Biomedical Applications and the Probabilistic Framework

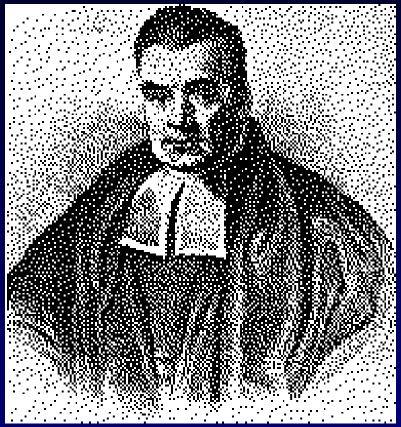
Peter Sykacek

http://www.sykacek.net

Talk Overview

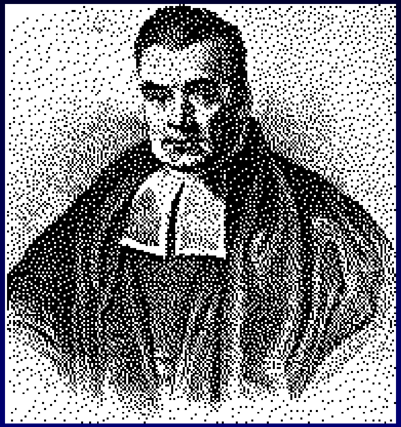
- Motivation of Probabilistic Concepts
- BCI, current practice & shortcomings
- Probabilistic Kalman Filter
- Adaptive BCI
- Gene Discovery
- DAG for Bayesian Marker Identification
- Gene Selection
- Discussion of Model Selection

Probabilistic Motivations



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

Probabilistic Motivations



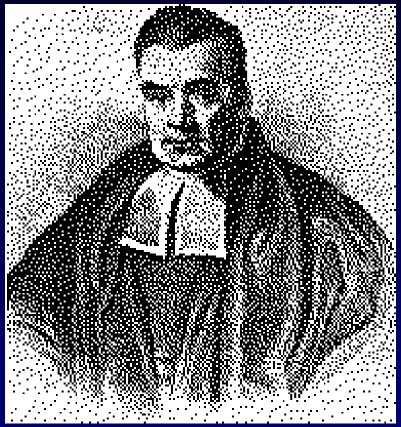
Thomas Bayes (1701 - 1763)

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

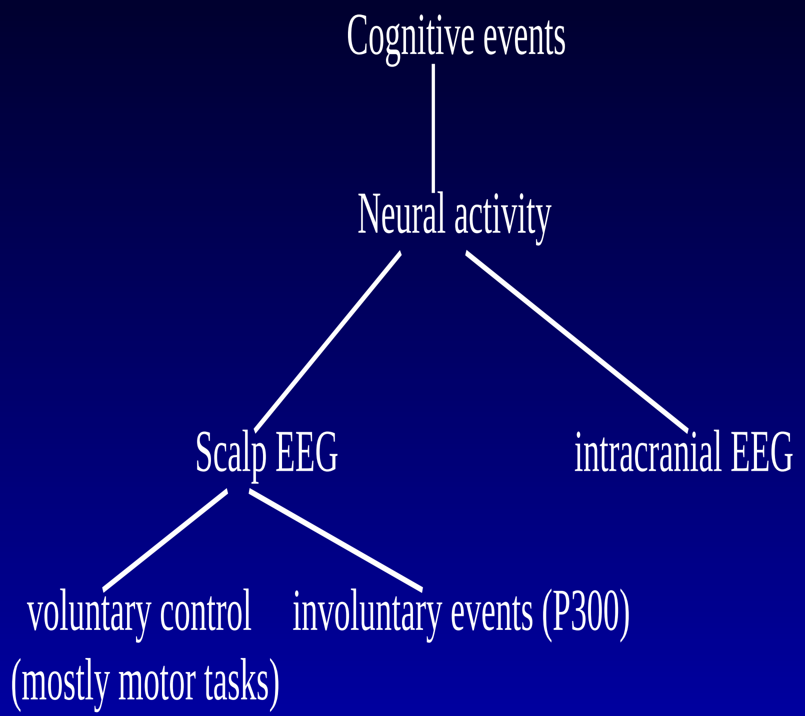
$\alpha_{opt} = \mathrm{argmax}_{\alpha} < u(\alpha) >$, where $< u(\alpha) >= \int_x u(\alpha, x)p(x|\mathcal{D})dx$.

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.

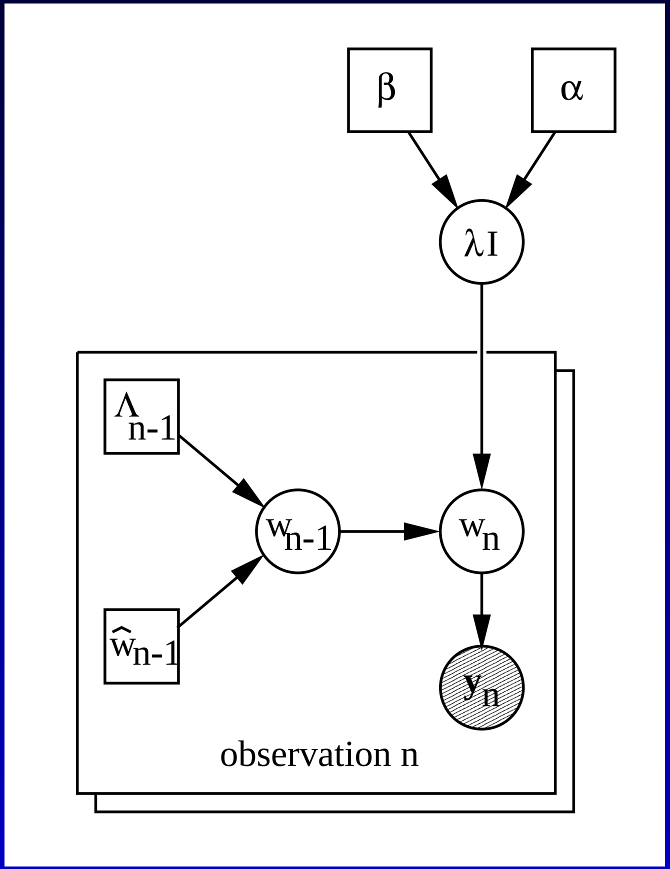
What if

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

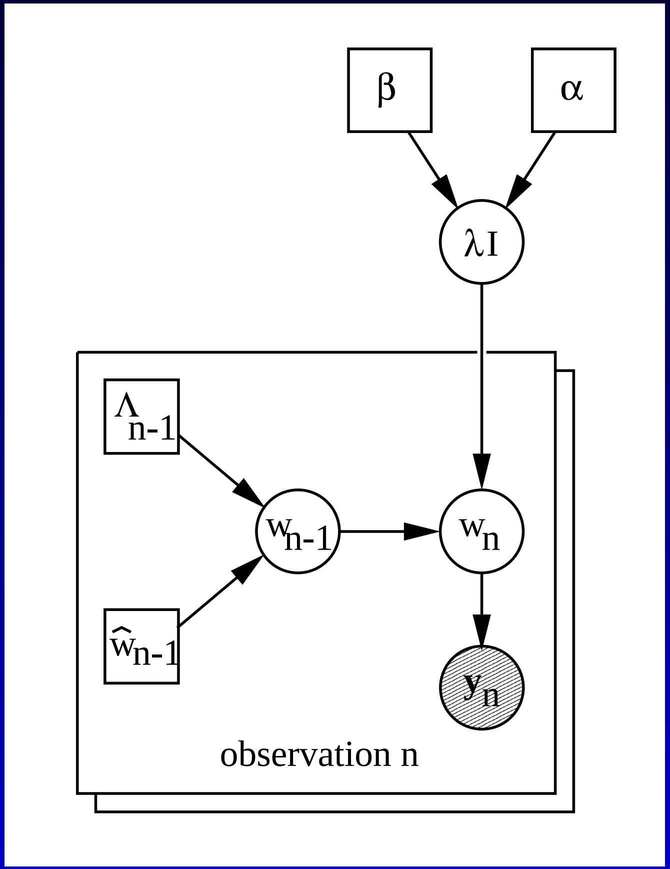
Assuming stationarity must be wrong !

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

Probabilistic Kalman Filter

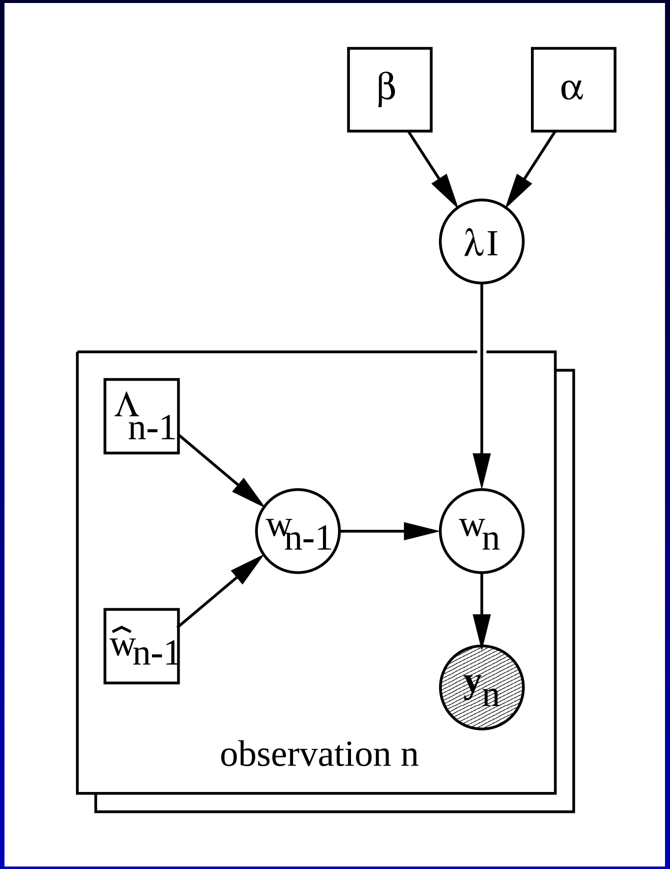


Probabilistic Kalman Filter



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

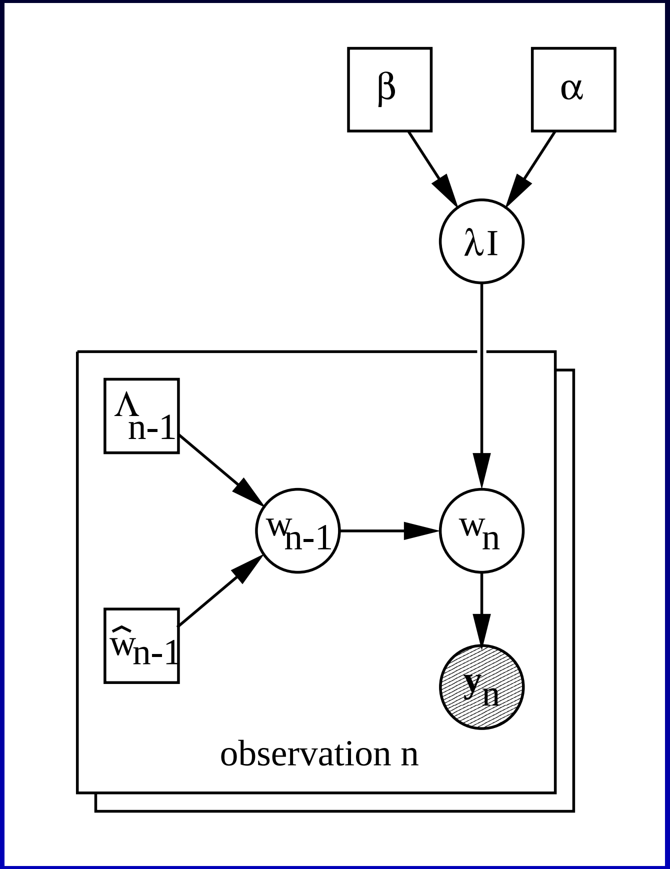
Probabilistic Kalman Filter

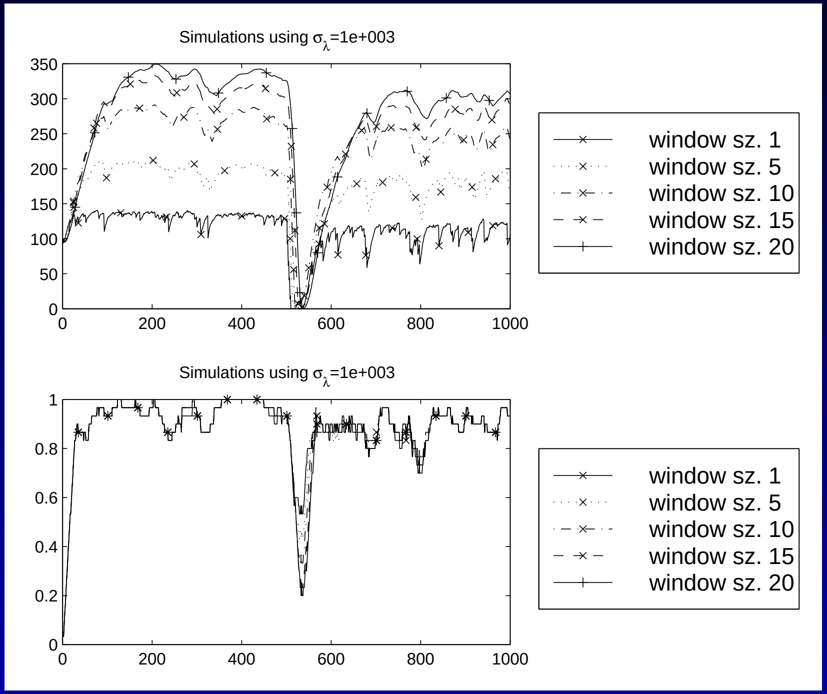


Key: get λ right (may regard $1/\lambda$ as learning rate)

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).

jump 2 TOC

Adaptive BCI

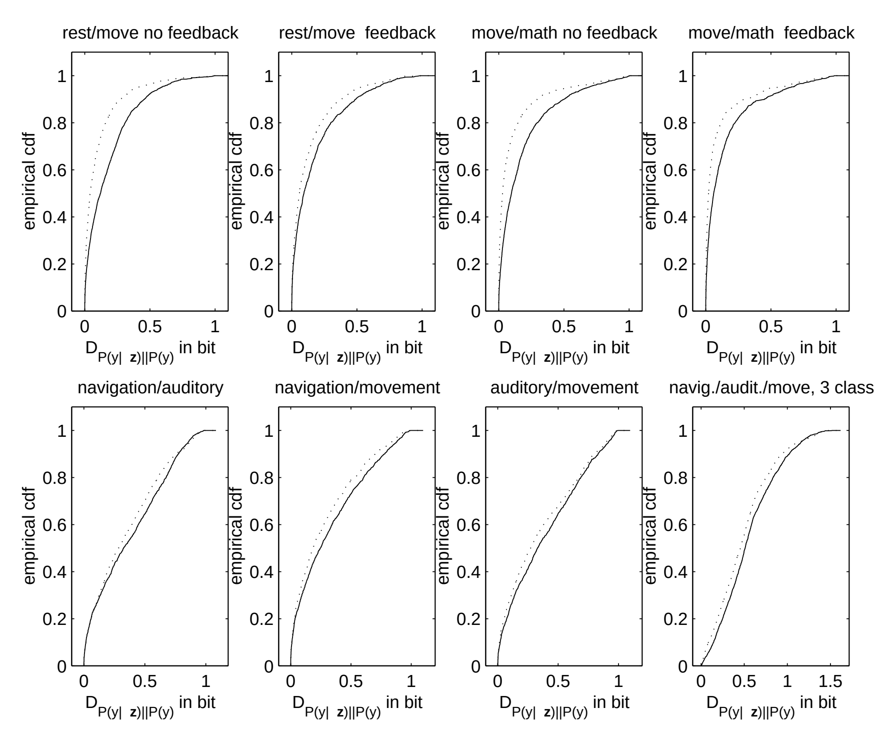
by variational Kalman filtering.

BCI: data driven prediction of cognitive state from EEG measurements.

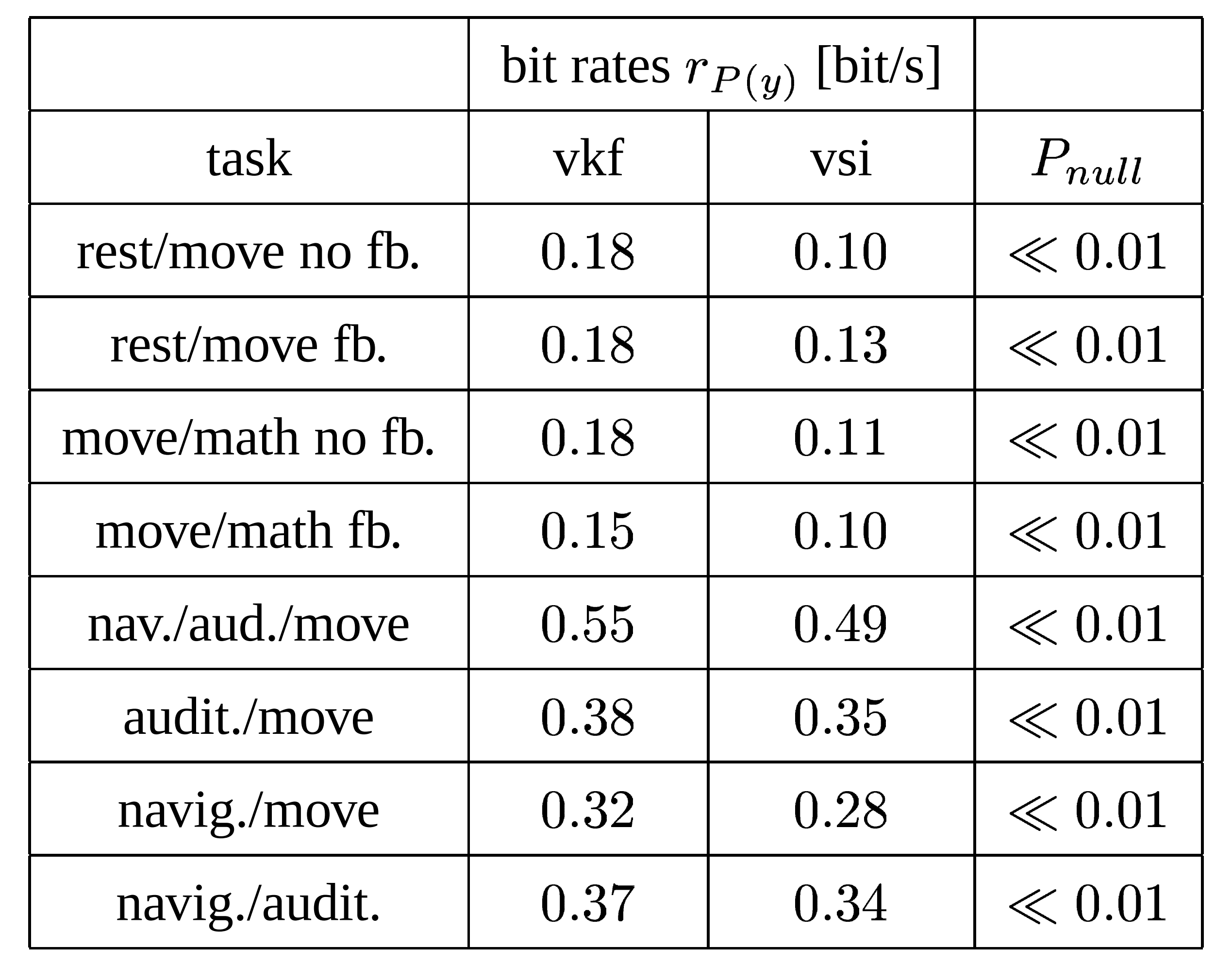
Working hypothesis: EEG dynamics during a cognitive task are subject to temporal variation (learning effects, fatigue ...)

Represent EEG segments by z-transformed reflection coefficients.

Mutual information, adaptive method and identical “stationary” model.



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: adaptive methods increase BCI bandwidths even on short time scales.

Gene discovery

Discovering “important” genes (or proteins) from microarray datasets can be classified as

- Identification of all differentially expressed genes.
- Identification of reliable (sets) of marker genes.

Current practise for the first: classical methods (e.g. t-test on differences of means) or probabilistic approaches with one indicator variable for each gene.

The second is typically done by conventional feature subset selection. As a result we obtain a set of genes that was found by heuristic search.

Bayesian Marker Identification

Missing in FSS: How good are other explanations?

Interpret microarray data as classification problem of “genetic” regressors w.r.t. a discrete response.

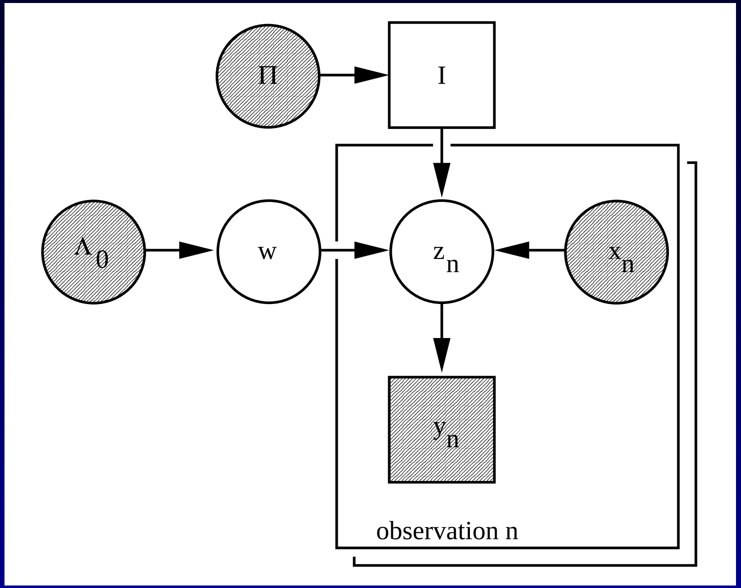
$->$ Bayesian variable selection provides this information. However: hopeless, unless we constrain the dimensionality.

Simplified attempt: $->$ Find distribution over individual genes.

Probabilities result from the marginal likelihood of each model.

$$P(I|\mathcal{D}) = \frac{\int_{\boldsymbol{w}} p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w}|I)P(I)}{\sum_I \int_{\boldsymbol{w}} p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w}|I)P(I)d\boldsymbol{w}},$$

DAG for Marker Identification

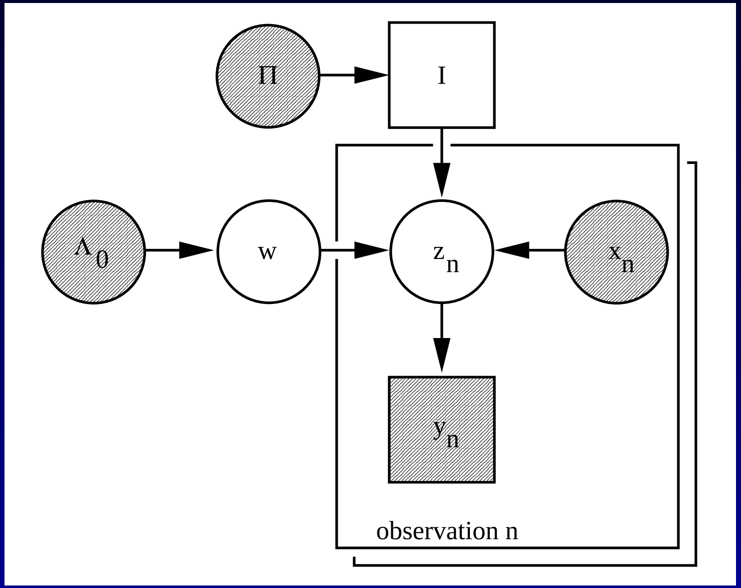


Latent variable probit GLM.

z_n is a one dimensional Gaussian random variable with mean $\boldsymbol{w}^T\boldsymbol{x}_n$ and precision 1.

$$P(y_n \equiv 1|z_n) = \begin{cases} 1, \text{ if } z_n > 0 \\ 0, \text{ if } z_n \leq 0 \end{cases}$$

DAG for Marker Identification



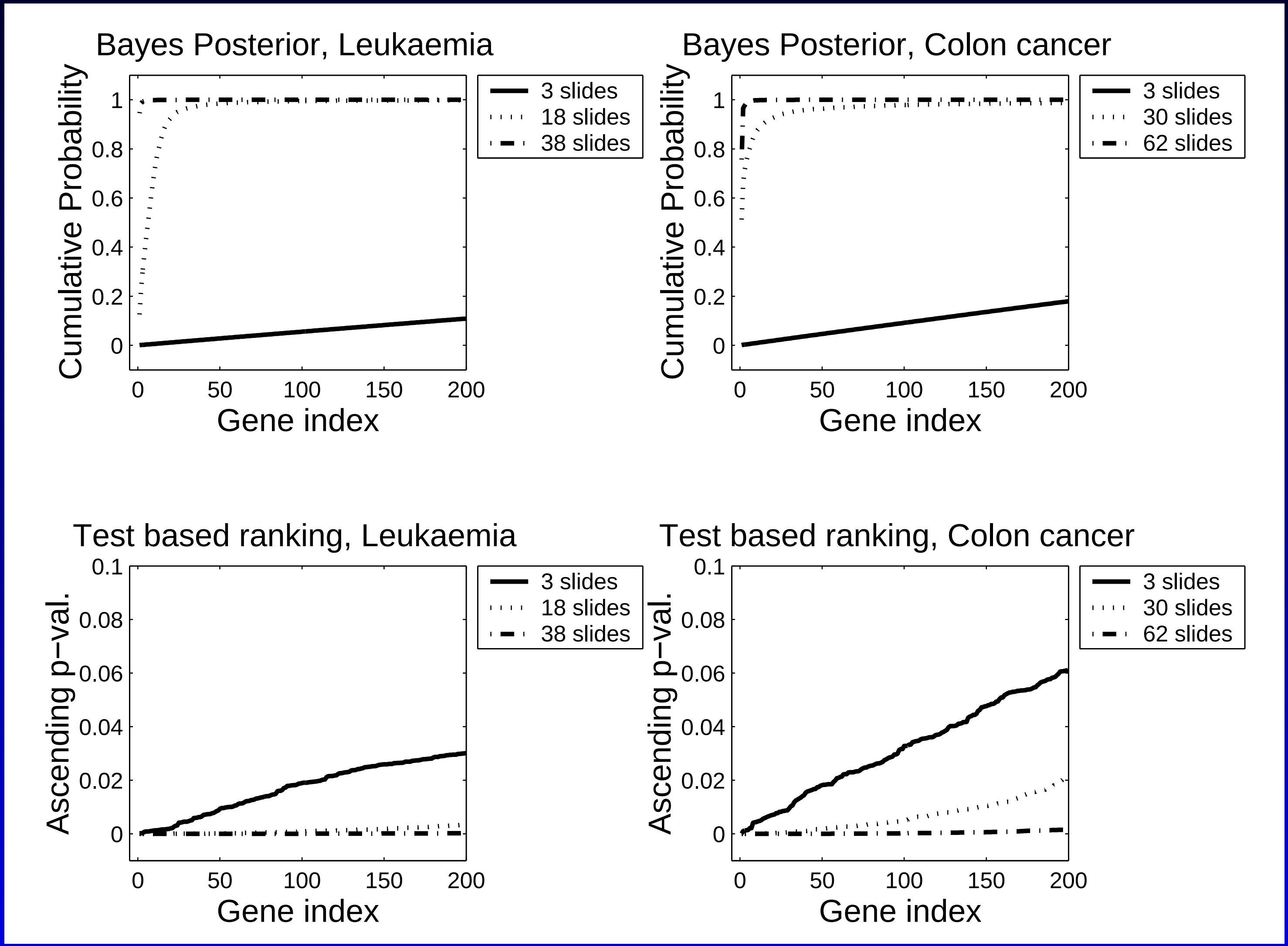
Latent variable probit GLM.

z_n is a one dimensional Gaussian random variable with mean $\boldsymbol{w}^T\boldsymbol{x}_n$ and precision 1.

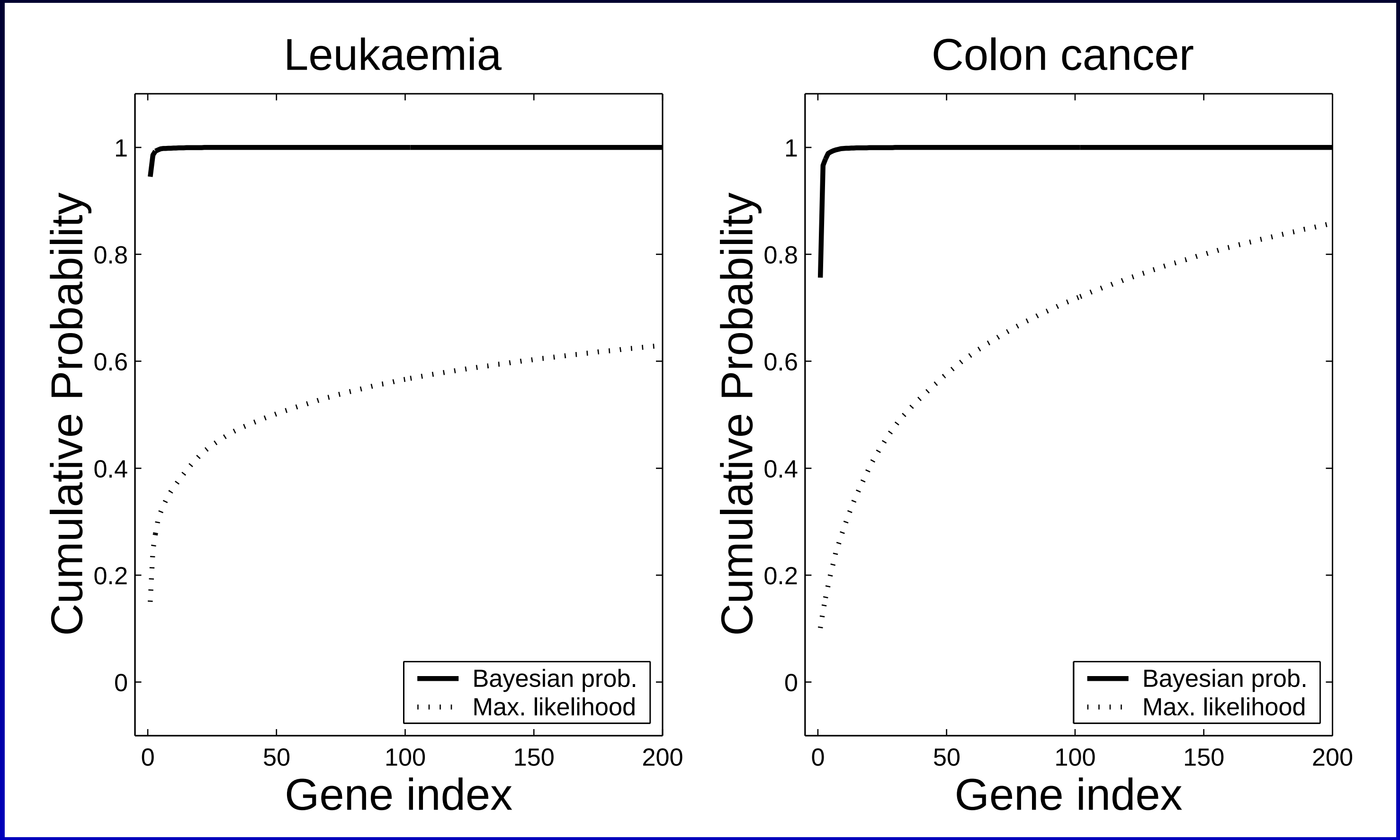
$$P(y_n \equiv 1|z_n) = \begin{cases} 1, \text{ if } z_n > 0 \\ 0, \text{ if } z_n \leq 0 \end{cases}$$

Inference can be done by a variational method (systematic error) or by sampling (random error). The latter allows to integrate over $\boldsymbol{w}$ analytically and we draw from z_n and I only.

Asymptotic Behaviour



Comparison with ML



Results differ since Bayesian model posteriors take “complexity” (ref. Hochreiter’s “flat minima”) into account.

Selection and Gen. Accuracy

Most probable regressors selected at a 0.99 threshold

Acc. no.	description	$P(I\|\mathcal{D})$
Colon Cancer (Alon et. al.)		
Z50753	Uroguanylin	0.76
R87126	Myosin	0.21
M63391	desmin gene	0.01
M36634	vasoact. pept.	0.01
Leukaemia (Golub et. al.)		
X95735	Zyxin	0.93
M55150	FAH Fumarylac.	0.05
M27891	CST3 Cystatin C	0.01

Generalization accuracy

Dataset	B. probit	“indifference”
Colon	84%	74% to 94%
Leukaemia	88%	91% to 96%

No “better” results in literature $->$ confirms model.

Biology confirms Uroguanylin (cell apoptosis) as important in colon cancer development.

Selection and Gen. Accuracy

Most probable regressors selected at a 0.99 threshold

Acc. no.	description	$P(I\|\mathcal{D})$
Colon Cancer (Alon et. al.)		
Z50753	Uroguanylin	0.76
R87126	Myosin	0.21
M63391	desmin gene	0.01
M36634	vasoact. pept.	0.01
Leukaemia (Golub et. al.)		
X95735	Zyxin	0.93
M55150	FAH Fumarylac.	0.05
M27891	CST3 Cystatin C	0.01

Generalization accuracy

Dataset	B. probit	“indifference”
Colon	84%	74% to 94%
Leukaemia	88%	91% to 96%

No “better” results in literature $->$ confirms model.

Biology confirms Uroguanylin (cell apoptosis) as important in colon cancer development.

But: Meaning of the probabilities?

Discussion

Quoting $P(I|\mathcal{D}) -> \mathcal{M}$-closed model selection with zero-one utility.

Our approach should assume an $\mathcal{M}$-open scenario.

Under asymptotic normality, $P(I|\mathcal{D})$ degenerates on $I_i \in \mathcal{M}$ that minimizes $\int p(y|\boldsymbol{w}_t) \log(p(y|\hat{\boldsymbol{w}}_i)/p(y|\boldsymbol{w}_t))dy$).

If the predictive distribution of a new observation is of interest, B&S's suggest to use a logarithmic score function for $\mathcal{M}$-open model comparison.

$$\int \log(p(y|I_i, \mathcal{D}))p(y|\mathcal{D})dy$$

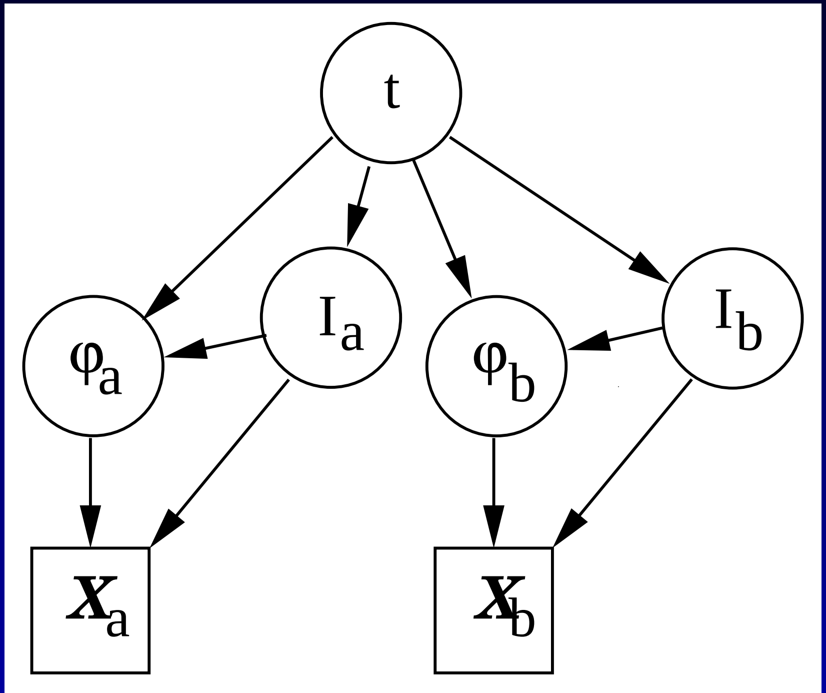
(e.g. cross validation estimate, still to be done)

A simple idea:

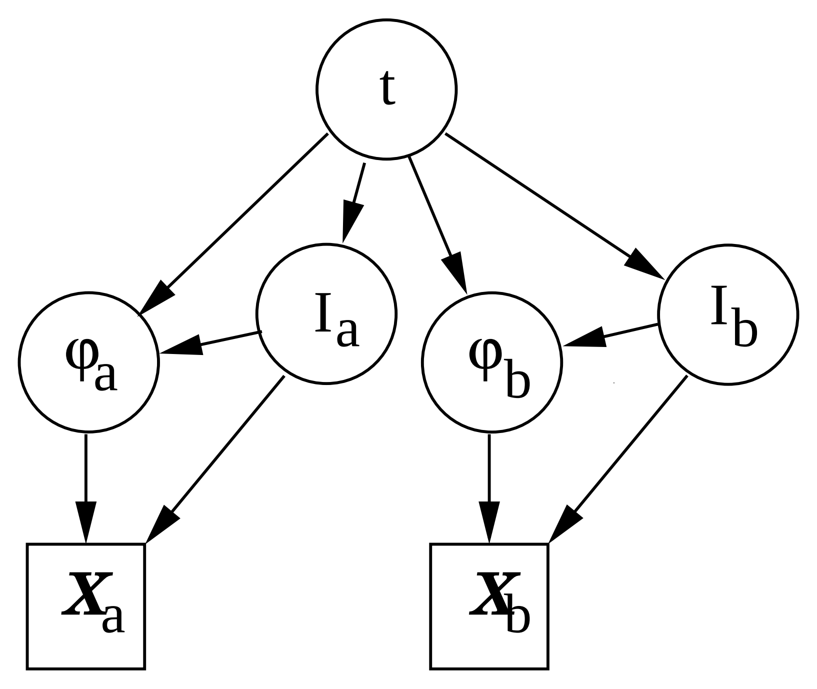
the world is one probabilistic model.

- 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

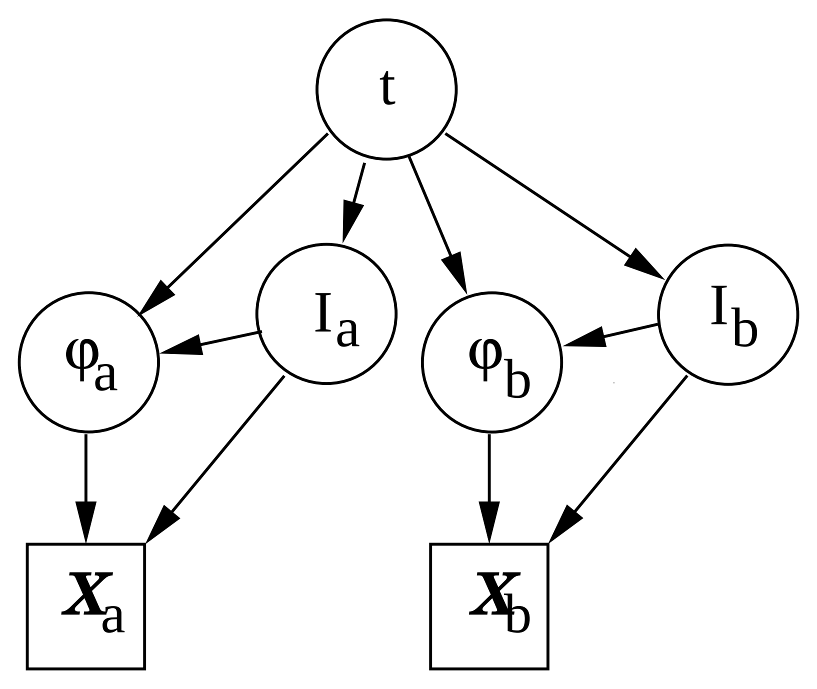


Bayes' Consistent Models

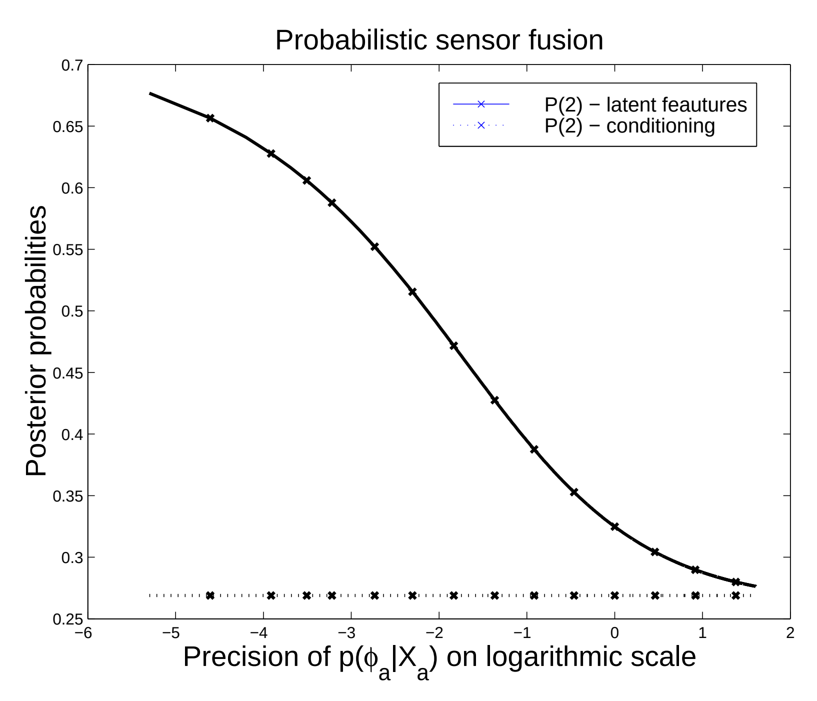


Expected utility requires to integrate over all unknown variables, including φ_a, φ_b, I_a and I_b that represent a feature space.

Bayes’ Consistent Models



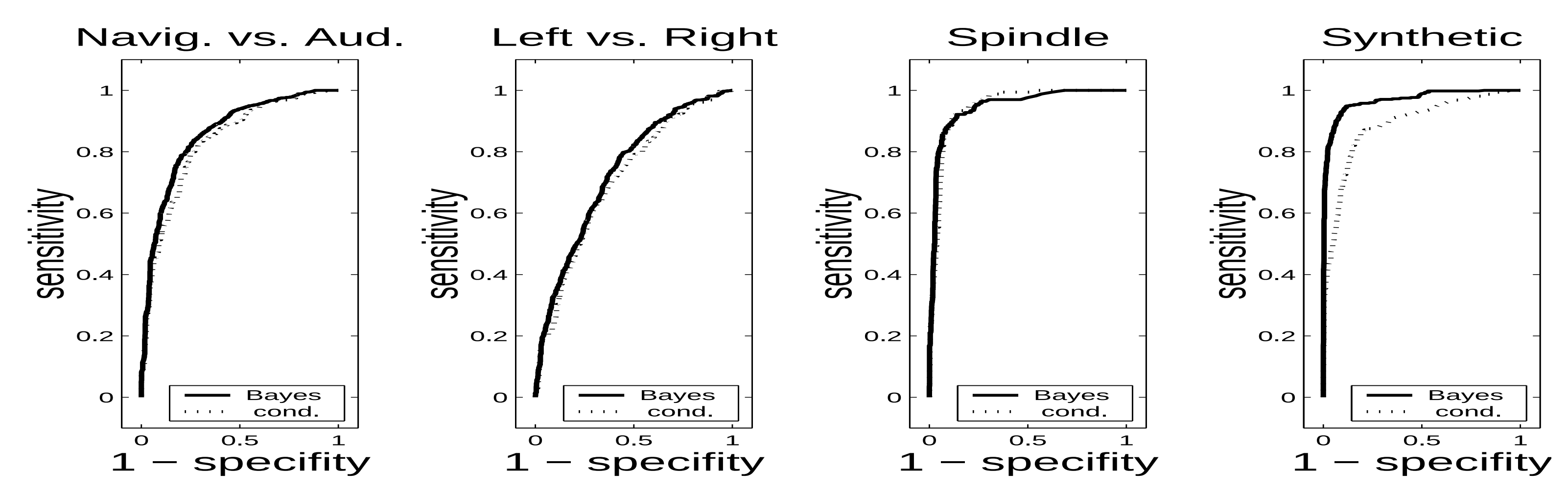
Expected utility requires to integrate over all unknown variables, including φ_a, φ_b, I_a and I_b that represent a feature space.



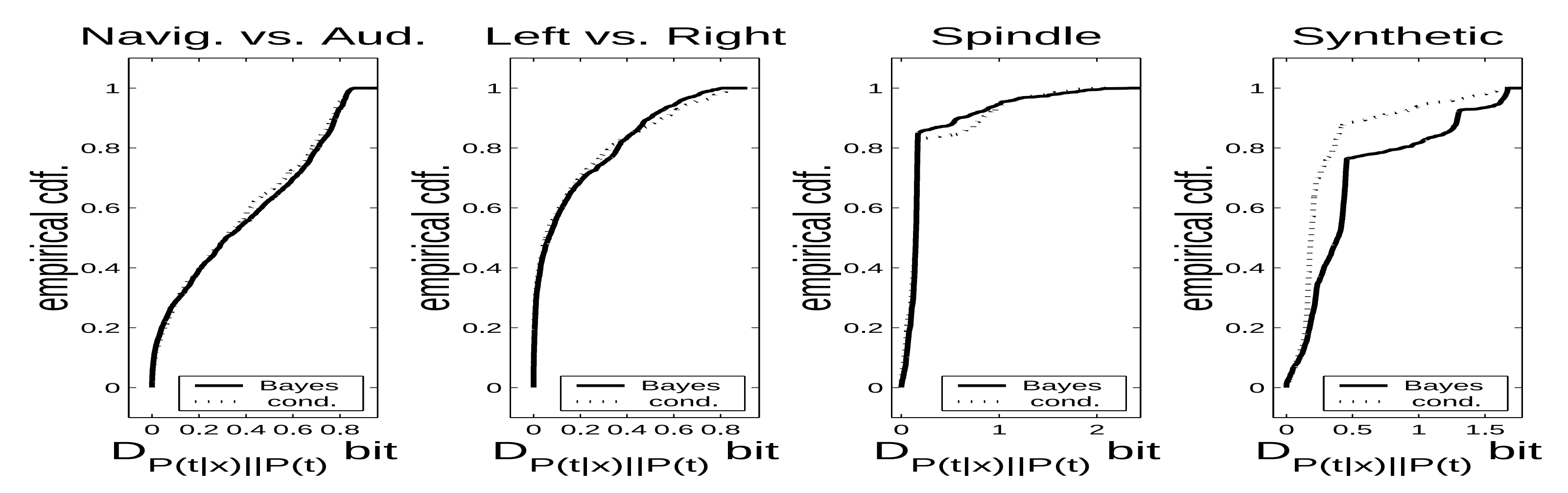
Decisions depend on (un)certainty and may thus change.

Time Series Classification

ROC Curves



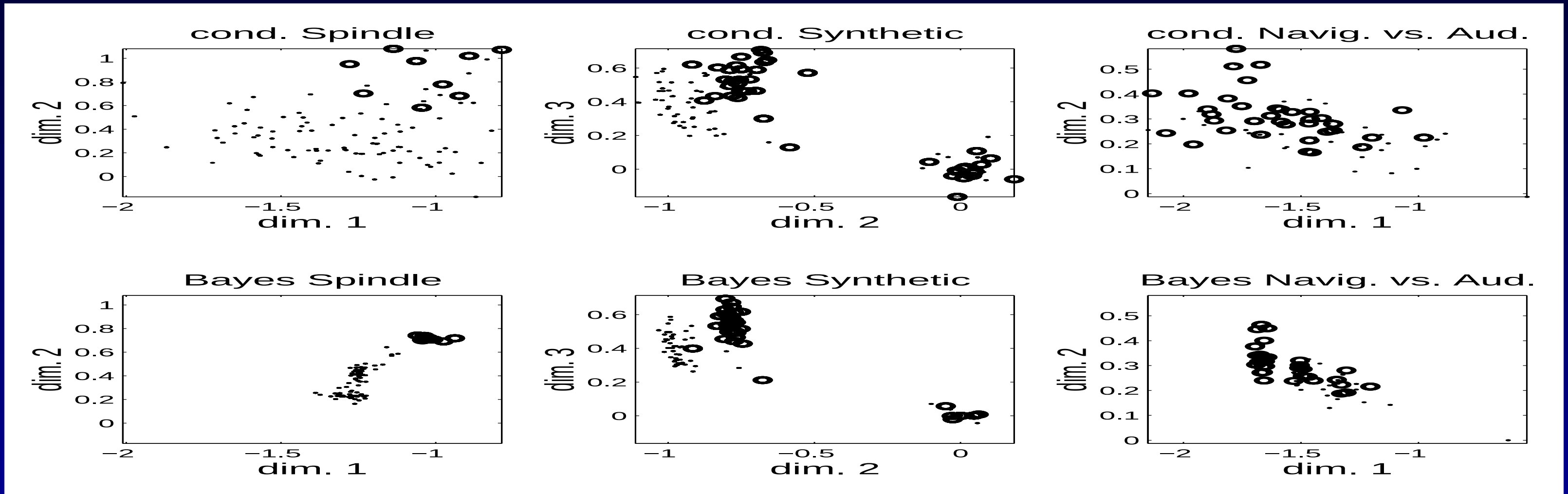
Kullback Leibler Divergence



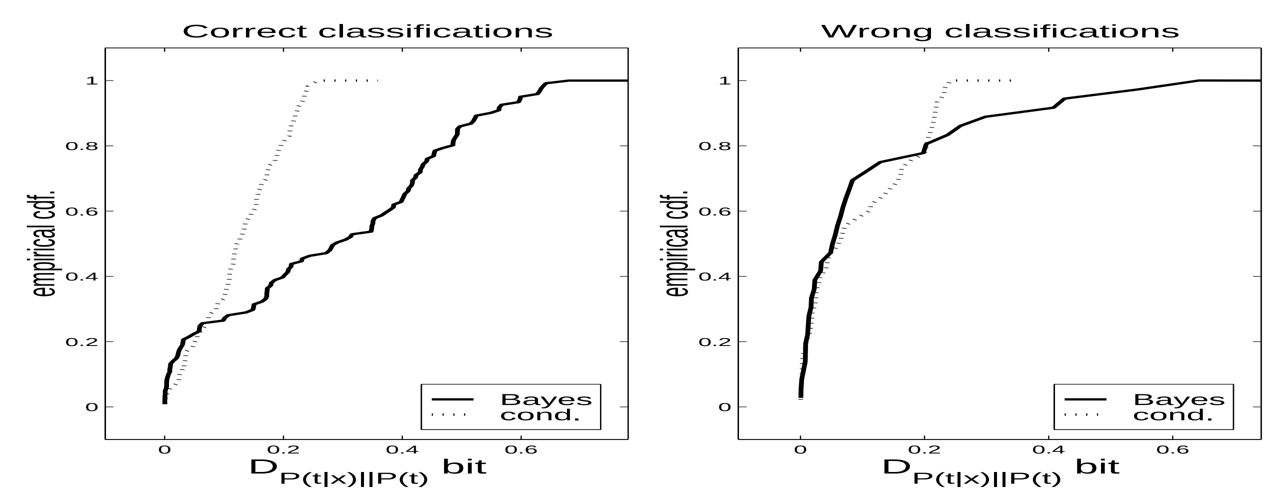
jump 2 TOC

More Results

Expected feature values



Kullback Leibler Divergence for “Artefacts”



Variational Kalman Filter

The logarithmic model evidence for a window of size N is

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

This is not a probabilistic structure! (need Rauch Tung Striebel smoother)

Variational Kalman Filter

The logarithmic model evidence for a window of size N is

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

This is not a probabilistic structure! (need Rauch Tung Striebel smoother)

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

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

Lower Bounds

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

back to vkf

jump 2 TOC

Lower Bounds

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

back to vkf

Lower Bounds

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

$$\begin{aligned}
-0.5 \log |\boldsymbol{\Lambda}_{n-1}^{-1} + \lambda^{-1} \boldsymbol{I}| &\geq \frac{d}{2} \log \lambda - \frac{1}{2} \log |\nu \boldsymbol{\Lambda}_n^{-1} + \boldsymbol{I}| \\
&- \frac{1}{2}(\lambda - \nu)\mathrm{tr}(\nu \boldsymbol{I} + \boldsymbol{\Lambda}_n)^{-1},
\end{aligned}$$

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

back to vkf

jump 2 TOC