_w/2} \beta^\alpha \Gamma(\alpha_q) \Gamma(\alpha_r)}{\beta_q + \frac{1}{2} (w_t - w_{t-1})^T (w_t - w_{t-1})} \right)^{-\alpha_q} dw_{t-1} \]

(4)

where \( p(q) \) is a conjugate Inverted Gamma distribution and \( IG(\cdot; \alpha, \beta) \) denotes the Inverted Gamma distribution with hyper-parameters \( \alpha \) and \( \beta \) in which \( \alpha_q = \alpha_q + N_w/2 \) where \( N_w \) is the dimension of the coefficients \( w_t \). In order to obtain the conditional Gaussian distribution (which enables the use of the computationally efficient KF framework), we use the Laplace approximation [See appendix 5.1] so that we have

\[
\approx N \left( w_t ; w_{t-1}, \left( \frac{\alpha_q}{\beta_q} \right)^{-1} \right). \]

(5)

Therefore, the posterior for the prediction is rewritten by

\[
p(w_t|z_{1:t-1}, x_{1:t-1}) \\
\approx \int_{w_{t-1}} N(w_t; \mu_{t-1|t-1}, \Sigma_{t-1|t-1}) \\
\times \int_{w_{t-1}} \left( \frac{\alpha_q}{\beta_q} \right)^{-1} dw_{t-1}. \]

(6)

Since the posterior is assumed to be Gaussian

\[
p(w_t|z_{1:t-1}, x_{1:t-1}) \approx N(w_t; \mu_{t-1|t-1}, \Sigma_{t-1|t-1}), \]

we hence have

\[
\mu_{t|t-1} = \mu_{t-1|t-1} \]

\[
\Sigma_{t|t-1} = \Sigma_{t-1|t-1} + \left( \frac{\alpha_q}{\beta_q} \right)^{-1}. \]

(8)

### 2.2 Filtering \( p(w_t|z_{1:t}, x_{1:t}) \)

Our goal is to explore the distribution \( p(w_t|z_{1:t}, x_{1:t}) \) and hence we need to infer the likelihood between \( z_t \) and \( w_t \). We can rewrite this by adding an auxiliary variable which links \( z_t \) and \( w_t \). The rewritten equation with the auxiliary variable is given by

\[
p(w_t|z_{1:t}, x_{1:t}) \\
= \int_{R_{yt}} p(w_t, y_t, R|z_{1:t}, x_{1:t}) dR dy_t \\
= \int_{R_{yt}} p(w_t, y_t, R, z_t, x_t|z_{1:t-1}, x_{1:t-1}) dR dy_t \\
= \int_{R_{yt}} p(z_t|y_t)p(x_t)p(y_t|w_t, x_t, R)p(R) \\
\times p(w_t|z_{1:t-1}, x_{1:t-1}) \\
= \int_{y_t} \left( \int_{R} p(y_t|w_t, x_t, R)p(R) dR \right) dy_t \\
= \int_{y_t} \left( \int_{R} p(y_t|w_t, x_t, R)p(R) dR \right) dy_t \\
= \int_{y_t} \left( \int_{R} p(z_t|y_t)p(y_t|w_t, x_t, R)p(R) dR \right) dy_t \\
= \int_{y_t} \left( \int_{R} p(z_t|y_t)p(y_t|w_t, x_t, R)p(R) dR \right) dy_t \\
= \int_{y_t} \left( \int_{R} p(z_t|y_t)(2\pi)^{-1/2} \beta^\alpha \Gamma(\alpha_r) \beta^\beta \Gamma(\alpha_r) dy_t \right) \\
\times \left( \frac{\alpha_q}{\beta_q} \right)^{-1/2} \beta^\alpha \Gamma(\alpha_q) \beta^\beta \Gamma(\alpha_r) dy_t. \]

(9)

where, as before \( p(R) \) is a conjugate prior using the Inverted Gamma distribution in the same family where \( \alpha_r = \alpha_r + 1/2 \) and \( \beta_r = \beta_r + 1/2(y_t - \varphi^T w_t)^2 \). Then, Eq. (9) is no longer linear due to two factors: the likelihood and the marginalisation over \( R \). Using a Laplace approximation, however, as detailed in the prediction step, we
can evaluate a Gaussian, and hence analytic, approximation given by
\[
(2\pi)^{-1/2} \frac{\beta^{\alpha^r}}{\Gamma(\alpha^r)} \frac{\Gamma(\alpha^r)}{\beta^{2/2} \sigma^r_{i}^2} \approx N\left(y_i; \varphi_T^T w_t, \left( \frac{\alpha^r}{\beta} \right)^{-1}\right).
\]
After obtaining this Gaussian approximation, we can hence make an approximation of the logistic function and Gaussian convolution by using a probit function \([8]\) given by
\[
\int g(y_i) N\left(y_i; \varphi_T^T w_t, R^* \right) dy_i \approx g\left(k(R^*) \varphi_T^T w_t\right)
\]
for \(z_t = 1\) where \(R^* = \left( \frac{\alpha^r}{\beta} \right)^{-1}\) and \(k(R^*) = \left( 1 + \frac{z_t R^*}{\beta} \right)^{-0.5}\). Therefore, we can have
\[
p(w_t | z_{1:t}, x_{1:t}) \approx N(w_i; \mu_{t|0}, \Sigma_{t|0}) g\left(k(R^*) \varphi_T^T w_t\right)^{z_t} \times \left[1 - g\left(k(R^*) \varphi_T^T w_t\right)\right]^{1-z_t} \\
\approx N(w_i; \mu_{t|0}, \Sigma_{t|0}) N(w_i; \mu_{*}, \Sigma_{*})
\]
where
\[
\mu_{*} = -\log(1/z_t - 1)\left(k(R^*) \varphi^*_T\right)^{-1}
\]
\[
\Sigma_{*} = \left\{ z_t (1 - z_t)k(R^*)^2 \varphi^*_T \varphi^*_T \right\}^{-1}
\]
If the value of \(z_t\) is 1 or 0 in Eq. (13), the calculation is numerically unstable so that we change the observations slightly by
\[
z_t = \max(0.1, \min(0.9, z_t)). \quad (14)
\]
Finally, we can have
\[
p(w_t | z_{1:t}, x_{1:t}) = N(w_t; \mu_{t|1}, \Sigma_{t|1})
\]
where
\[
\Sigma_{t|1} = \left( \Sigma_{*}^{-1} + \Sigma_{t|0}^{-1} \right)^{-1}
\]
\[
\mu_{t|1} = \Sigma_{t|1} \left( \Sigma_{*}^{-1} \mu_{*} + \Sigma_{t|0}^{-1} \mu_{t|0} \right).
\]

3 Results for experimental data set

3.1 Data acquisition

Data used in this experiment consisted of two channels of EEG, recorded at 256Hz placed over the central portion of the head and one channel of muscle electrical activity (EMG), recorded at 1024Hz over the muscles of the right fore-arm. The EMG was then down-sampled to 256Hz and muscle contraction strength for movement and non-movement detection was evaluated via a simple windowed peak and trough detection; this then formed a movement / non-movement label.

3.2 Feature extraction and basis formation

The second reflection coefficient of a second-order autoregressive (AR) model [9] were calculated over each EGG signal once every 78ms using a sliding one-second-long window, forming a set of feature vectors \(x_t\). These vectors were projected into a non-linear latent space using a set of Gaussian basis functions, providing the stream of \(\varphi z_t\).

In this paper, hyper-parameters prior distributions over \(Q\) and \(R\) are given by \(\alpha_q = 2\), \(\beta_q = 100\), \(\alpha_r = 5\), and \(\beta_r = 0.1\). These hyper-parameters are chosen with experiences in this paper but we can also estimate the hyper-parameters in time series. This still gives better explanation of the \(Q\) and \(R\) compared to direct estimation of \(Q\) and \(R\).

For comparison, fixed \(\tilde{Q}\) and \(\tilde{R}\) are chosen by
\[
\tilde{q} = E(q | \alpha_q, \beta_q) = \int q p(q | \alpha_q, \beta_q) dq
\]
and
\[
\tilde{R} = E(R | \alpha_r, \beta_r) = \int R p(R | \alpha_r, \beta_r) dR
\]
where \(\tilde{Q} = \tilde{q} I\). In addition, we used 10 basis functions \((N_b = 10)\). Also, the \(p(u_0) = N(0, q_0)\) where \(q_0 = 0\) and \(\Sigma_0 = q_0 I\) respectively. Here \(q_0 = 10000000\).

Fig. 1 shows the comparison of four different methods based on Laplace approximation to make Gaussian distribution. The first plot of the figure shows the labels: inactive \((z_t = 0)\), active \((z_t = 1)\) and unknown labels \((z_t = -1)\) in the movement respectively. Red dot line stands for the ground truth and we changed partly the ground truth into \(z_t = -1\) after the 3000th sample. Blue line stands for the partly observed labels. As we can see, while our EKF with \(\tilde{Q}\) and \(\tilde{R}\) does not find the underlying labels in the missing labels, EKF with marginalisation of \(Q\) and \(R\) finds them. We have checked such a bad performance in EKF with \(\tilde{Q}\) and \(\tilde{R}\) comes from the numerical changes in Eq. (14). However, even the \(Q\) and \(R\) are underestimated, our proposed marginalised EKF obtain reasonable results. By integrating out the \(Q\) and \(R\), we can reduce difficulty of non-stationary EEG signals.

Fig. 2 demonstrates the comparison of the cross entropy of the four different method. The cross entropy of the \(t\)th sample is defined by
\[
\mathbf{E}_i^{cross} = -\log p(z_i | \tilde{z}_i) = -\log \left\{ \tilde{z}_i \left(1 - \tilde{z}_i\right)^{1-z_i} \right\} \quad (17)
\]
where \(z_t\) and \(\tilde{z}_t\) represent for underlying labels and the posterior of estimated labels respectively. As shown in this figure, two modified EKFs with marginalised \(Q\) has much smaller mean and variance of the errors between real ground truth of the labels and estimated labels by four methods.
Figure 1: Comparison of four different EKF algorithms: a) trajectories with true labels (red dots) and missing labels (blue solid), b) $\tilde{Q}$ and $\tilde{R}$ are used, c) $\tilde{Q}$ is used but $R$ is marginalised, d) $Q$ is marginalised but $\tilde{R}$ is used, and e) both $Q$ and $R$ are marginalised.
Using the first derivative of the $L$ where

\[ \frac{\delta^2 L}{\delta w_t^2} \big|_{w_t=w_t-1} = \frac{\alpha_q^*}{\beta_q}. \]  

Now, we have an approximated distribution

\[ \pi(w_t) \approx N \left( w_t; w_{t-1}, \left( \frac{\alpha_q^*}{\beta_q} \right)^{-1} \right). \]  

### References


