Probabilistic Methods in BCI Research

Peter Sykacek

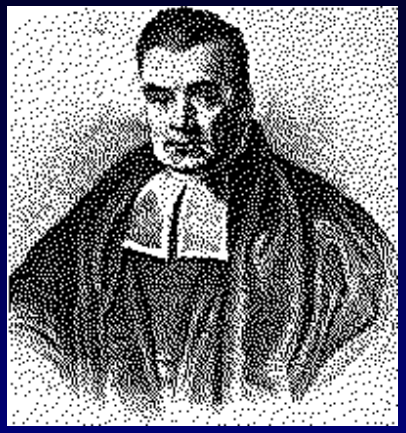
Depts. of Genetics & Pathology,
University of Cambridge

http://www.sykacek.net

Talk Overview

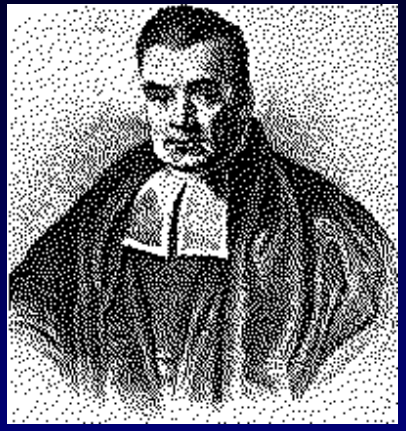
- Motivation of Probabilistic Concepts
- BCI, current practice & shortcomings
- Adaptive BCI
- A Note on AR modelling
- GLM based Classification
- Probabilistic Kalman Filter
- Cognitive Issues
- Computer Simulations
- Conclusion

Probabilistic Motivations



Thomas Bayes (1701 - 1763)
Learning from data using a decision theoretic framework

Probabilistic Motivations

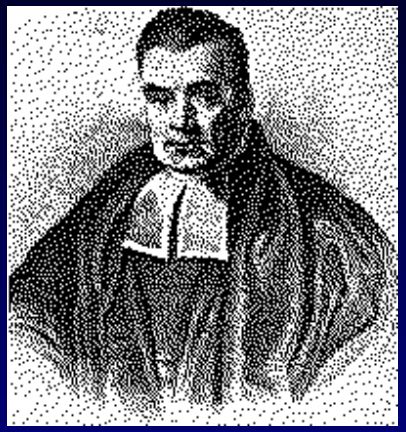


Thomas Bayes (1701 - 1763)
Learning from data using a decision theoretic framework

$$p(x|\mathcal{D}) = \frac{p(\mathcal{D}|x)p(x)}{p(\mathcal{D})}$$

First consequence: we must revise beliefs according to Bayes theorem

Probabilistic Motivations



Thomas Bayes (1701 - 1763)
Learning from data using a decision theoretic framework

$$p(x|\mathcal{D}) = \frac{p(\mathcal{D}|x)p(x)}{p(\mathcal{D})}$$

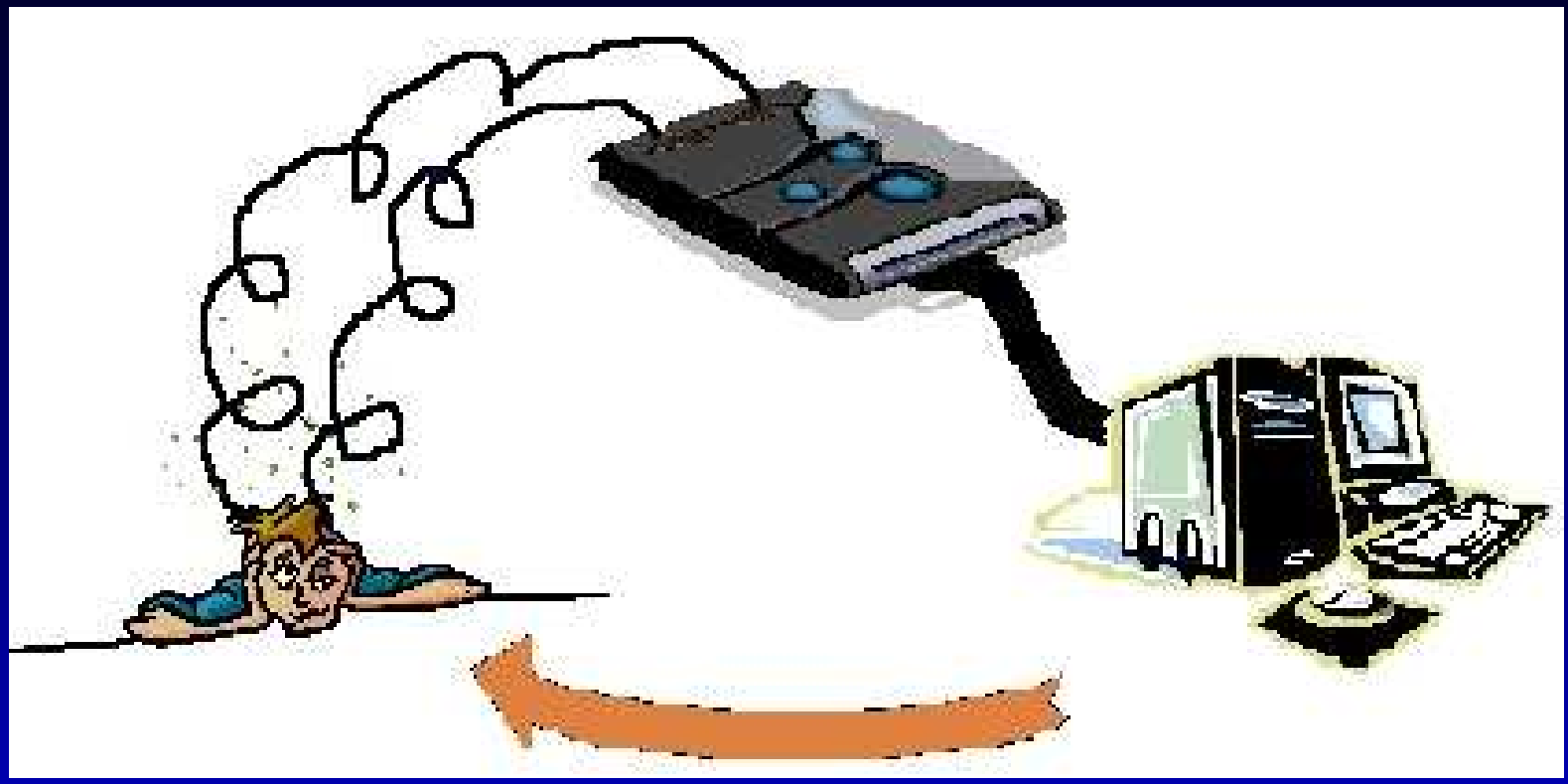
First consequence: we must revise beliefs according to Bayes theorem

$\alpha_{opt} = \text{argmax}_\alpha < u(\alpha) >$, where

$< u(\alpha) >= \int_x u(\alpha, x)p(x|\mathcal{D})dx.$

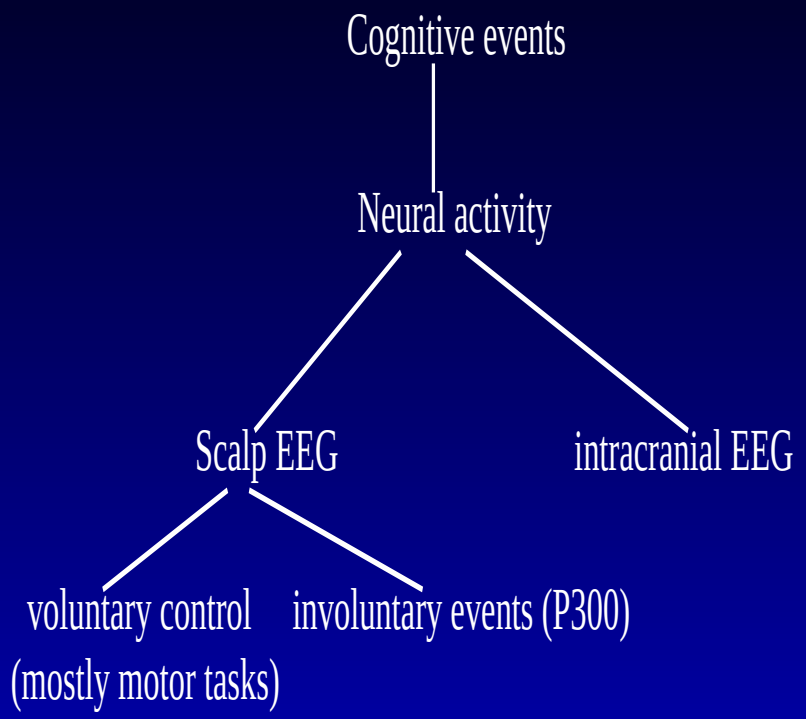
Second consequence: Decisions by maximising expected utilities

Brain Computer Interface



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

Classification of BCIs



intracranial EEG $->$ high spatial and temporal resolution; highly invasive!; 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 & 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 !

$->$ Probabilistic method for “adaptive” BCI.

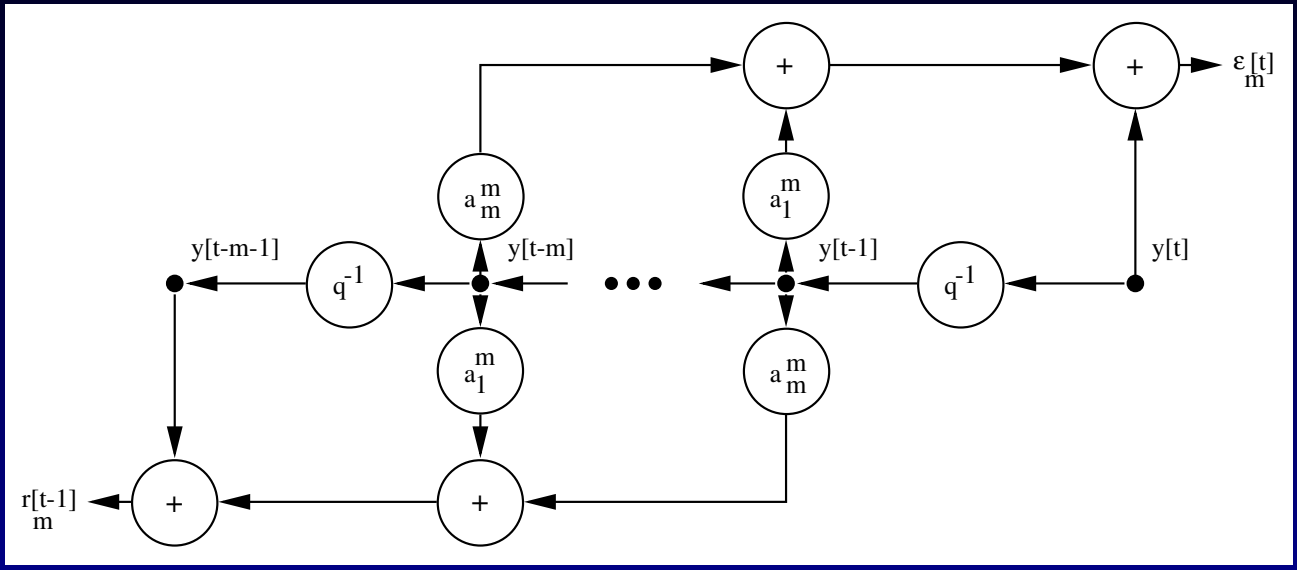
Our Approach

- Adaptive BCI refers to a brain computer interface (BCI) which is built around adaptively inferred classification.
- Two stage approach: extract features from EEG and predict probabilities of cognitive states.
- Feature extraction by AR-models
- Adaptive classification by variational Kalman filter

Probabilities $->$ optimal bit rate

Real time requirement $->$ easy to compute

Autoregressive (AR) modelling



Parameterise by reflection coefficients

$$\hat{\rho}_m = -\frac{\boldsymbol{r}_m^{\mathsf{T}} \boldsymbol{\epsilon}_m}{\boldsymbol{r}_m^{\mathsf{T}} \boldsymbol{r}_m}$$

Fishers z-transform by stability argument $->$

$$\boldsymbol{x}_n = [\operatorname{arctanh}(\hat{\rho}_{1,n}), \ldots, \operatorname{arctanh}(\hat{\rho}_{p,n})]^T$$

GLM

Reflection Coefficients

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

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 $\mathcal{Y}_n$:

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

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

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

A GLM classifier

$$\begin{aligned}
\boldsymbol{\phi}_n &= \begin{bmatrix} 1 \\ \boldsymbol{x}_n \\ \boldsymbol{\varphi}(\boldsymbol{x}_n; \boldsymbol{w}_{\boldsymbol{\varphi}}) \end{bmatrix} \\
\eta_n &= \boldsymbol{\phi}_n^T \boldsymbol{w} \\
P(y_n | \boldsymbol{w}, \boldsymbol{w}_\varphi, \boldsymbol{x}_n) &= \frac{1}{1 + \exp((2y_n - 1)\eta_n)}, \quad \text{with}
\end{aligned}$$

$\boldsymbol{\phi}_n$: 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):

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

Kalman filter tracking

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

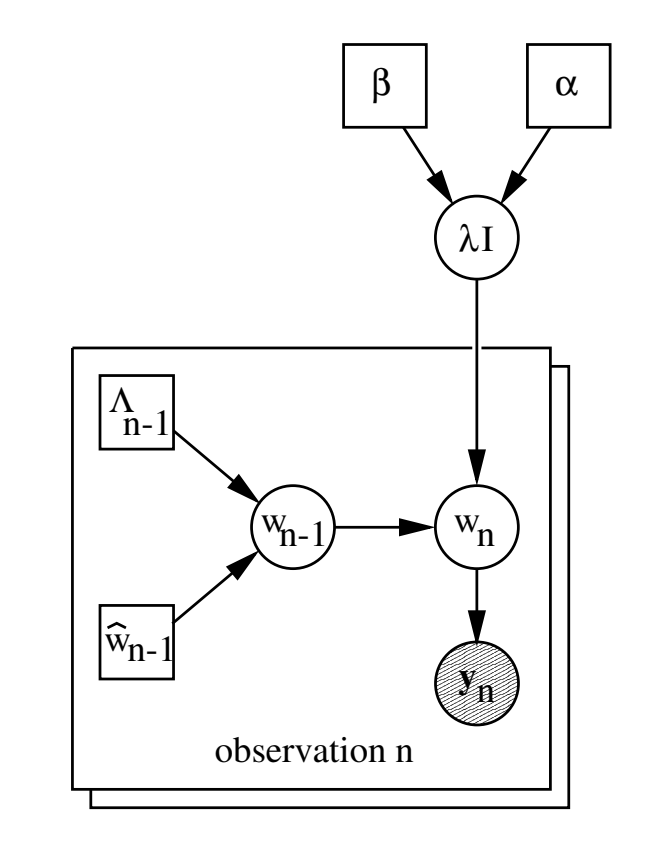
$p(\boldsymbol{w}_{n-1})$
$p(\boldsymbol{w}_n|\boldsymbol{w}_{n-1}, \lambda\boldsymbol{I})$ for times $n \geq 1$
$p(y_n|\boldsymbol{x}_n, \boldsymbol{w}_n)$ for times $n \geq 1$, where

$p(\boldsymbol{w}_{n-1})$, Gaussian "prior" at time $n-1$.

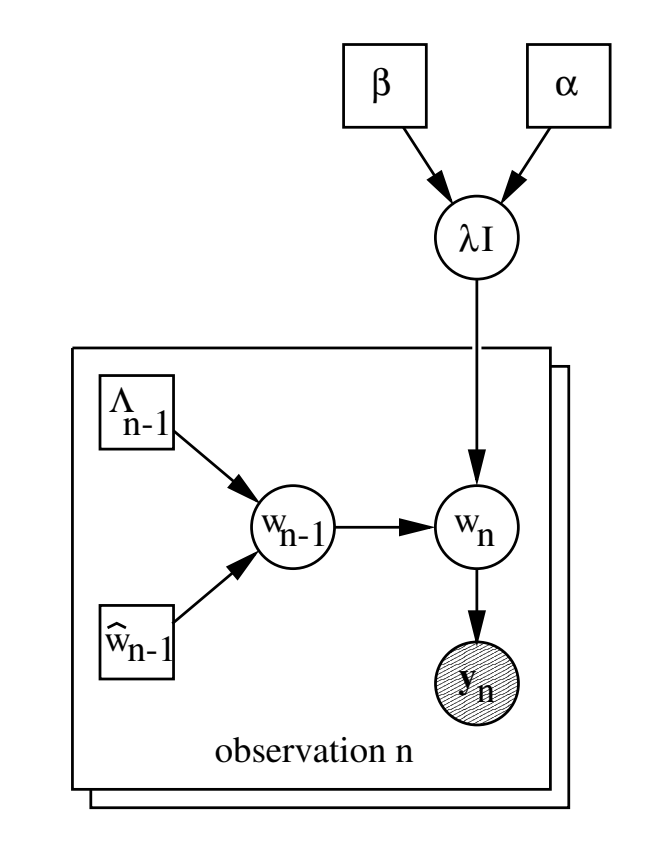
$p(\boldsymbol{w}_n|\boldsymbol{w}_{n-1}, \lambda\boldsymbol{I})$, Gaussian "state noise" with mean $\boldsymbol{w}_{n-1}$ and precision $\lambda\boldsymbol{I}$.

$p(y_n|\boldsymbol{x}_n, \boldsymbol{w}_n)$, observation noise model.
if linear and Gaussian $->$ Kalman filter.
Here logit link and nonlinear non Gaussian.

Probabilistic Kalman Filter

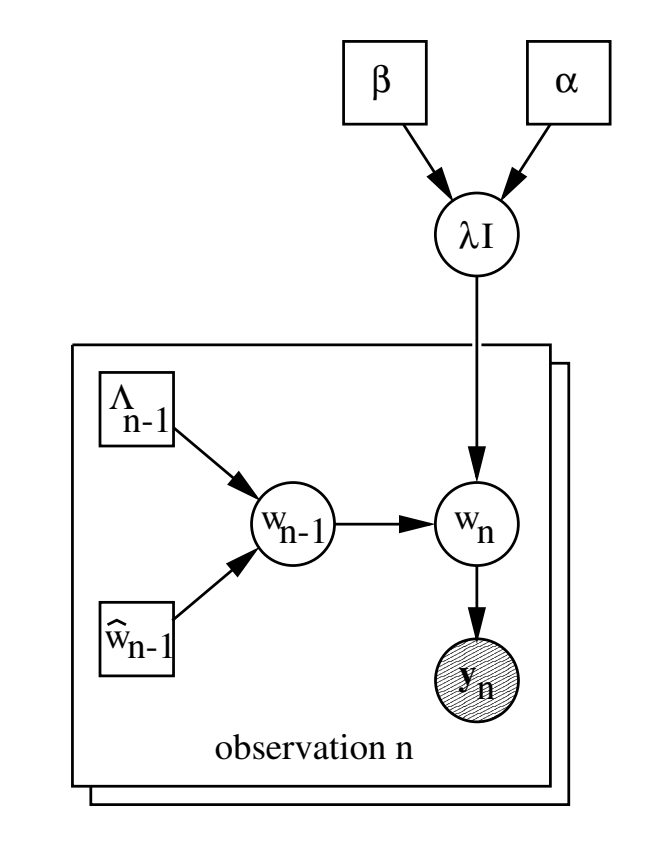


Probabilistic Kalman Filter



Key: get λ right (may regard $1/\lambda$ as learning rate)

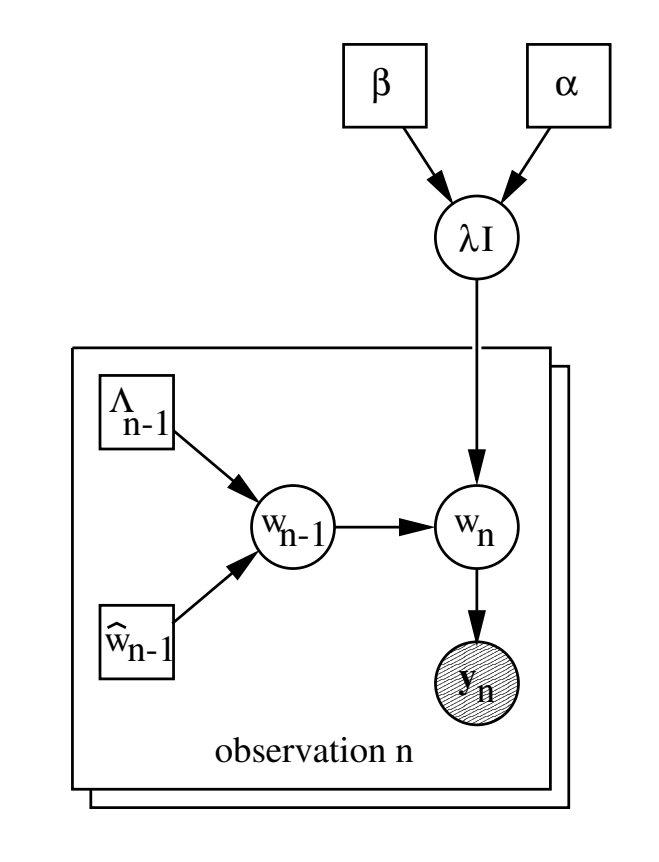
Probabilistic Kalman Filter

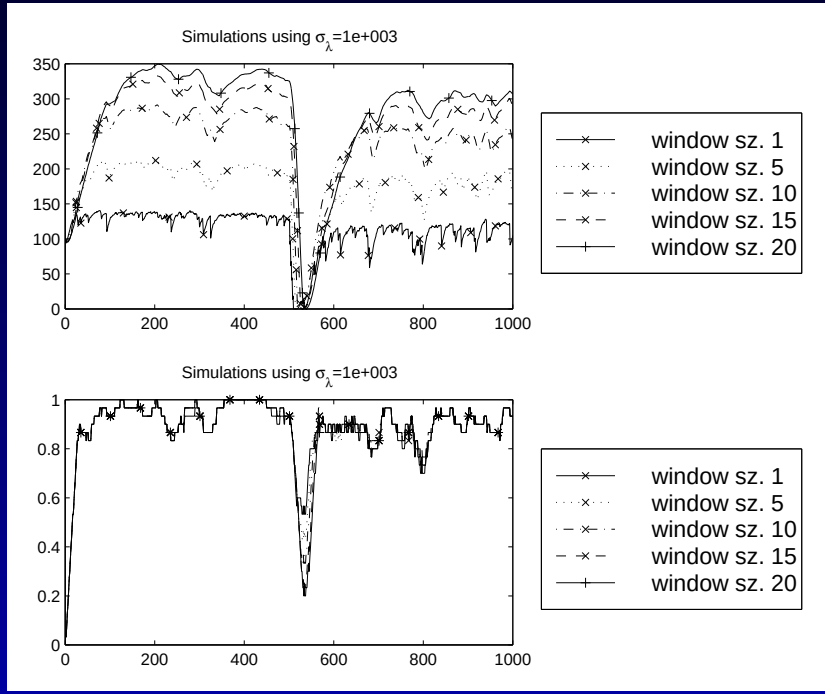


Key: get λ right (may regard $1/\lambda$ as learning rate)

Classification: Non linear and non Gaussian, some eqns.

Probabilistic Kalman Filter



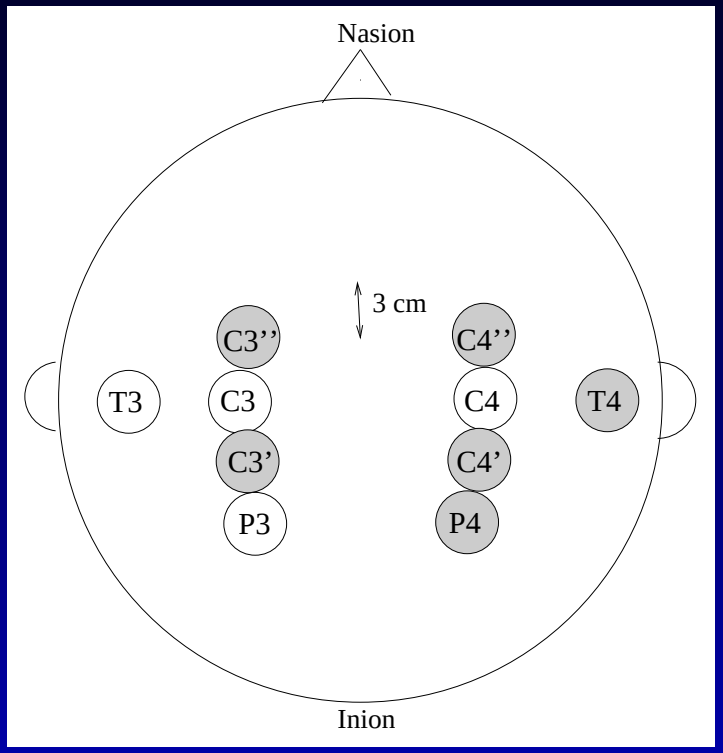


Key: get λ right (may regard $1/\lambda$ as learning rate)

Classification: Non linear and non Gaussian, some eqns.

Illustration of $< \lambda >$ and "instantaneous" generalization error for B. D. Ripley's synthetic data with artificial non-stationarity (swap labels after sample 500).

Cognitive Issues



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)

Original task setting:

rest EEG, math task and imagined movement; 8 young and healthy subjects; 12 repetitions of each task for about 8 seconds

Modified task setting:

spatial imagination, auditory imagination, left and right imagined movement; 10 young and healthy subjects; 10 repetitions for about 10 seconds

Data Recording:

EEG band pass filtered (0.1 Hz - 100 Hz) and sampled at 384 Hz, 12 Bit resolution.

Computer Simulations

- Extract 3 reflection coefficients per second EEG and channel. Predict probability of state and update parameters.
- Comparison with equivalent static classifier. Half of entire experiment $-$ $>$ training data.
- We measure generalization accuracy on independent test data and check for statistical significance using Mc. Nemar's test (a test for paired experiments).
- Estimate BCI's bit rate (channel capacity).

Generalisation Accuracy

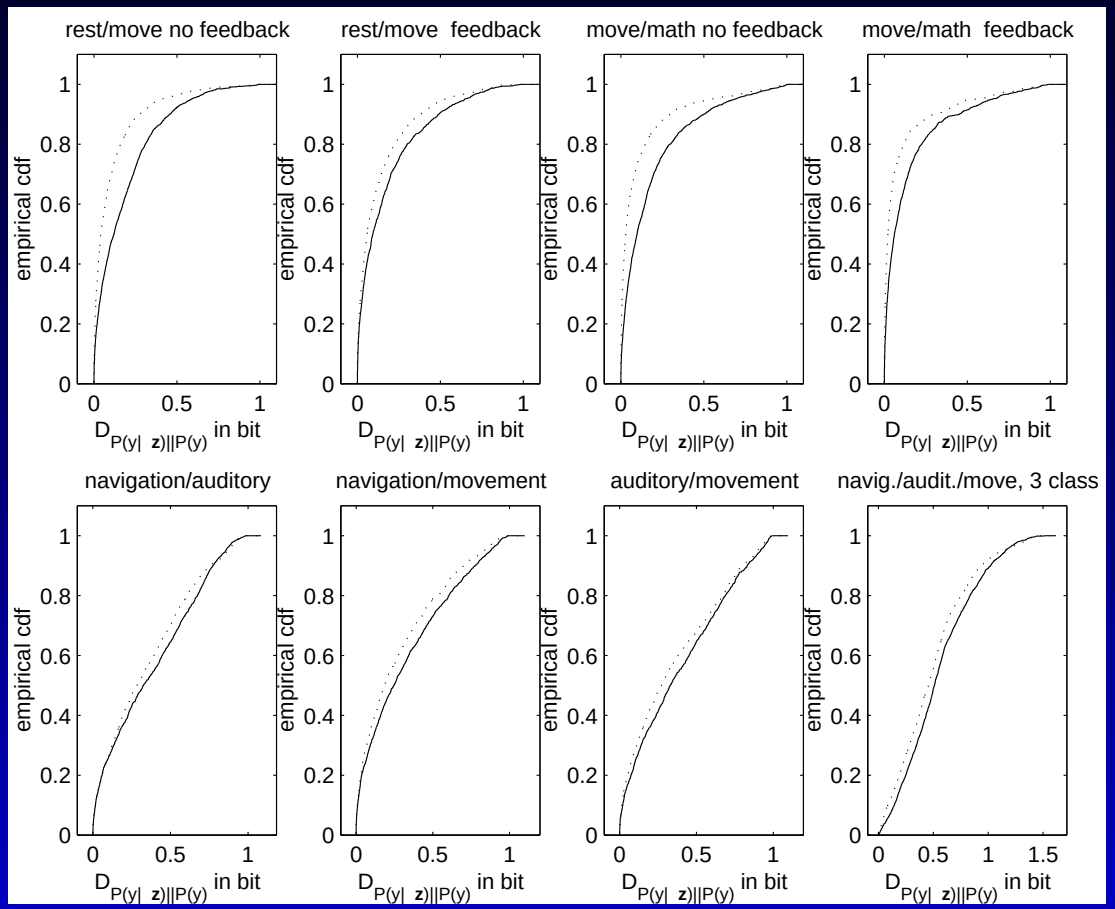
Experiment 1:

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	$\ll 0.01$

Experiment 2:

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	0.75	0.73	$\ll 0.01$

CDF over KL divergence



Measures BCI's channel capacity

Empirical cdf. over KL divergence between prior probabilities of cog. states and posteriors.

dotted line –> static method, solid line –> variational Kalman filter.

KL divergences of vkf are larger.

Communication Bandwidth

	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	$\ll 0.01$
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 a truly adaptive BCI which we infer using a novel algorithm based on variational Bayes.
- An empirical comparison using generalisation accuracy and bit rate show that the proposed method improves over static classification.
- We thus suggest that in order to achieve optimal bit rates BCI's should be based on concepts of adaptive learning.
- Since all calculations can be done in real time, the variational Kalman filter is a promising technique for a fully adaptive BCI.

Acknowledgements

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 and S. Roberts Engineering Science, University of Oxford

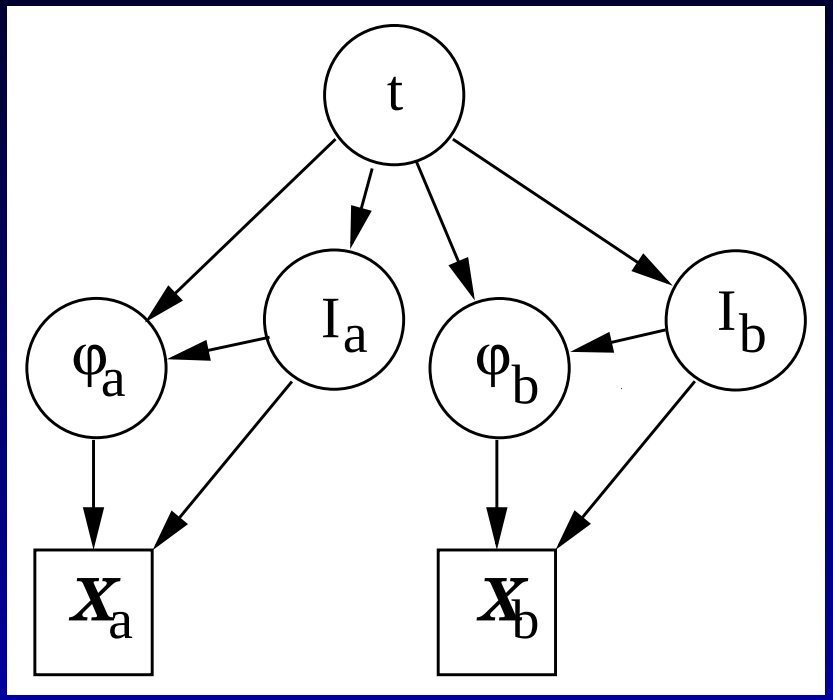
Most of this work was supported by the BUPA grant Nr. F46/399.

A simple idea:

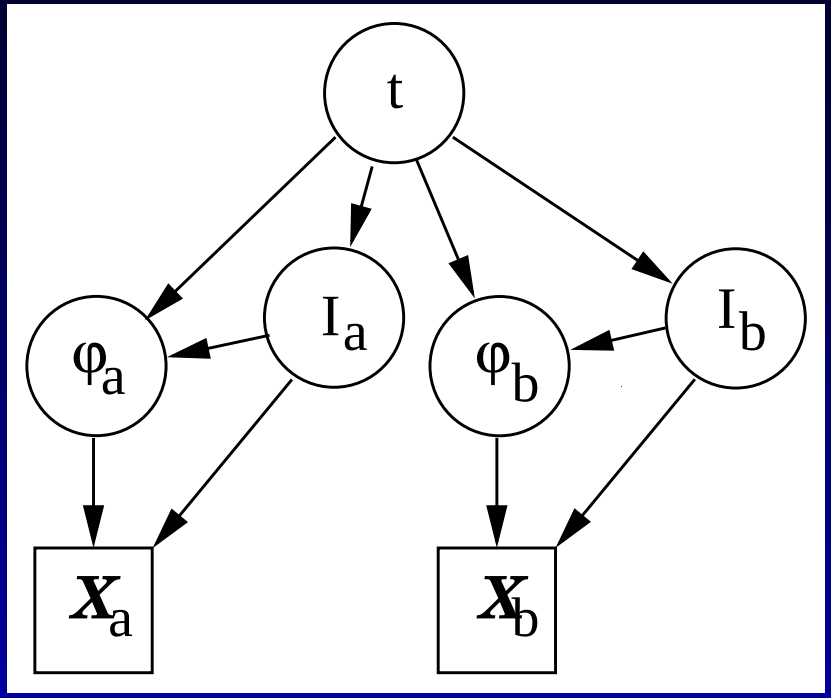
the world is one probabilistic model.

- Applications often require hierarchical structure: a feature extraction part and a probabilistic model.
- Classical approach: treat both parts separately and thus regard features as sufficient statistic of the data. $->$ Features are deterministic variables.
- Our suggestion: treat such hierarchical settings as one probabilistic model. $->$ Feature extraction is a representation in a latent space.

Bayes' Consistent Models

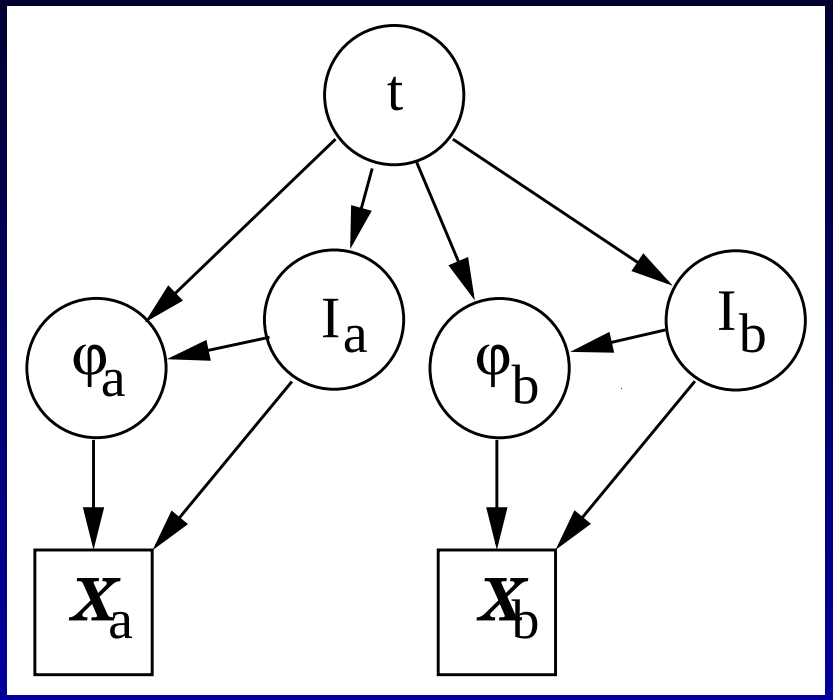


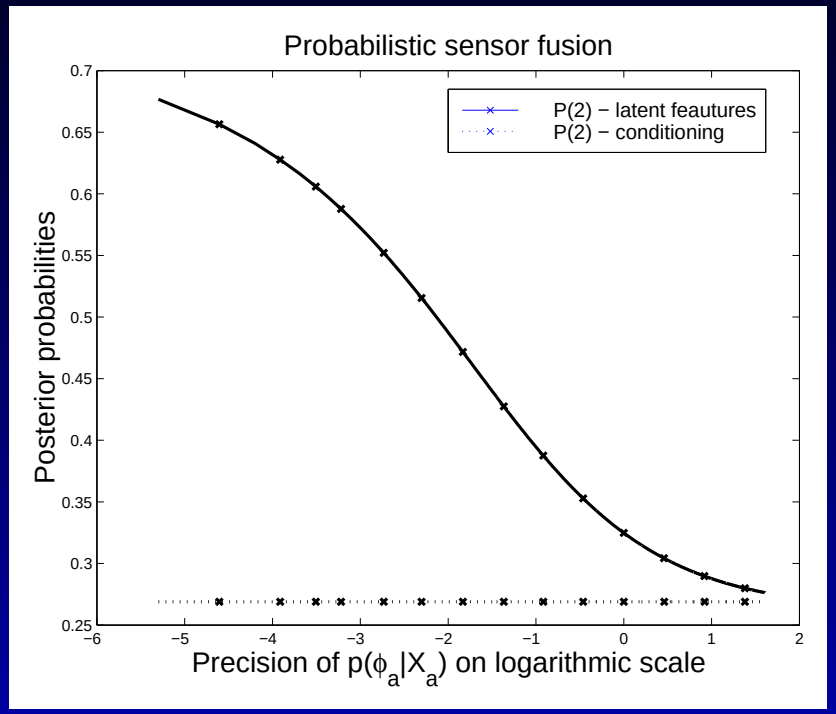
Bayes' Consistent Models



Expected utility requires to integrate over all unknown variables, including φ_a, φ_b, I_a and I_b that represent a feature space.

Bayes' Consistent Models



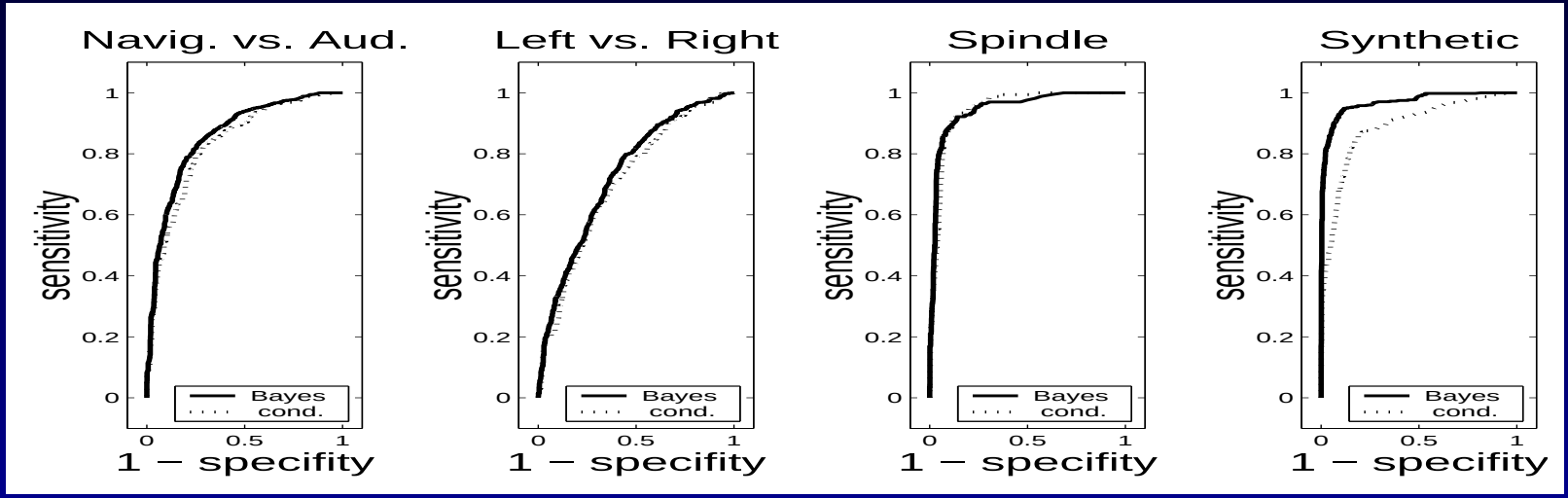


Expected utility requires to integrate over all unknown variables, including φ_a, φ_b, I_a and I_b that represent a feature space.

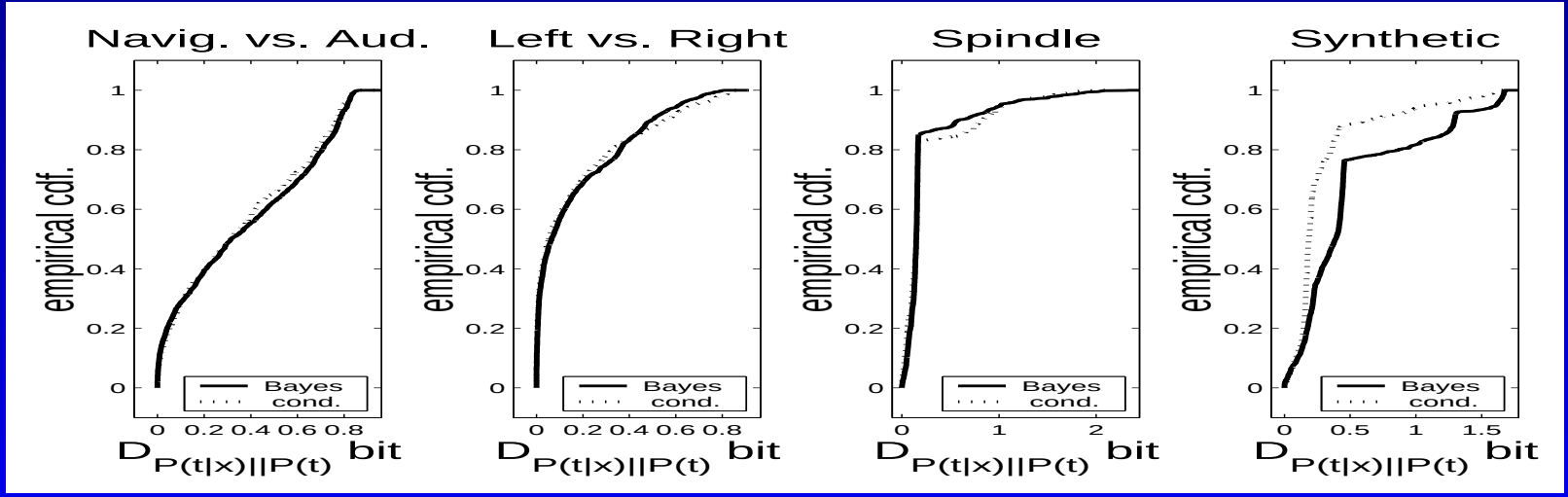
Decisions depend on (un)certainty and may thus change.

Time Series Classification

ROC Curves

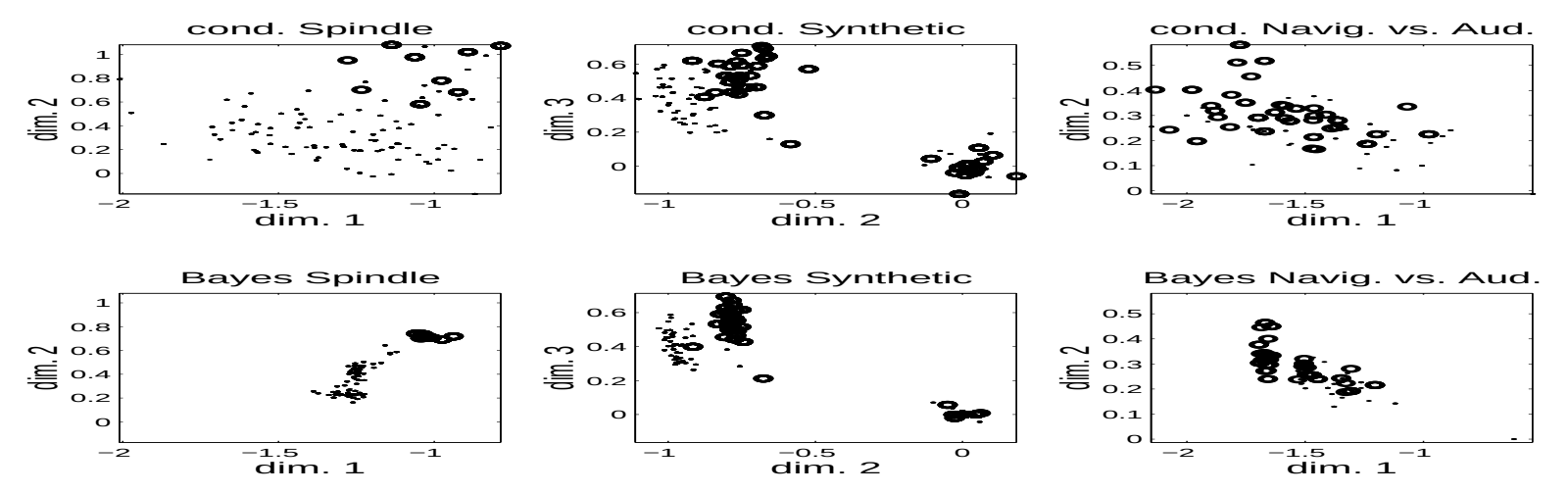


Kullback Leibler Divergence

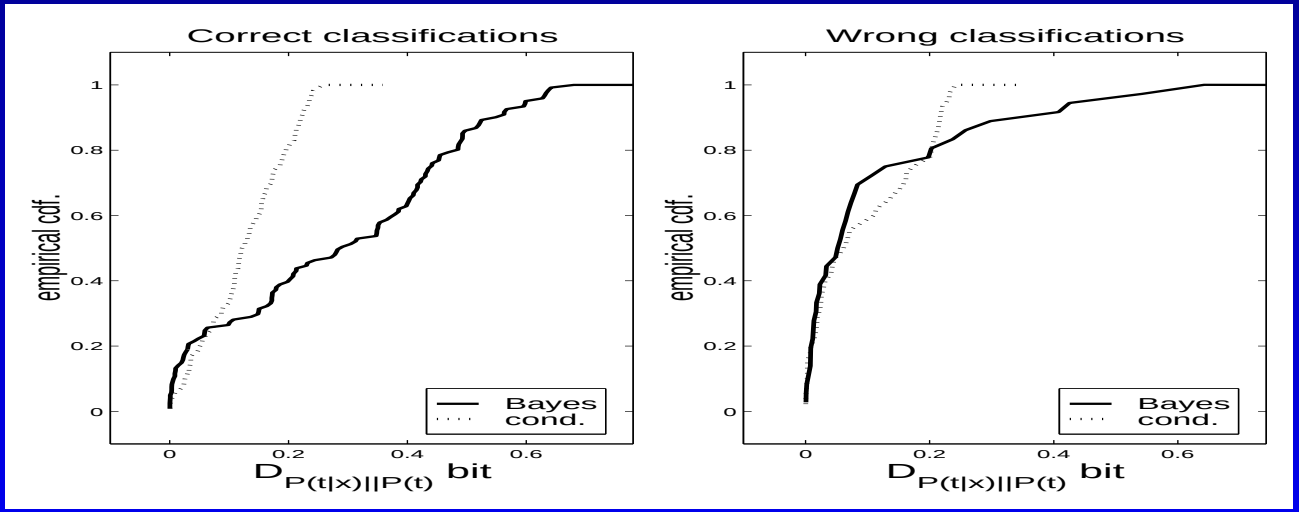


More Results

Expected feature values



Kullback Leibler Divergence for “Artefacts”



Variational Kalman Filter

The logarithmic model evidence for a window of size N is

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

Variational Kalman Filter

The logarithmic model evidence for a window of size N is

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

Plug in distributions and integrate over $\boldsymbol{w}_{n-1}$

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

Windowed KF no probabilistic model! (Rauch Tung Striebel smoother!)

Lower Bounds

$$\begin{aligned}\log(P(y_n|\boldsymbol{\phi}_n,\boldsymbol{w}_n)) &\geq -\frac{(2y_n-1)\boldsymbol{\phi}_n^T\boldsymbol{w}_n}{2} - \log(2) - \log(\cosh(\frac{\xi_n}{2})) \\ &- \frac{\tanh(\frac{\xi_n}{2})}{4\xi_n}\left(\left(\frac{\boldsymbol{\phi}_n^T\boldsymbol{w}_n}{2}\right)^2 - \xi_n^2\right)\end{aligned}$$

back to vkf

Lower Bounds

$$\begin{aligned}
\log(P(y_n|\boldsymbol{\phi}_n, \boldsymbol{w}_n)) &\geq -\frac{(2y_n - 1)\boldsymbol{\phi}_n^T \boldsymbol{w}_n}{2} - \log(2) - \log(\cosh(\frac{\xi_n}{2})) \\
&- \frac{\tanh(\frac{\xi_n}{2})}{4\xi_n}\left(\left(\frac{\boldsymbol{\phi}_n^T \boldsymbol{w}_n}{2}\right)^2 - \xi_n^2\right) \\
-0.5\log|\boldsymbol{\Lambda}_{n-1}^{-1} + \lambda^{-1}\boldsymbol{I}| &\geq \frac{d}{2}\log\lambda - \frac{1}{2}\log|\nu\boldsymbol{\Lambda}_n^{-1} + \boldsymbol{I}| \\
&- \frac{1}{2}(\lambda - \nu)\mathrm{tr}(\nu\boldsymbol{I} + \boldsymbol{\Lambda}_n)^{-1},
\end{aligned}$$

back to vkf

Lower Bounds

$$\begin{aligned}
\log(P(y_n|\boldsymbol{\phi}_n, \boldsymbol{w}_n)) &\geq -\frac{(2y_n - 1)\boldsymbol{\phi}_n^T \boldsymbol{w}_n}{2} - \log(2) - \log(\cosh(\frac{\xi_n}{2})) \\
&- \frac{\tanh(\frac{\xi_n}{2})}{4\xi_n}\left(\left(\frac{\boldsymbol{\phi}_n^T \boldsymbol{w}_n}{2}\right)^2 - \xi_n^2\right) \\
-0.5\log|\boldsymbol{\Lambda}_{n-1}^{-1} + \lambda^{-1}\boldsymbol{I}| &\geq \frac{d}{2}\log\lambda - \frac{1}{2}\log|\nu\boldsymbol{\Lambda}_n^{-1} + \boldsymbol{I}| \\
&- \frac{1}{2}(\lambda - \nu)\mathrm{tr}(\nu\boldsymbol{I} + \boldsymbol{\Lambda}_n)^{-1},
\end{aligned}$$

$$\begin{aligned}
&-0.5(\boldsymbol{w}_n - \hat{\boldsymbol{w}}_{n-1})^T(\boldsymbol{\Lambda}_{n-1}^{-1} + \lambda^{-1}\boldsymbol{I})^{-1}(\boldsymbol{w}_n - \hat{\boldsymbol{w}}_{n-1}) \geq \\
&\quad -0.5(\boldsymbol{w}_n - \hat{\boldsymbol{w}}_{n-1})^T(\boldsymbol{\Lambda}_{n-1}^{-1} + \nu^{-1}\boldsymbol{I})^{-1}(\boldsymbol{w}_n - \hat{\boldsymbol{w}}_{n-1}) \\
&\quad -0.5(\lambda - \nu)(\boldsymbol{w}_n - \hat{\boldsymbol{w}}_{n-1})^T(\nu\boldsymbol{\Lambda}_{n-1}^{-1} + \boldsymbol{I})^{-2}(\boldsymbol{w}_n - \hat{\boldsymbol{w}}_{n-1})
\end{aligned}$$

back to vkf