Adaptive BCI based on variational Bayes: an empirical evaluation

P. Sykacek[1], S. Roberts[1] and M. Stokes[2]

[1]Robotics Research Group, Department of Engineering Science, University of Oxford, Parks Road, Oxford OX1 6PJ, UK

[2]Research Department, Royal Hospital for Neurodisability, Putney, London, UK.

psyk@robots.ox.ac.uk

http://www.robots.ox.ac.uk/∼parg/

Presented at the BCI workshop in Albany NY, 12-17 June 2002.

Definition and Motivation

- *Adaptive BCI* refers to a brain computer interface (BCI) which is built around adaptively infered classification.
- The proposed method is a classical two stage approach: We first extract *features* from consecutive segments of EEG and build a classifier that translates these features into *probabilities* of cognitive states.
- Previous research suggests that human behaviour *adapts* while performing BCI experiments. $->$ Learning effects?
- Thus *static* methods which are infered only once using some training data can *not* be expected to result in BCI's with *maximal bit rates*.

Feature extraction by autoregressive (AR) processes

$$y[t] = \sum_{m=1}^{p} a_m y[t-m] + \epsilon[t], \text{ with}$$

a_m: m-th order AR coefficient, $y[t]$: a sample of EEG time series, $\epsilon[t]$: sample of white noise.

We extract 3 *reflection coefficients* ρ_m from an EEG segment $\mathcal{Y}$:

$$p(\rho_m|\mathcal{Y}) = \frac{1}{\sqrt{2\pi}s} \exp\left(-\frac{1}{2s^2}(\rho_m - \widehat{\rho}_m)^2\right), \text{ with} \qquad (1)$$

$$\text{m.p. value } \widehat{\rho}_m = -\frac{\boldsymbol{r}_m^{\mathsf{T}}\boldsymbol{\epsilon}_m}{\boldsymbol{r}_m^{\mathsf{T}}\boldsymbol{r}_m} \text{ and variance } s^2 = \frac{1-(\widehat{\rho}_m)^2}{(N-1)}$$

and combine them for the n-th window to feature vector $\boldsymbol{x}_n$.

A generalized nonlinear classifier

$$\boldsymbol{\phi}_n = \begin{bmatrix} 1 \\ \boldsymbol{x}_n \\ \boldsymbol{\varphi}(\boldsymbol{x}_n; \boldsymbol{w}_{\boldsymbol{\varphi}}) \end{bmatrix} \tag{2}$$

$$\eta_n = \boldsymbol{\phi}_n^T \boldsymbol{w} \tag{3}$$

$$P(y_n|\boldsymbol{w}, \boldsymbol{w}_\varphi, \boldsymbol{x}_n) = \frac{1}{1 + \exp((2y_n - 1)\eta_n)}, \quad \text{with} \tag{4}$$

$\boldsymbol{\phi}_n$: projection into nonlinear feature space, y_n: response variable (cognitive state), $\boldsymbol{w}$ and $\boldsymbol{w}_\varphi$: model coefficients.

conditioning on $\boldsymbol{w}_\varphi$ we have likelihood (data of size N):

$$p(\mathcal{D}_N|\boldsymbol{w}) = \prod_{n=1}^{N} P(y_n|\boldsymbol{w}, \boldsymbol{x}_n), \tag{5}$$

Variational Kalman filtering

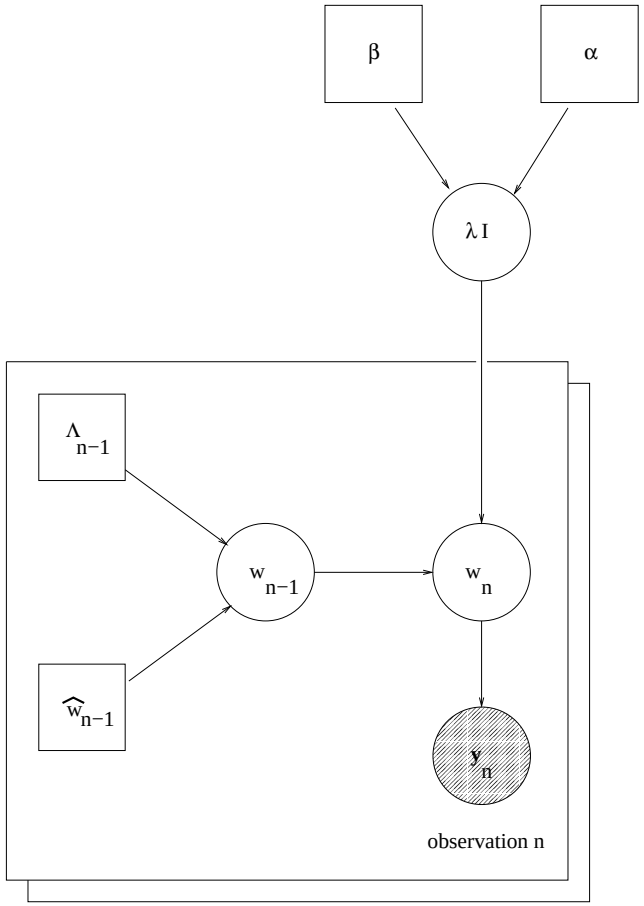
Probabilistic view of adaptive inference $->$ state space formulation of a first order Markov process.

$$
\begin{aligned}
& p(\boldsymbol{w}_{n-1}) \\
& p(\boldsymbol{w}_n|\boldsymbol{w}_{n-1}, \lambda \boldsymbol{I}) \text{ for times } n \geq 1 \\
& p(y_n|\boldsymbol{x}_n, \boldsymbol{w}_n) \text{ for times } n \geq 1, \text{ where}
\end{aligned} \tag{6}
$$

$\boldsymbol{w}_{n-1}$, $\boldsymbol{w}_n$: Gaussian distributed parameters of classifier at consecutive time instances n (EEG segments!), λ: precision of Gaussian process noise *most important parameter!!*.

linear Gaussian case $->$ Kalman filter. Here nonlinear due to logistic sigmoid $->$ propose to use *variational Kalman filter*.

Adaptive learning as probabilistic model



Directed acyclic graph describing a hierarchical model for adaptive inference. Posterior $p(\boldsymbol{w}_{n-1}|\hat{\boldsymbol{w}}_{n-1}, \Lambda_{n-1})$ and $\lambda \boldsymbol{I}$ define prior for $\boldsymbol{w}_n$. Use non informative proper Gamma prior for hyper parameter λ. For inference assume constant *adaption rate* within a window.

Log evidence of a segment of size N

Assumption in Kalman filtering: successive pairs $(\boldsymbol{w}_{n-1}, \boldsymbol{w}_n)$ are statistically independent.

$$\begin{aligned}
\log(p(\mathcal{D}_N)) &= \log\Big(\int_\lambda \prod_{n=1}^{N} \Big[\int_{\boldsymbol{w}_n} (2\pi)^{-\frac{d}{2}} |\boldsymbol{\Lambda}_{n-1}^{-1} + \lambda^{-1}\boldsymbol{I}|^{-\frac{1}{2}} \\
&\times \exp(-0.5(\boldsymbol{w}_n - \hat{\boldsymbol{w}}_{n-1})^T (\boldsymbol{\Lambda}_{n-1}^{-1} + \lambda^{-1}\boldsymbol{I})^{-1} (\boldsymbol{w}_n - \hat{\boldsymbol{w}}_{n-1})) \\
&\times (1 + \exp((2y_n - 1)\boldsymbol{\phi}_n^T \boldsymbol{w}_n))^{-1} d\boldsymbol{w}_n\Big] \\
&\times \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{(\alpha-1)} \exp(-\beta\lambda) d\lambda\Big)
\end{aligned} \tag{7}$$

Equation (7) suggests a variational approximation of $p(\boldsymbol{w}_1, ..., \boldsymbol{w}_N, \lambda)$ by $Q(\lambda)\prod_n Q(\boldsymbol{w}_n)$ with $Q(\lambda)$: Gamma distribution and $Q(\boldsymbol{w}_n)$: Gaussian distributions. Note: Equation (7) is *not* yet conjugate with this suggestion! We need additional lower convex bounds.

Finally: Negative free energy (lower bound of (7))

$$F(\{Q(\boldsymbol{w}_n),\ \forall\ n\}, Q(\lambda), \{\xi_n,\ \forall\ n\}, \nu) = \int_\lambda Q(\lambda)\Big\{\alpha \log(\beta) \tag{8}$$

$$- \log(\Gamma(\alpha)) + (\alpha - 1)\log(\lambda) - \beta\lambda + \sum_{n=1}^{N}\Big[\int_{\boldsymbol{w}_n} Q(\boldsymbol{w}_n)\Big(-\frac{d}{2}\log(2\pi)$$

$$+\frac{d}{2}\log(\lambda) - \frac{1}{2}\log|\nu\boldsymbol{\Lambda}_{n-1}^{-1} + \boldsymbol{I}| - \frac{1}{2}(\lambda - \nu)\mathrm{tr}(\nu\boldsymbol{I} + \boldsymbol{\Lambda}_{n-1})^{-1}$$

$$-0.5(\boldsymbol{w}_n - \hat{\boldsymbol{w}}_{n-1})^T(\boldsymbol{\Lambda}_{n-1}^{-1} + \nu^{-1}\boldsymbol{I})^{-1}(\boldsymbol{w}_n - \hat{\boldsymbol{w}}_{n-1})$$

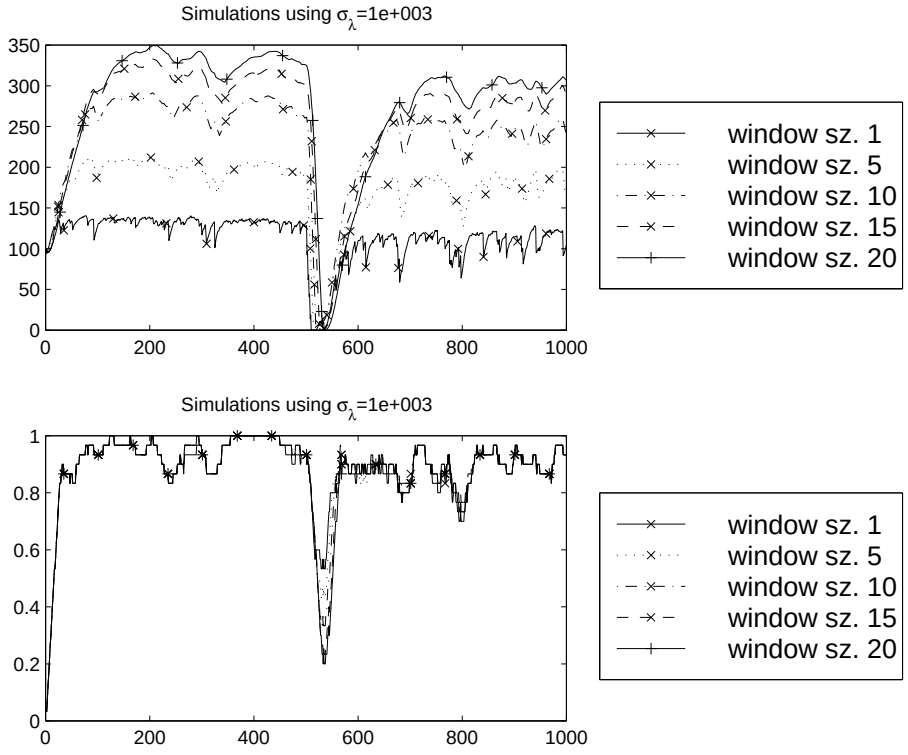
$$-0.5(\lambda - \nu)(\boldsymbol{w}_n - \hat{\boldsymbol{w}}_{n-1})^T(\nu\boldsymbol{\Lambda}_{n-1}^{-1} + \boldsymbol{I})^{-2}(\boldsymbol{w}_n - \hat{\boldsymbol{w}}_{n-1})$$

$$-\frac{(2y_n - 1)\boldsymbol{w}_n^T\boldsymbol{\phi}_n}{2} - \log(2) - \log(\cosh(\frac{\xi_n}{2})) - \frac{\tanh(\frac{\xi_n}{2})}{4\xi_n}$$

$$\times\Big(\Big(\frac{\boldsymbol{w}_n^T\boldsymbol{\phi}_n}{2}\Big)^2 - \xi_n^2\Big) - \log(Q(\boldsymbol{w}_n))\Big)d\boldsymbol{w}_n\Big] - \log(Q(\lambda))\Big\}d\lambda,\ \text{which}$$

we maximize w.r.t parameters of all $Q(\boldsymbol{w}_n)$, $Q(\lambda)$, all ξ_n and ν.

Tracking and stationary accuracy on synthetic data



Non stationarity by switching class labels in the second half of the data. Above graph: Expectation $< \lambda >_{Q(\lambda)} = \frac{\widehat{\alpha}}{\widehat{\beta}}$ which corresponds to adaptation rate. Second graph: instantaneous generalization accuracy estimated in a window of size 30. Gamma Prior over λ with expectation 100 and variance 10^6.

Based on such experiments: reasonable compromise between stationary accuracy and tracking with a window size to 10 and $\alpha = 0.01$ and $\beta = 10^{-4}$.

BCI experiments

- Comparison with equivalent static classifier that is infered with sequential variational inference (non adaptive method).

- We measure *generalization accuracy* on independent test data and check for statistical significance using Mc.Nemar's test (a test for paired experiments).

- We also estimate the BCI's *bit rate* and check for significance of different bit rates using a Kolmogorov-Smirnov test.

Data - common properties in both studies

- Recorded using an ISO-DAM amplifier using gain 10^4 and analogue band pass filter with pass band between 0.1 Hz and 100 Hz. We sample at 384 Hz and 12 bit resolution.
- Feature extraction is based on the proposed method, extracting 3 reflection coefficients for each electrode and second. These features are labeled according to the cognitive task the subject was supposed to do in the respective time interval.
- In order to make the comparison fair, we use a two fold cross testing by split the data obtained in every trial in 2 halfs and allow the classifier to converge using on one half before assessing the performance on the other half. All results are *within subject* though averaged over all subjects that participate in the study.

Generalization accuracy study one

Characteristics: 8 subjects, two sets of trials 1.) no cognitive activity (rest EEG) vs. imagined movements and 2.) a mathematical task vs. imagined movements, differential electrodes at C3'-C4' with reference behind left mastoid, 10 repetitions of each task done for 10 seconds, once without and once with visual feedback.

Predicting for every second *without reject option* we get:

Cognitive task	Generalization results		
	vkf	vsi	P_{null}
rest/move, no feedback	0.69	0.61	0.00
rest/move, feedback	0.71	0.70	0.39
move/math, no feedback	0.69	0.62	0.00
move/math, feedback	0.64	0.60	0.00

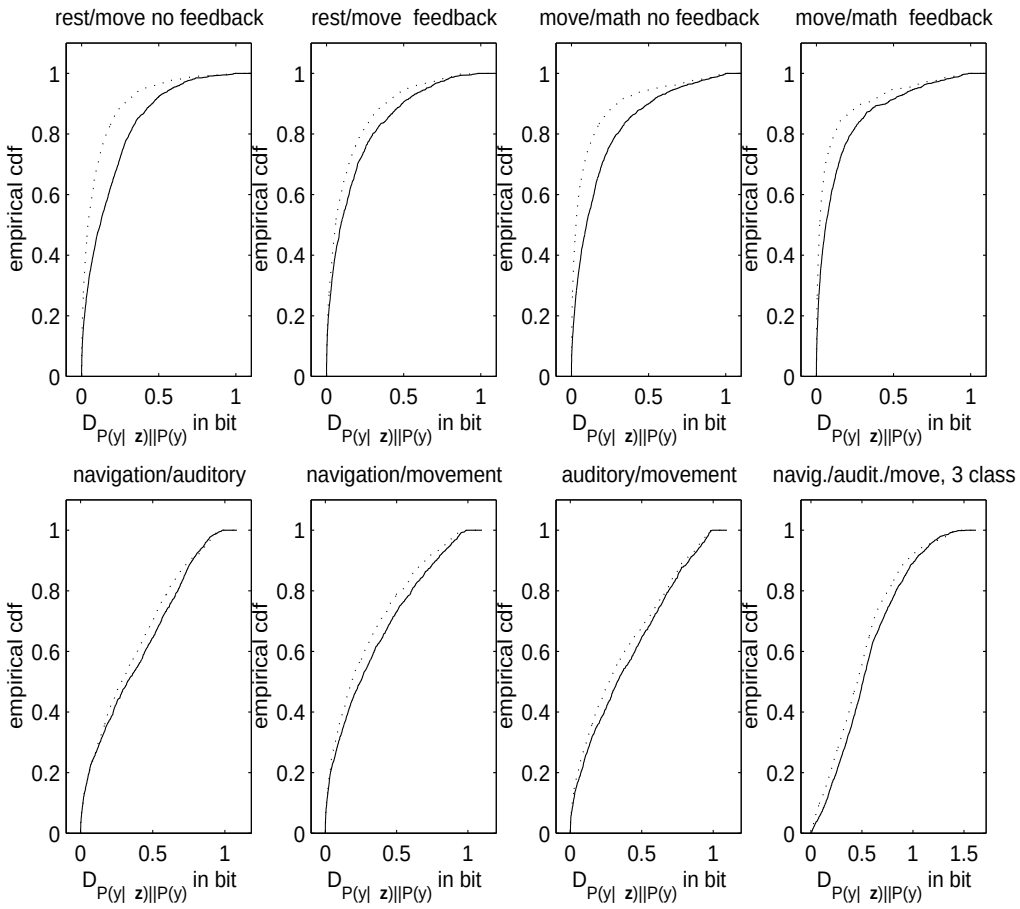
Generalization accuracy study two

Characteristics: 10 subjects, three sets of trials assessing combinations of a navigation task, an auditory imagination and imagined movements, 2 differential electrode sites C3'-C3" (left motor area for right motor imagination) and T4-P4 (right temperoparietal for spatial and auditory tasks) with reference lateral to left mastoid, 10 repetitions of each task done for 7 seconds, no feedback.

We predict for every second without reject option:

Cognitive task	Generalization results		
	vkf	vsi	P_{null}
navigation/auditory,	0.86	0.85	0.02
navigation/movement	0.80	0.80	0.31
auditory/movement	0.78	0.76	0.00
navig./audit./move, 3 class	0.75	0.73	0.00

Empirical cdfs over "information distance"



Empirical cdf. over Kullback Leibler (KL) divergences between prior probabilities of cog. states and posteriors obtained by BCI classifier. dotted line $->$ static method, solid line $->$ variational Kalman filter. KL divergences of vkf are larger. Kolmogorov-Smirnov test based on these cdfs.

Comparing average bit rates

Study one			
	bit rates $r_{P(y)}$ [bit/s]		
task	vkf	vsi	P_{null}
rest/move no fb.	0.18	0.10	4.210^{-26}
rest/move fb.	0.18	0.13	5.510^{-57}
move/math no fb.	0.18	0.11	5.610^{-8}
move/math fb.	0.15	0.10	1.310^{-47}
Study two			
	bit rates $r_{P(y)}$ [bit/s]		
task	vkf	vsi	P_{null}
nav./aud./move	0.55	0.49	4.010^{-3}
audit./move	0.38	0.35	4.010^{-4}
navig./move	0.32	0.28	9.010^{-4}
navig./audit.	0.37	0.34	6.010^{-7}

Conclusion

- We propose in this work a truly adaptive BCI which we infer using a novel algorithm based on variational Bayes.
- An empirical comparison using *generalization accuracy* and *bit rate* show that the proposed method improves over equivalent static classification. The differences were found to be highly significant.
- We thus suggest that in order to achieve *optimal bit rates* BCI's should be based on concepts of adaptive learning.
- Since all calculations of the proposed algorithm can be done in *real time*, the *variational Kalman filter* is a promising technique towards a fully adaptive BCI.