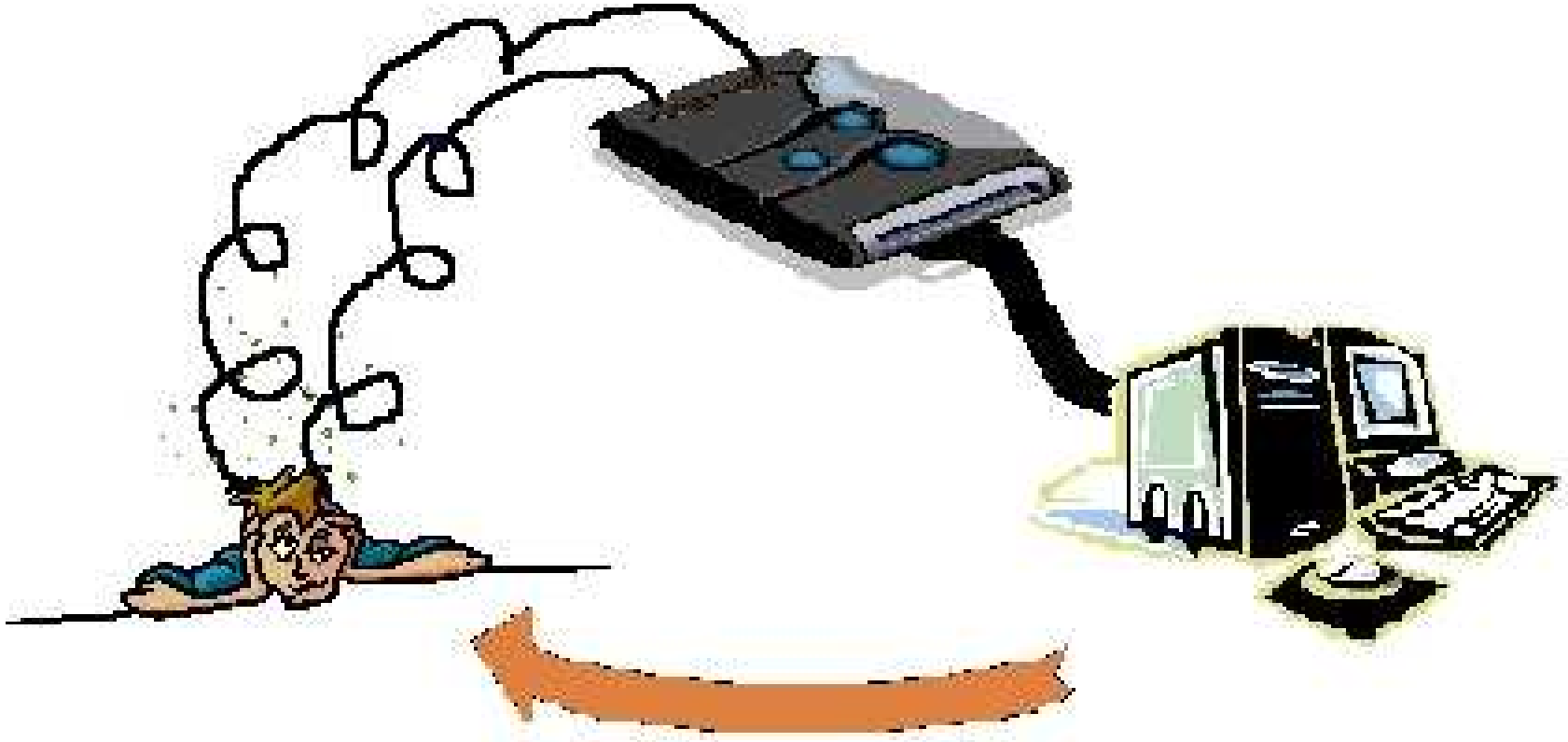
Towards adaptive BCI

Collaborators:
I. Johnsrude and A. M. Owen, MRC Cognition & Brain Sciences Unit, University of Cambridge
E. Curran University of Keele
M. Stokes Research Department, Royal Hospital for Neuro-disability, Putney, London
W. Penny Wellcome Department of Imaging Neurosciene, University College London
M. Gibbs, P. Sykacek and S. Roberts Engineering Science, University of Oxford

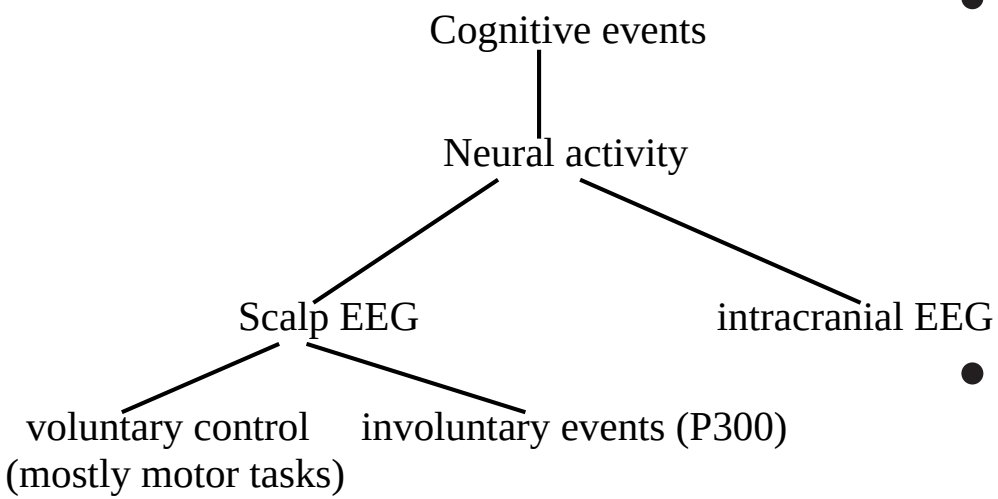
Peter Sykacek is currently supported by grant Nr. F46/399 kindly provided by the BUPA foundation.

A Brain Computer Interface



Computer is controlled *directly* by *cortical activity*.

Classification of BCI's



- intracranial EEG $->$ high spatial and temporal resolution; high interference with patient; allows 2-d control of artificial limb.
- surface EEG $->$ low spatial and temporal resolution; no permanent interference with patient; slow! at most 20 bit per minute and task.

$->$ focus on BCI's based on scalp recordings.
$->$ low bit rates; last resort if no other communication possible

BCI with almost no adaptation

- P300 based: L. A. Farwell and E. Donchin, $->$ User intention is embedded within a sequence of symbols. The correct symbol leads to "surprise" and triggers a P300.

- filter & threshold: N. Birbaumer etal. , $->$ threshold slow cortical potentials; J.R. Wolpaw etal., $->$ threshold moving average in an appropriate pass band e.g. μ-rhythm.

These principles rely mostly on user training.

BCI with "static" pattern recognition

- Extract representation of EEG "waveforms" (e.g. low pass filtered time series; spectral representation)
- Parameterize supervised classification implicitly assuming stationarity.

What if

Technical setup changes during operation?
(e.g. electrolyte changes impedance)
User learns from feedback?
User shows fatigue?

Assuming stationarity must be wrong !

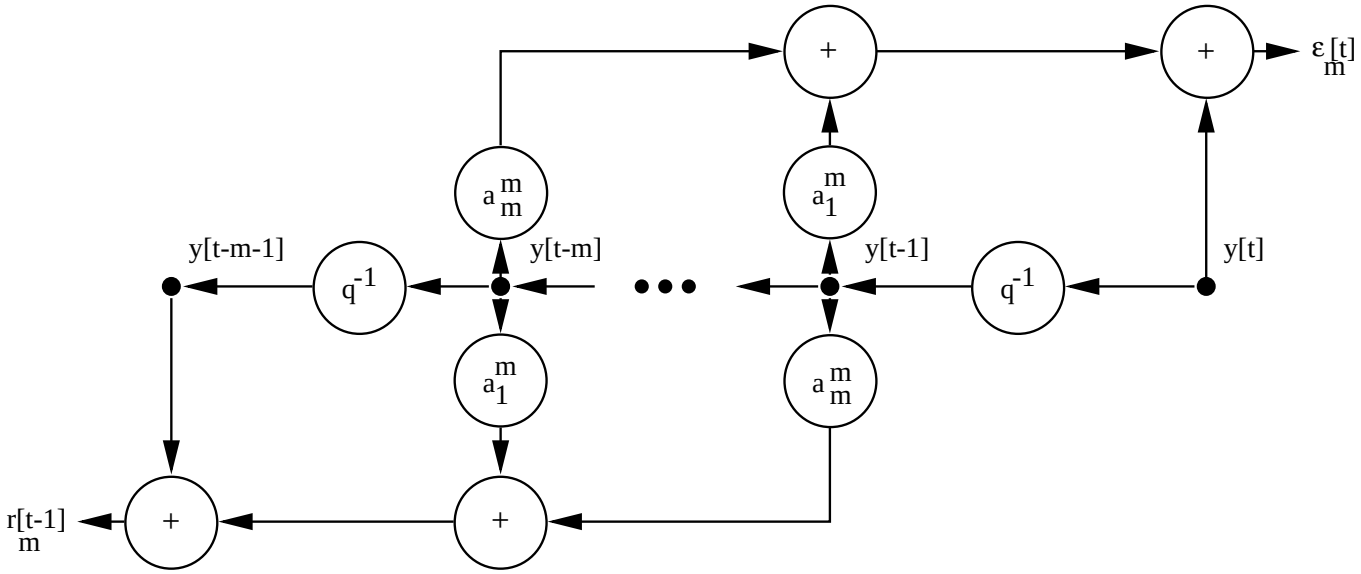
$->$ Propose "adaptive" BCI.

The architecture

- Adaptive BCI refers to a brain computer interface (BCI) which is built around adaptively inferred classification.
- The proposed method is a 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.
- Feature extraction by suitably transformed AR-coefficients.
- Adaptive classification by variational Kalman filter.

For practicality we need to keep computational cost down.

Feature extraction by autoregressive (AR) processes



... represented by reflection coefficients.

... and corresponding equations

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

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

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

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

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

and use $\boldsymbol{x}_n = [\text{arctanh}(\widehat{\rho}_{1,n}), ..., \text{arctanh}(\widehat{\rho}_{p,n})]^T$ to represent $\mathcal{Y}_n$.

A generalized nonlinear classifier (or RBF network)

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

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

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

$\boldsymbol{\phi}_n$: 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

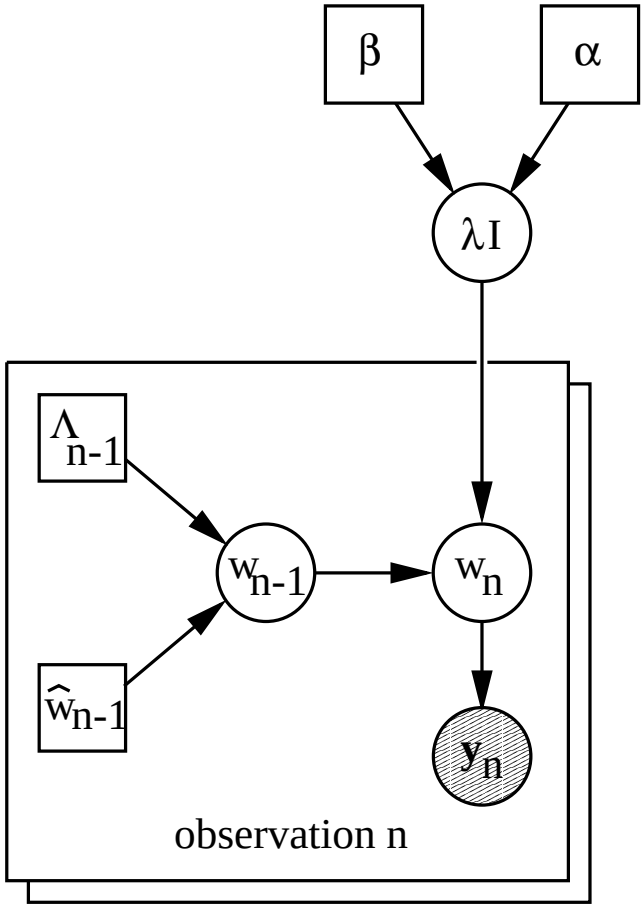
Probabilistic view of adaptive inference $->$ state space formulation of a first order Markov process.

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

$\boldsymbol{w}_{n-1}$, $\boldsymbol{w}_n$: 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 marginal likelihood for one observation

$$\log(p(\mathcal{D}_n)) = \log\Big(\int_\lambda \int_{\boldsymbol{w}_{n-1}} \int_{\boldsymbol{w}_n} p(\boldsymbol{w}_{n-1}|\mathcal{D}_{n-1}) p(\boldsymbol{w}_n|\boldsymbol{w}_{n-1}, \lambda \boldsymbol{I}) P(y_n|\boldsymbol{w}_n, \boldsymbol{\phi}_n) p(\lambda|\alpha,\beta) d\boldsymbol{w}_n d\boldsymbol{w}_{n-1} d\lambda\Big)$$

Model for a single observation y_n.
Inference: integrate w.r.t. $\boldsymbol{w}_{n-1}$ and approximate $p(\boldsymbol{w}_n, \lambda)$ by a product $Q(\lambda)Q(\boldsymbol{w}_n)$. NOT simple mean field since we approximate a partially marginalized distribution.

Problem: inference of λ based on a single observation.
$->$ Follow the previously proposed "windowed" Kalman filter (Jazwinsky 69).

Windowed Kalman filter

Independence of successive pairs of state vectors $(\boldsymbol{w}_{n-1}, \boldsymbol{w}_n)$
and ignore anti-causal informnation flow.
The logarithmic model evidence for a window of size N is then

$$\log(p(\mathcal{D}_N)) = \log\Bigg(\int_\lambda \prod_{n=1}^{N} \Bigg[\int_{\boldsymbol{w}_{n-1}} \int_{\boldsymbol{w}_n} p(\boldsymbol{w}_{n-1}|\mathcal{D}_{n-1}) \tag{7}$$
$$p(\boldsymbol{w}_n|\boldsymbol{w}_{n-1}, \lambda \boldsymbol{I}) P(y_n|\boldsymbol{w}_n, \boldsymbol{\phi}_n) d\boldsymbol{w}_n d\boldsymbol{w}_{n-1}\Bigg] p(\lambda|\alpha, \beta) d\lambda\Bigg).$$

Note: there is no more a corresponding probabilistic structure!
(The correct model is the "Rauch Tung Striebel" smoother.)

Filling in distributions:

$$
\begin{aligned}
\log(p(\mathcal{D}_N)) \;=\; & \log\Bigg(\int_\lambda \prod_{n=1}^{N}\Bigg[\int_{\boldsymbol{w}_n} (2\pi)^{-\frac{d}{2}}|\boldsymbol{\Lambda}_{n-1}^{-1}+\lambda^{-1}\boldsymbol{I}|^{-\frac{1}{2}} \qquad (8)\\
& \times \exp(-0.5(\boldsymbol{w}_n-\hat{\boldsymbol{w}}_{n-1})^T(\boldsymbol{\Lambda}_{n-1}^{-1}+\lambda^{-1}\boldsymbol{I})^{-1}(\boldsymbol{w}_n-\hat{\boldsymbol{w}}_{n-1}))\\
& \times (1+\exp((2y_n-1)\boldsymbol{\phi}_n^T\boldsymbol{w}_n))^{-1}d\boldsymbol{w}_n\Bigg]\\
& \times \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{(\alpha-1)}\exp(-\beta\lambda)d\lambda\Bigg)
\end{aligned}
$$

Remember: approximate $p(\boldsymbol{w}_n,\lambda)$ as $Q(\lambda)Q(\boldsymbol{w}_n)$ where $Q(\lambda)$: Gamma distribution and $Q(\boldsymbol{w}_n)$: Gaussian distributions.
No conjugacy! Bound below using Jaakkola & Jordans bound for the logistic cdf and our bound for "Gamma conjugacy" (introduce variational parameters ξ_n and ν).

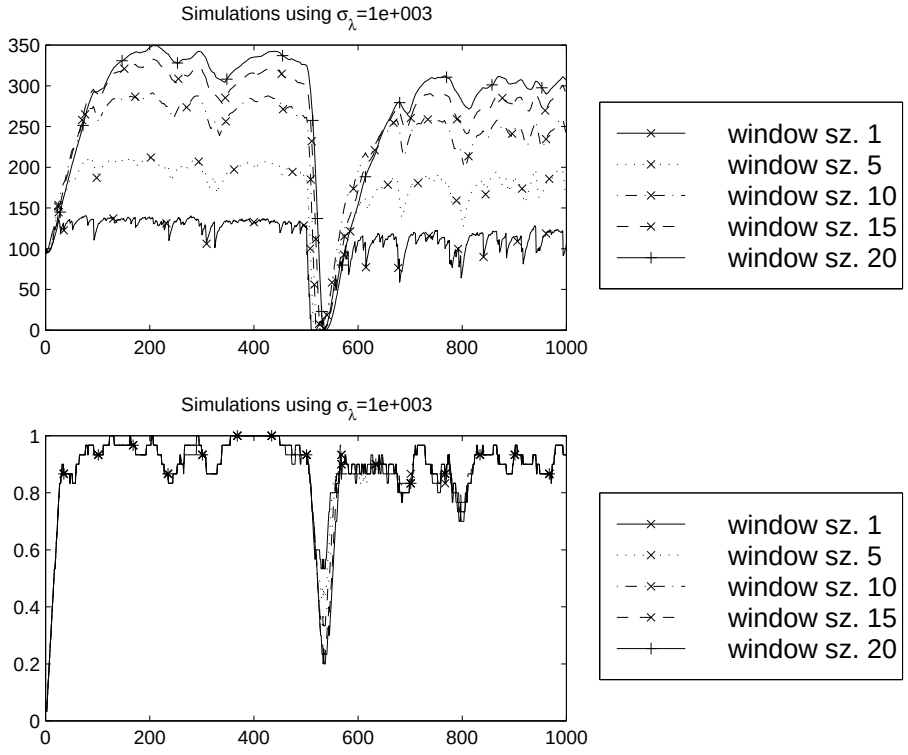
Inference steps of the variational Kalman filter

update(ξ_N)
update($\boldsymbol{w}_N$)
update(λ)
REPEAT
    $\forall\ n$ update($\boldsymbol{w}_n$)
    $\forall\ n$ update(ξ_N)
    update(λ)
    update(ν)
    $\boldsymbol{\delta} \sim N(\mathbf{0}, \boldsymbol{\Delta})$
    $\boldsymbol{w}'_\varphi = \boldsymbol{w}_\varphi + \boldsymbol{\delta}$ % propose new basis coefficients
    with probability P_{acc}, $\boldsymbol{w}_\varphi = \boldsymbol{w}'_\varphi$
UNTIL convergence(F)

Predicting probabilities with the variational Kalman filter.

$\boldsymbol{\phi}_N =$ basisfunction($\boldsymbol{x}_N$)
REPEAT
    update(ξ_N)
    update($\boldsymbol{w}_N$)
UNTIL convergence(F)
calculate($\tilde{P}$)

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 inferred with sequential variational inference (non adaptive method).

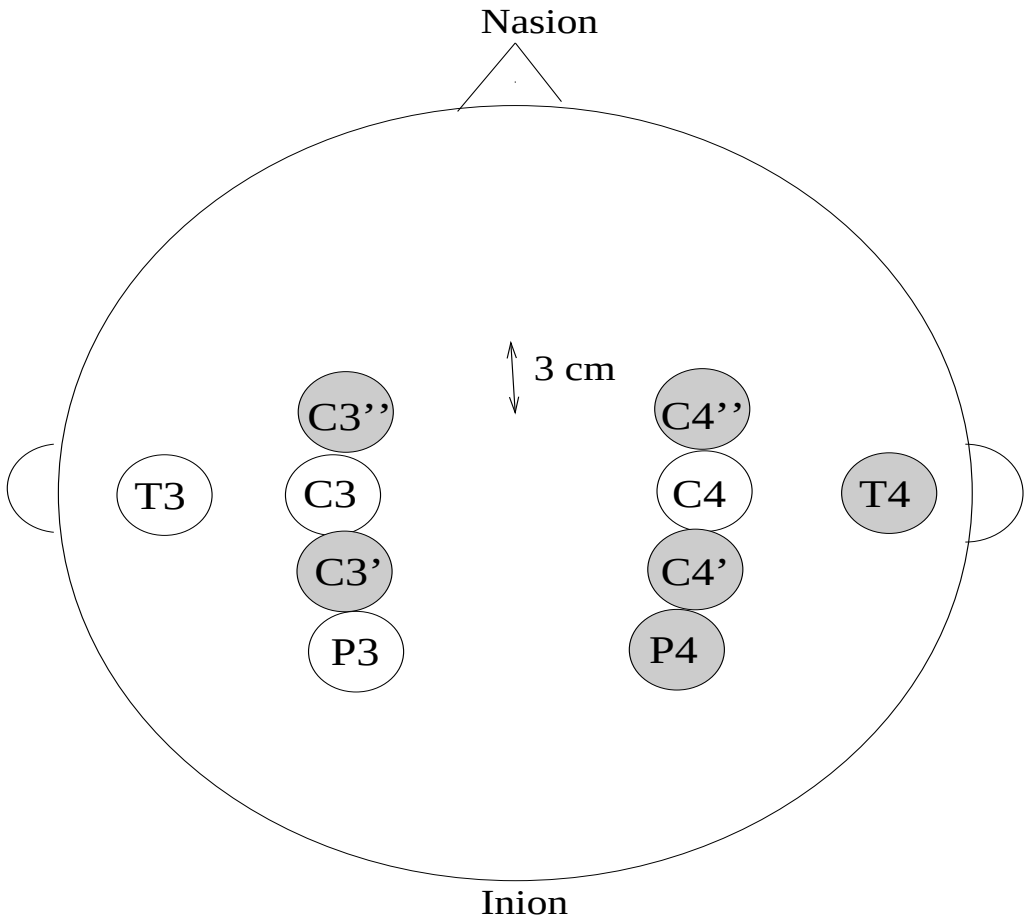
- We measure generalization accuracy on independent test data and check for statistical significance using Mc. Nemar's test (a test for paired experiments).

- 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 labelled 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 halves 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.

Electrode positions



Augmented 10-20 positions at T4, P4 (right tempero-parietal for spatial and auditory tasks), C3', C3'' (left motor area for right motor imagery) and C4', C4'' (right motor area for left motor imagery)

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	$\ll 0.01$
rest/move, feedback	0.71	0.70	0.39
move/math, no feedback	0.69	0.62	$\ll 0.01$
move/math, feedback	0.64	0.60	$\ll 0.01$

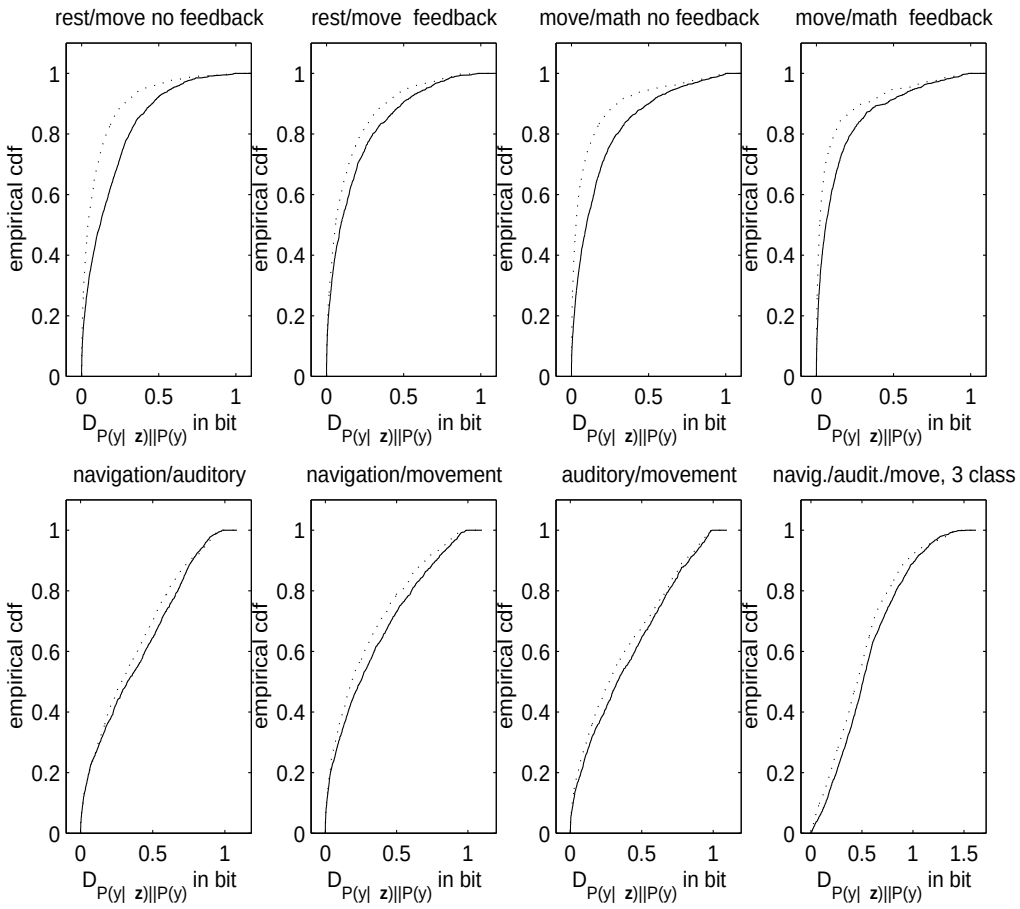
Generalization accuracy study two

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 temporo-parietal 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	$\ll 0.01$
navig./audit./move, 3 class	0.75	0.73	$\ll 0.01$

Empirical cdfs over KL divergence



Empirical cdf. over Kullback Leibler (KL) divergence between prior probabilities of cog. states and posteriors obtained by the 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	$\ll 0.01$
rest/move fb.	0.18	0.13	$\ll 0.01$
move/math no fb.	0.18	0.11	$\ll 0.01$
move/math fb.	0.15	0.10	$\ll 0.01$
Study two			
	bit rates $r_{P(y)}$ [bit/s]		
task	vkf	vsi	P_{null}
nav./aud./move	0.55	0.49	$\ll 0.01$
audit./move	0.38	0.35	$\ll 0.01$
navig./move	0.32	0.28	$\ll 0.01$
navig./audit.	0.37	0.34	$\ll 0.01$

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.

Recent BCI references

E. Curran and M. Stokes, "Learning to control brain activity: A review of the production and control of EEG components for driving brain-computer interface (BCI) systems". *Brain and Cognition*, pp. 326–335, 2003.

E. Curran, P. Sykacek, M. Stokes, S. Roberts, W. Penny, I. Johnsrude and A. Owen. "Cognitive tasks for driving a brain computer interfacing system: a pilot study". *IEEE Trans. Neur. Syst. and Rehab. Eng.*, to appear, 2004.

P. Sykacek and S. Roberts. "Adaptive classification by variational Kalman filtering", In *Advances in Neural Processing Systems 15*, MIT Press, 2003.

P. Sykacek, S. Roberts, M. Stokes, E. Curran, M. Gibbs and L. Pickup. "Probabilistic methods in BCI research". *IEEE Trans. Neur. Syst. and Rehab. Eng.*, 2003.

P. Sykacek, S. Roberts and M. Stokes. "Adaptive BCI based on variational Bayesian Kalman filtering: an empirical evaluation", *IEEE Trans. Biomed. Eng.*, to appear, 2004.

Please see http://www.robots.ox.ac.uk/parg/ for our online versions!