Probabilistic data fusion for time series classification

P. Sykacek, S. J. Roberts

Robotics Research Group, Department of Engineering Science,
University of Oxford, Parks Road, Oxford OX1 6PJ, UK
psyk@robots.ox.ac.uk
sjrob@robots.ox.ac.uk
http://www.robots.ox.ac.uk/~parg/

Outline

- Define probabilistic sensor fusion and propose a model for this purpose.
- Illustration of properties of such models w.r.t. Bayesian theory and information fusion.
- Propose a "sensor fusing" model for time series classification.
- Short discussion of a MCMC approach for inference of the proposed model
- Experimental evaluation and conclusion.

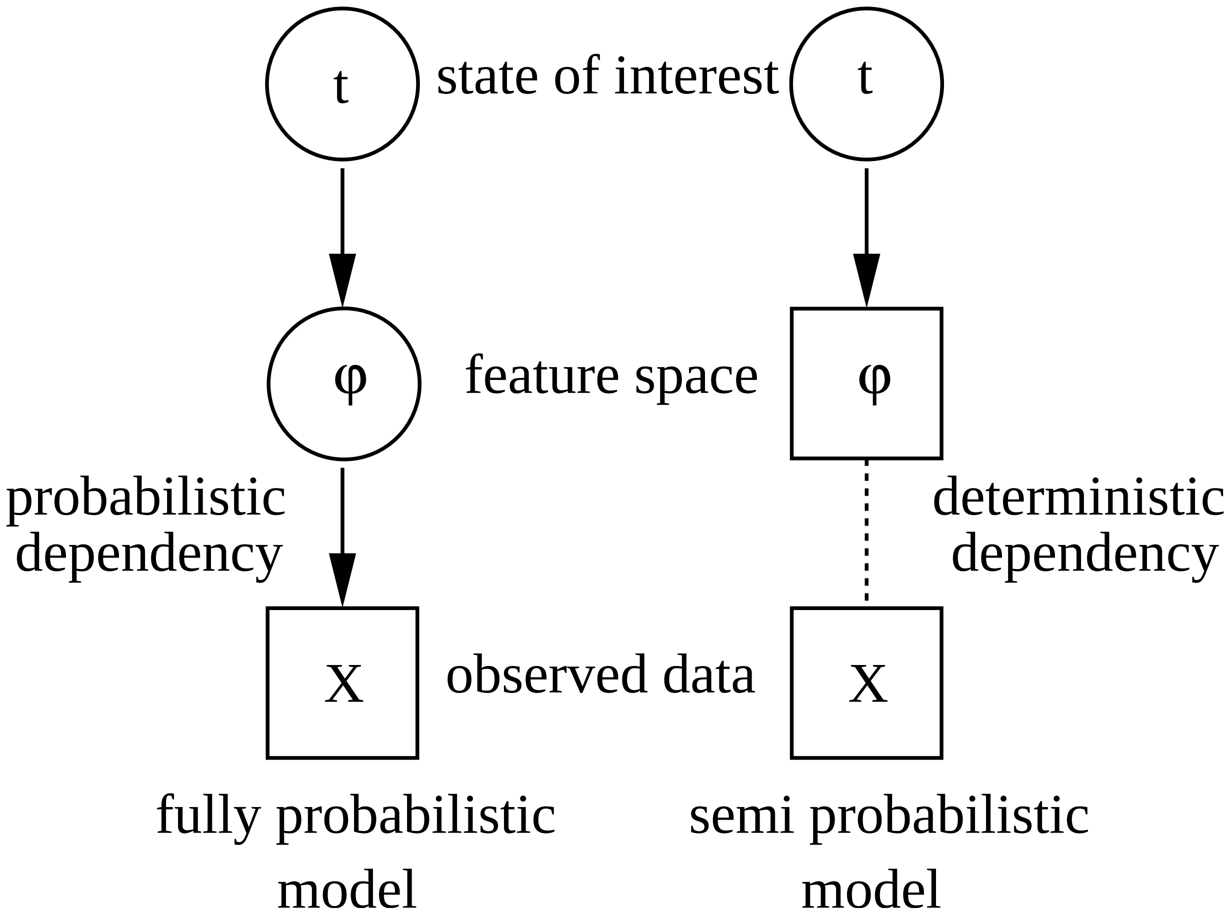
A simple idea: the world is one probabilistic model

- 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.

Probabilistic sensor fusion

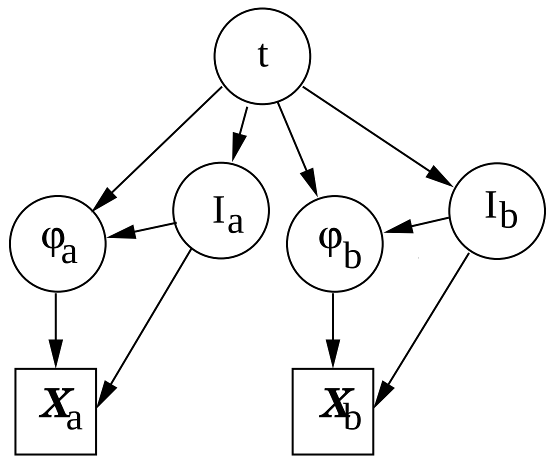


Some Bayesian motivations for our suggestion

- We infer features from limited amount of data. $->$ Predicting the state of interest (t), considers parameter and model uncertainty. Sensor fusion is based on certainty of information.

- The idea provides also a consistent prior in the feature space. Classical settings fail since their priors neglect infromation obtained from previous observations.

Marginal inference in a naïve Bayes' model

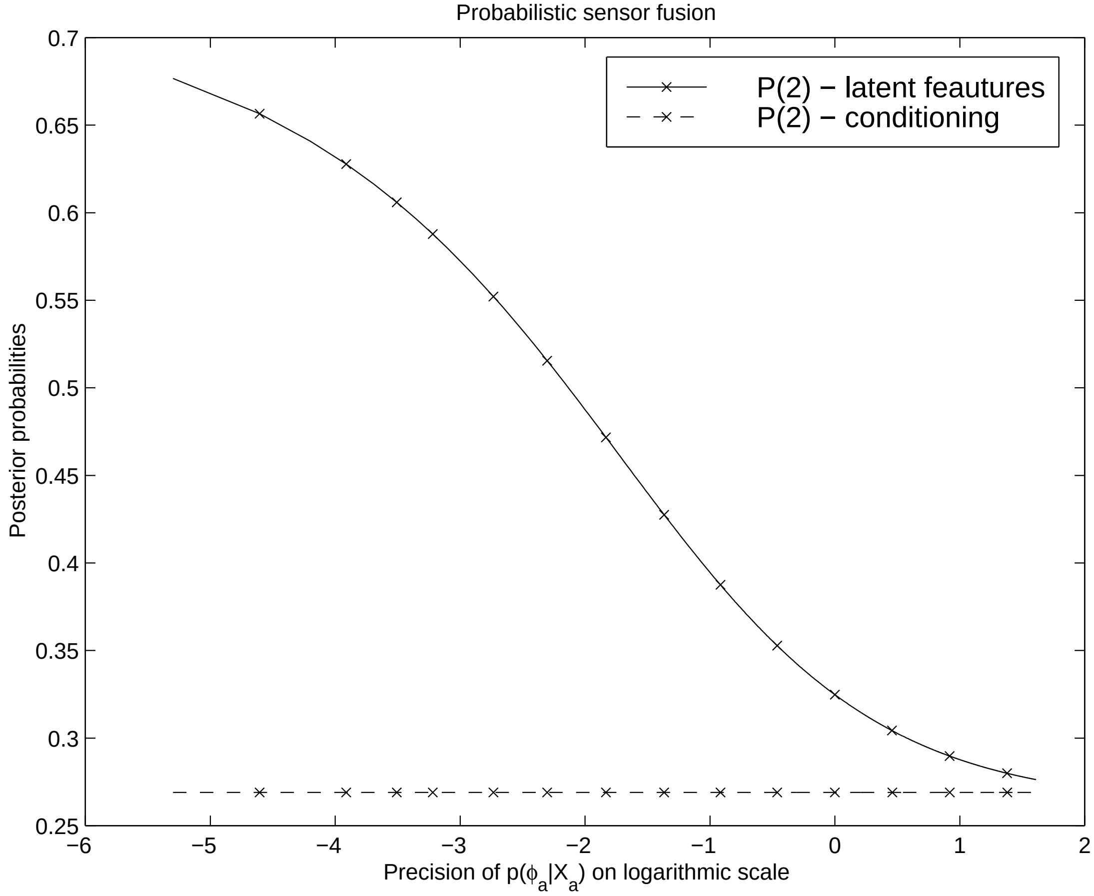


Suppose we want the probability $P(t|\mathcal{X}_a, \mathcal{X}_b)$ for the above DAG:

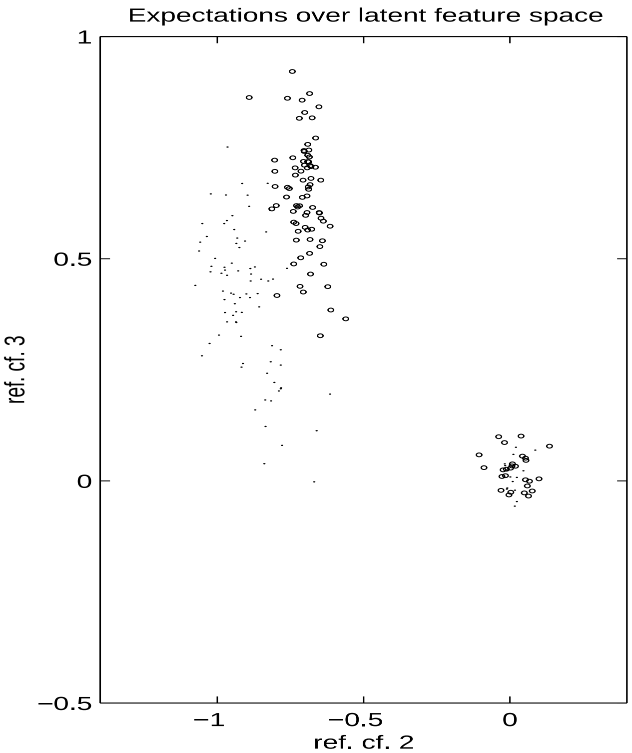
$$P(t|\mathcal{X}_a, \mathcal{X}_b) = \frac{p(\mathcal{X}_a)p(\mathcal{X}_b)}{p(\mathcal{X}_a, \mathcal{X}_b)} \frac{1}{P(t)} \left(\sum_{I_a} \int_{\boldsymbol{\varphi}_a} P(t|\boldsymbol{\varphi}_a, I_a) p(\boldsymbol{\varphi}_a, I_a|\mathcal{X}_a) d\boldsymbol{\varphi}_a \right.$$

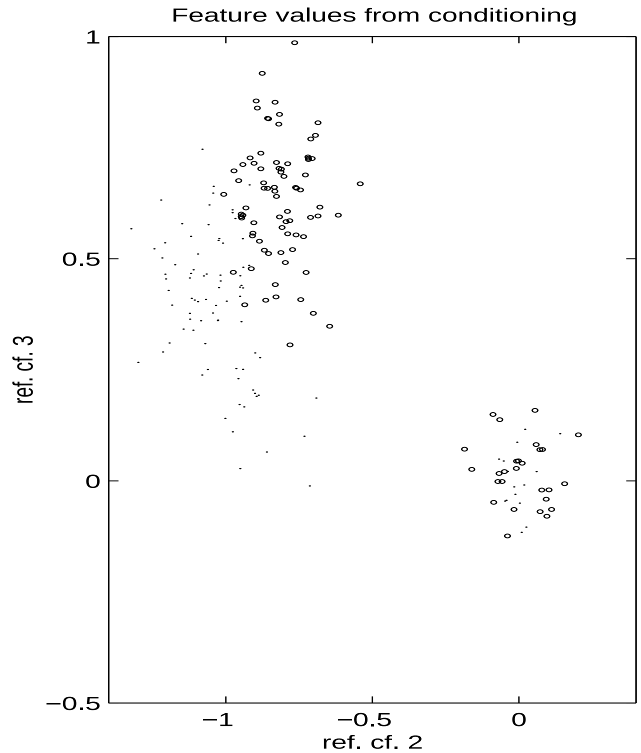
$$\left. \times \sum_{I_b} \int_{\boldsymbol{\varphi}_b} P(t|\boldsymbol{\varphi}_b, I_b) p(\boldsymbol{\varphi}_b, I_b|\mathcal{X}_b) d\boldsymbol{\varphi}_b \right), \quad (1)$$

Illustration of certainty based $P(t|\mathcal{X}_a, \mathcal{X}_b)$ vs. conditioning



Similar: expected values in the latent space



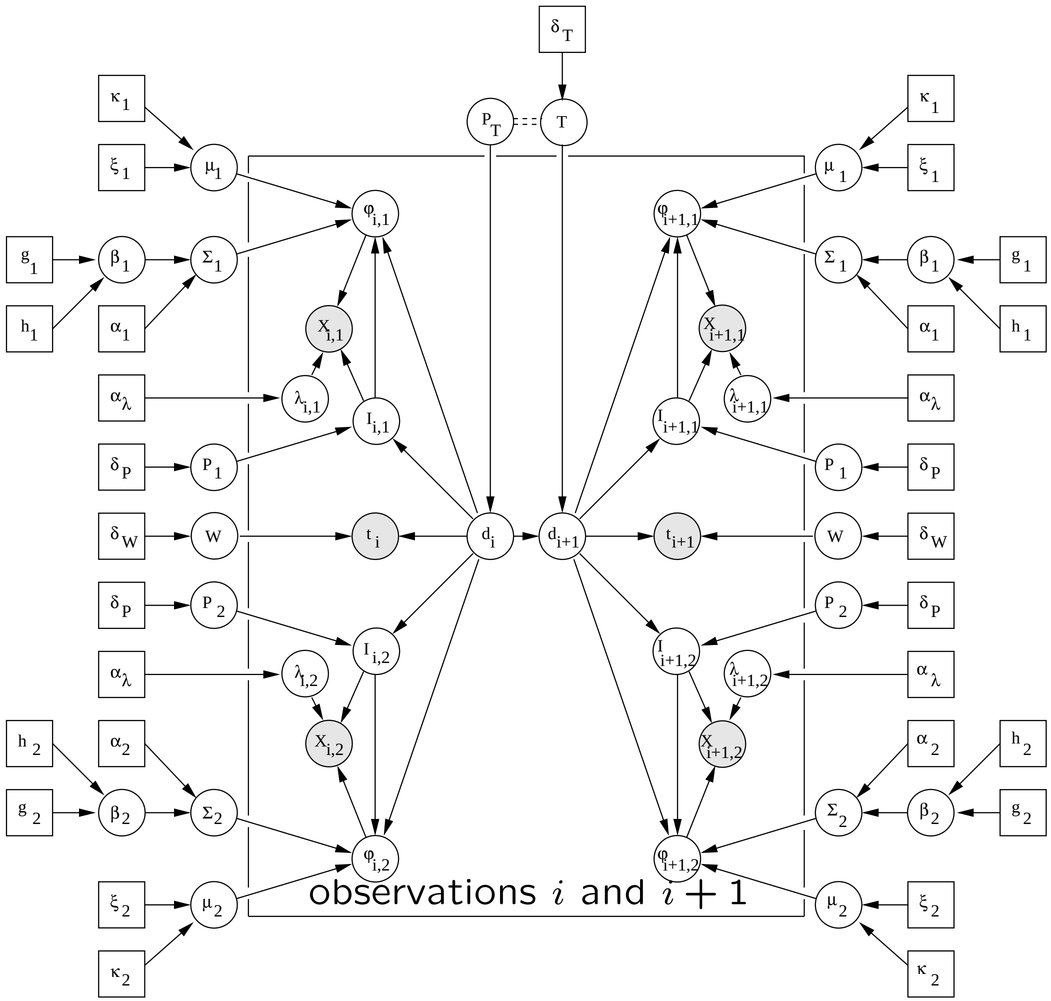


A probabilistic model for time series classification

- The objective is to classify successive segments of multivariate time series data
- To allow for temporal correlations we use a hidden Markov model like architecture
- The latent feature space is modeled by diagonal Gaussian and Multinomial distributions
- Class labels are modeled by Multinomial distributions

$->$ Gaussian and Multinomial observations hidden Markov model (GMOHMM)

The GMOHMM


δT
PT
T
κ1
ξ1
μ1
g1
β1
Σ1
h1
α1
αλ
δP
P1
δW
W
P2
h2
α2
g2
β2
Σ2
ξ2
μ2
κ2
φi,1
Xi,1
λi,1
Ii,1
ti
di
di+1
ti+1
Ii,2
λi,2
Xi,2
φi,2
φi+1,1
Xi+1,1
λi+1,1
Ii+1,1
Ii+1,2
λi+1,2
Xi+1,2
φi+1,2
observations i and i + 1

Legend:

variable	meaning	variable	meaning
t_i	class label	d_i	state variable
$\boldsymbol{\varphi}_{i,s}$	feature variable	$I_{i,s}$	model indicator
$\lambda_{i,s}$	noise precision	s	sensor number
$\mathcal{X}_{i,s}$	time series segment	$\boldsymbol{T}$	transition probabilities
$\boldsymbol{W}$	class probabilities	$\boldsymbol{\mu}_s$	kernel means
$\boldsymbol{\Sigma}_s$	kernel covariances	$\boldsymbol{P}_s$	indicator probabilities
δ_T, δ_W, δ_P	prior counts	ξ_s, κ_s	Gaussian prior
α_s, β_s	Gamma prior	g_s, h_s	Gamma prior
α_λ	Jeffrey's prior	i	time index
$s_{i,s}$	sufficient statistics		

A feasible representation of the latent space

...based on second order statistics of the time series.

- We must use reflection coefficients (i.e. partial correlation coefficients) since unlike AR cfs. this representation does not depend on the model order.

- We want to model $p(\boldsymbol{\varphi}_{I_{i,s}}|d_i)$ by Gaussians. However $\boldsymbol{\varphi_{i,s}} \in \Re^{I_{i,s}}$ and assuming dynamic stability there is a mismatch with $\boldsymbol{\rho}_{i,s} \in [-1, 1]^{I_{i,s}} \subset \Re^{I_{i,s}}$.
  $->$ To adjust that, we use the homomorphic transformation $\boldsymbol{\varphi}_{i,s} = \text{artanh}(\boldsymbol{\rho}_{i,s})$.

AR model

$$x[t] = \sum_{m=1}^{I_{i,s}} a^m_{I_{i,s}} x[t-m] + \epsilon[t] \quad (2)$$

$y[t]$: sample of the time series; $I_{i,s}$: model order; $a^m_{I_{i,s}}$: m-th AR coefficient; $\epsilon[t]$: sample of i.i.d. white noise with precision $\lambda_{i,s}$.

... in lattice filter representation (Levinson-Durbin recursion):

$$\boldsymbol{a}_{I_{i,s}+1} = \begin{bmatrix} \boldsymbol{a}_{I_{i,s}} + \rho_{(I_{i,s}+1)} \boldsymbol{a}^{\circlearrowleft}_{I_{i,s}} \\ \rho_{(I_{i,s}+1)} \end{bmatrix} \quad (3)$$

Reparameterise $\boldsymbol{a}_{I_{i,s}}$ as reflection coefficients. $\rho_{I_{i,s}}$: $I_{i,s}$-th reflection coefficient; $\boldsymbol{a}_{I_{i,s}}{}^{\circlearrowleft}$: $I_{i,s}$-th order AR coefficient vector multiplied by an exchange matrix.

Inference (I): Conjugate priors

- Each component mean gets a Gaussian prior: $\boldsymbol{\mu}_{i,s} \sim \mathcal{N}_1(\boldsymbol{\xi}_s, \boldsymbol{\kappa}_s^{-1})$.
- We have diagonal covariance matrices $->$ each diagonal element has an independent Gamma prior: $\boldsymbol{\Sigma}_{i,s}[j,j]^{-1} \sim \Gamma(\alpha_s, \boldsymbol{\beta}_s[j])$.
- The hyper-parameters get Gamma priors: $\boldsymbol{\beta}_s[j] \sim \Gamma(g_s, \boldsymbol{h}_s[j])$.
- The state conditional class probabilities have Dirichlet priors: $\boldsymbol{W} \sim \mathcal{D}(\delta_W, .., \delta_W)$.
- The transition probabilities have Dirichlet priors: $\boldsymbol{T} \sim \mathcal{D}(\delta_T, .., \delta_T)$.
- The observation probabilities of model orders have Dirichlet priors $\boldsymbol{P}_s \sim \mathcal{D}(\delta_P, .., \delta_P)$.
- The precision $\lambda_{i,s}$ gets a Jeffrey's prior. That is the scale parameter a_λ is set to 0.

Inference (II): MCMC method

- Integrals occuring during inference not analytically tractable $->$ approch it with MCMC methods. Whenever possible use Gibbs updates (standard and easily found in literature).
- Focus here on updates for marginalizing the latent feature space.

We assume:

- admissible model orders between 0 and I_{max}.
- set $P(\text{move}(\mathcal{C}_I - > \mathcal{C}_{I+1})|\mathcal{C}_I) \equiv P(\text{move}(\mathcal{C}_{I+1} - > \mathcal{C}_I)|\mathcal{C}_{I+1})$.

marginalizing the latent feature space $->$ 2 move types minimum: a) within model class $\mathcal{C}_I$; b) between successive model classes $\mathcal{C}_I$ and $\mathcal{C}_{I+1}$.

Metropolis Hastings for updates within model class

A convenient proposal (likelihood ratio $\times$ proposal ratio is 1!):

$$\begin{aligned} \boldsymbol{\varphi}'_{i,s} &= \text{artanh}(\boldsymbol{\rho}(\boldsymbol{a}'_{i,s})) \qquad (4) \\ \text{where} & \\ \boldsymbol{a}'_{i,s} &\sim \mathcal{S}t_{\nu}(\hat{\boldsymbol{a}}, \boldsymbol{\Sigma}) \\ \text{with} & \\ \hat{\boldsymbol{a}} &= \boldsymbol{A}^{-1}\boldsymbol{r} \\ \boldsymbol{\Sigma} &= \boldsymbol{A}^{-1}\frac{(R_0 - \boldsymbol{r}^T\boldsymbol{A}^{-1}\boldsymbol{r})}{2\nu} \\ \nu &= N - I_{i,s} \end{aligned}$$

$\boldsymbol{A}$: $I_{i,s}$-dimensional sample auto-covariance matrix, R_0: sample variance, $\boldsymbol{r} = [R_1, ..., R_{I_{i,s}+1}]^T$: vector of sample auto-correlations at lags 1 to $I_{i,s}+1$; and N:number of samples in time series $\mathcal{X}_{i,s}$.

... gives prior ratio as acceptance probability:

$$a = \min\left(1, \frac{p(\boldsymbol{\varphi}'_{i,s})\left|\frac{\partial \boldsymbol{\varphi}'_{i,s}}{\partial \boldsymbol{a}'_{i,s}}\right|}{p(\boldsymbol{\varphi}_{i,s})\left|\frac{\partial \boldsymbol{\varphi}_{i,s}}{\partial \boldsymbol{a}_{i,s}}\right|}\right). \tag{5}$$

We may calculate the Jacobian in analogy with Levinson-Durbin recursion (3).

Reversible jump MC for moves between model classes

Partial proposal from $\mathcal{C}_{I_{i,s}}$ to $\mathcal{C}_{I_{i,s}+1}$ (only one new reflection coefficient):

$$\varphi'_{i,s} = [\boldsymbol{\varphi}_{i,s}, \text{artanh}(\rho)] \tag{6}$$

where

$$\rho \sim \mathcal{S}t_\nu(\hat{\rho}, \sigma)$$

with

$$\hat{\rho} = -\frac{k_2}{k_1}$$

$$\sigma = \sqrt{\frac{1-\hat{\rho}^2}{2(N-1)}}$$

$$\nu = N-1$$

$$k_1 = R_0 + 2\boldsymbol{a}_{i,s}^T \boldsymbol{r}_0 + \boldsymbol{a}_{i,s}^T \boldsymbol{A}_0 \boldsymbol{a}_{i,s}$$

$$k_2 = R_{I+2} + 2\boldsymbol{r}_0^T \boldsymbol{a}_{i,s}{}^{\circlearrowleft} + \boldsymbol{a}_{i,s}^T \boldsymbol{A}_0 \boldsymbol{a}_{i,s}{}^{\circlearrowleft}.$$

Acceptance probability of this move:

$$a = \min\left(1, \left(1 - \frac{k_1^2}{k_2^2}\right)^{-\frac{N-1}{2}} \sqrt{\frac{\pi}{2}} \frac{\Gamma(\frac{N-1}{2}) p(\boldsymbol{\varphi}'_{i,s}) p(I_{i,s}+1)}{\Gamma(\frac{N}{2}) p(\boldsymbol{\varphi}_{i,s})(1-\rho^2) p(I_{i,s})}\right), \quad (7)$$

For updates from $\mathcal{C}_{I_{i,s}+1}$ to $\mathcal{C}_{I_{i,s}}$, we drop the last dimension from $\boldsymbol{\varphi}_{i,s}$ and invert the second argument of the min operation in Equation (7).

Experiments

... to assess whether and what we gain by using a latent feature space. We use:

- synthetic data with and without artefacts
- single trial EEG with emphasis on classification of cognitive state of the brain, i.e. a brain computer interface.
- sleep EEG with emphasis on classification of sleep spindles.

and compare the classification performance of the latent feature space GMOHHM with the performance of the GMOHMM when conditioning on point estimates.

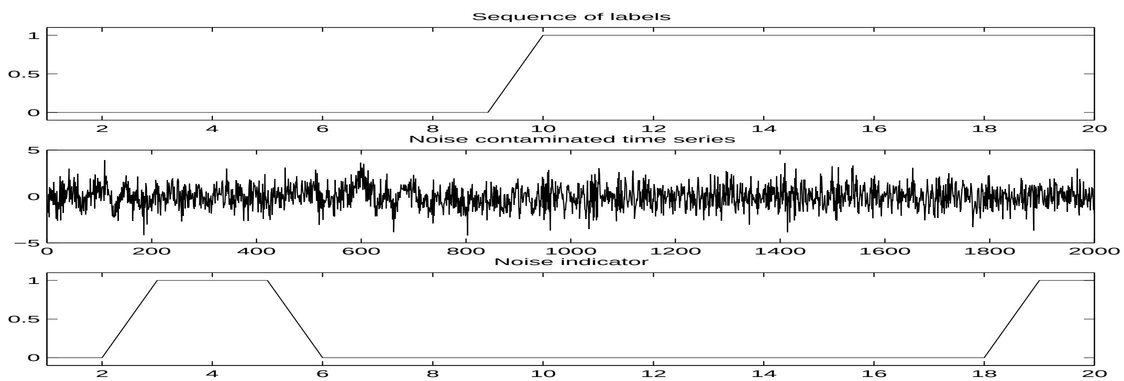
Synthetic data

Generate as target labels a state sequence (200 training, 600 test).

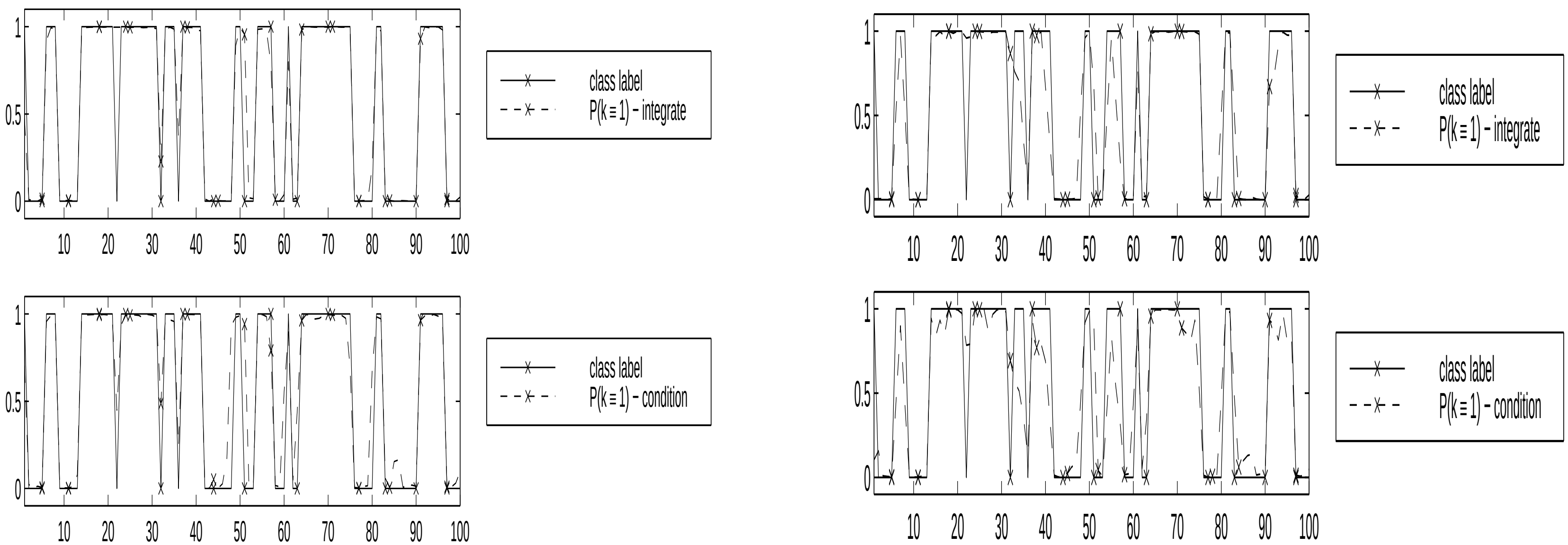
if state $\equiv$ 1: generate data using reflection cfs.: (0.9, -0.8, 0.5)

if state $\equiv$ 2: generate data using reflection cfs.: (0.9, -0.7, 0.6)

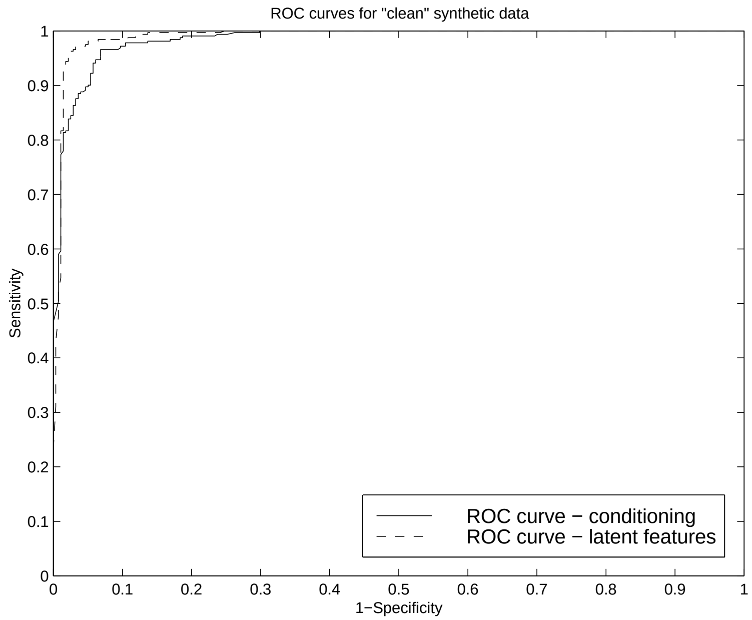
Each segment has 200 samples, generated with noise level $\sigma = 1$. Due to sampling effects we obtain a data set with Bayes error > 0. In order to get a more realistic problem, we use a second state sequence to replace 20% of the segments with white noise.

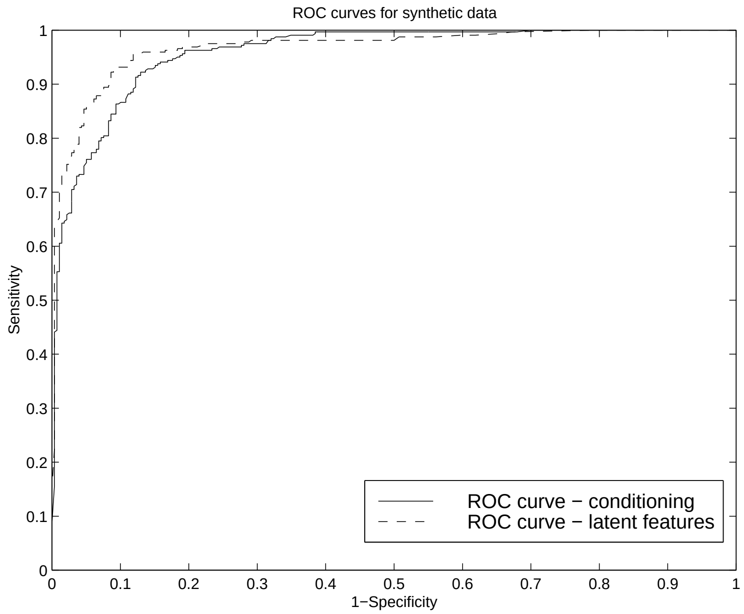


Probabilities on clean and noisy data



ROC curves on clean and noisy data





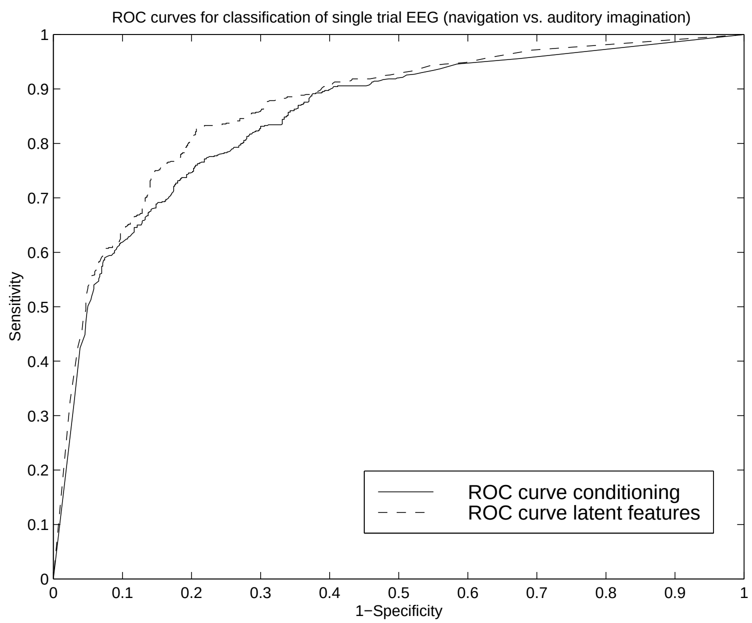
Classifying cognitive tasks (BCI)

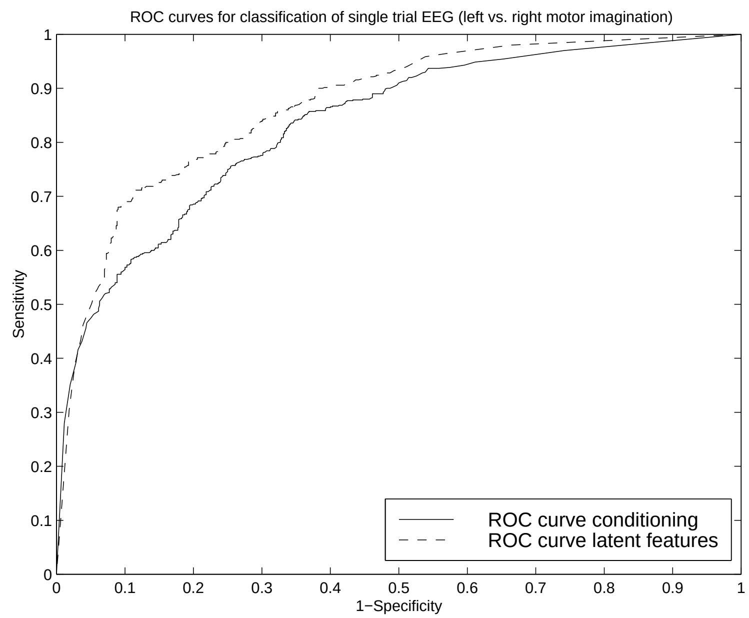
Settings of the cognitive experiment:

- Ten young healthy untrained subjects.
- Two cognitive task pairings: auditory-navigation (A) and left motor-right motor imagination (B).
- Three electrode sites: 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) and ground at left mastoid process.
- Silver-silver chloride electrodes, ISO-DAM system (gain 10^4, filter with pass band between 0.1 Hz and 100 Hz). Sampled with 384 Hz and 12 bit resolution.
- Each cognitive experiment was performed 10 times for 7 seconds.

Settings of the computer experiment:

- Evaluations are done by 10 fold cross validation.
- Realistic performance estimates $->$ classify all half second segments.
- No additional filtering.
- Draw 6000 samples from the posterior and regard the first 1000 samples as burn in.

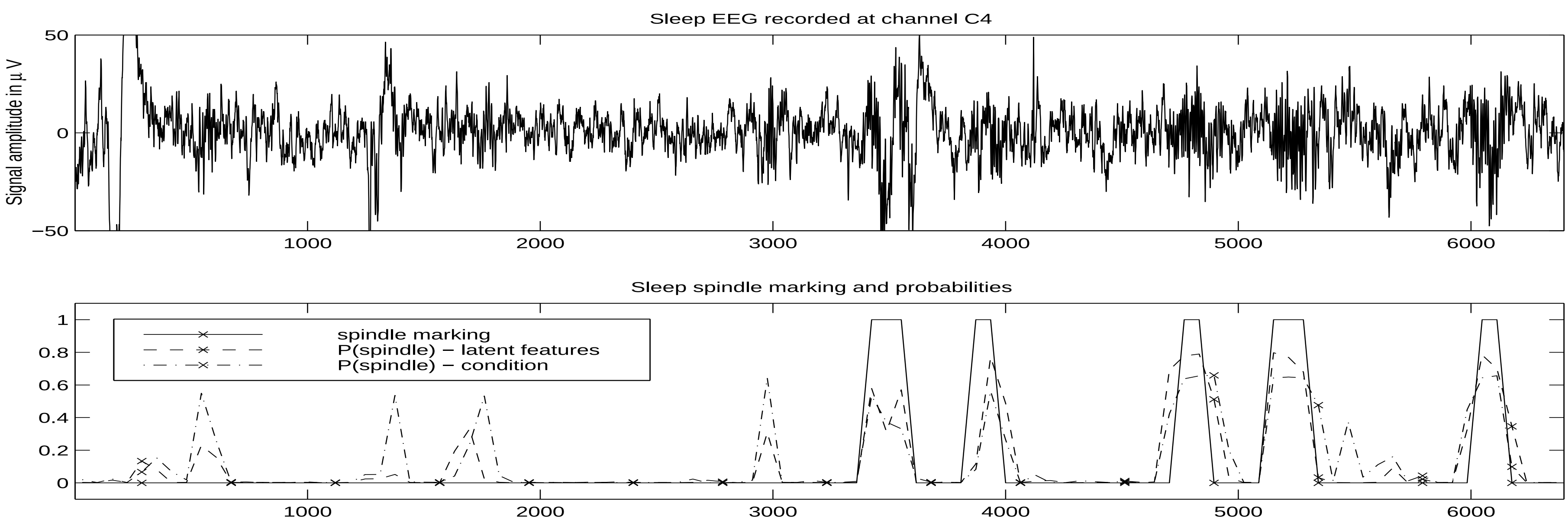




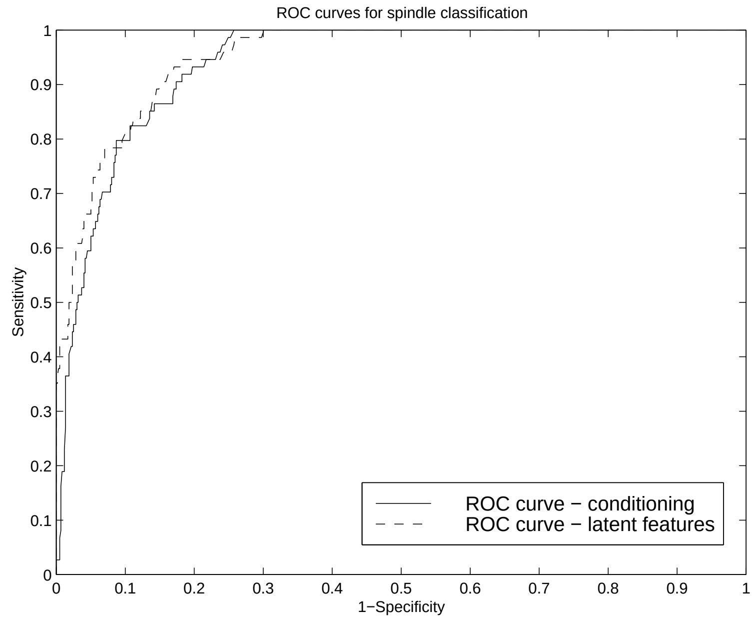
Classification of sleep spindles

Data: two subjects, 7 minutes of EEG each, 3 EEG channels (F4, C4 and P4). sampled at 102.4Hz.

Objective: classify segments (64 samples) for sleep spindles.



... and ROC curve



Summary generalization errors

task	condition	integrate	sig. level
synthetic clean	5.5%	3.3%	$p = 0.02$
synthetic noisy	12.2%	9.8%	$p = 0.02$
left vs. right motor	26%	23%	$p < 0.01$
auditory vs. navigation	24.5%	20%	$p < 0.01$
spindle	8.8%	7.3%	$p = 0.045$

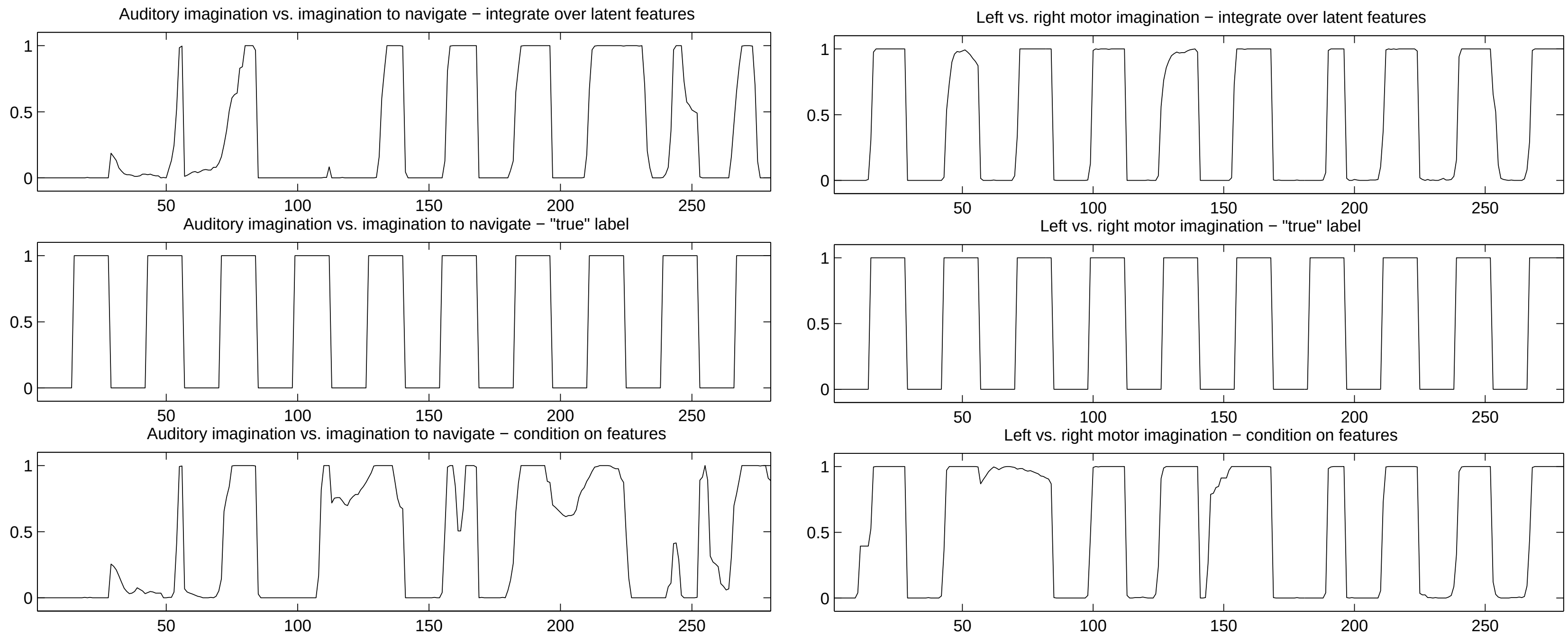
Discussion

- Generalization results confirm that latent feature spaces are not only a Bayesian curiosity.
- Disadvantage is an increased computational complexity. Sweeps remain $O(n)$ (n... number of samples) but latent feature space requires larger number of samples.
- Method not feasible for large problems (e.g. all night sleep analysis) or online application (e.g. BCI)
- Downsizing is an issue! (e.g. mean field methods instead of MCMC).

Related work

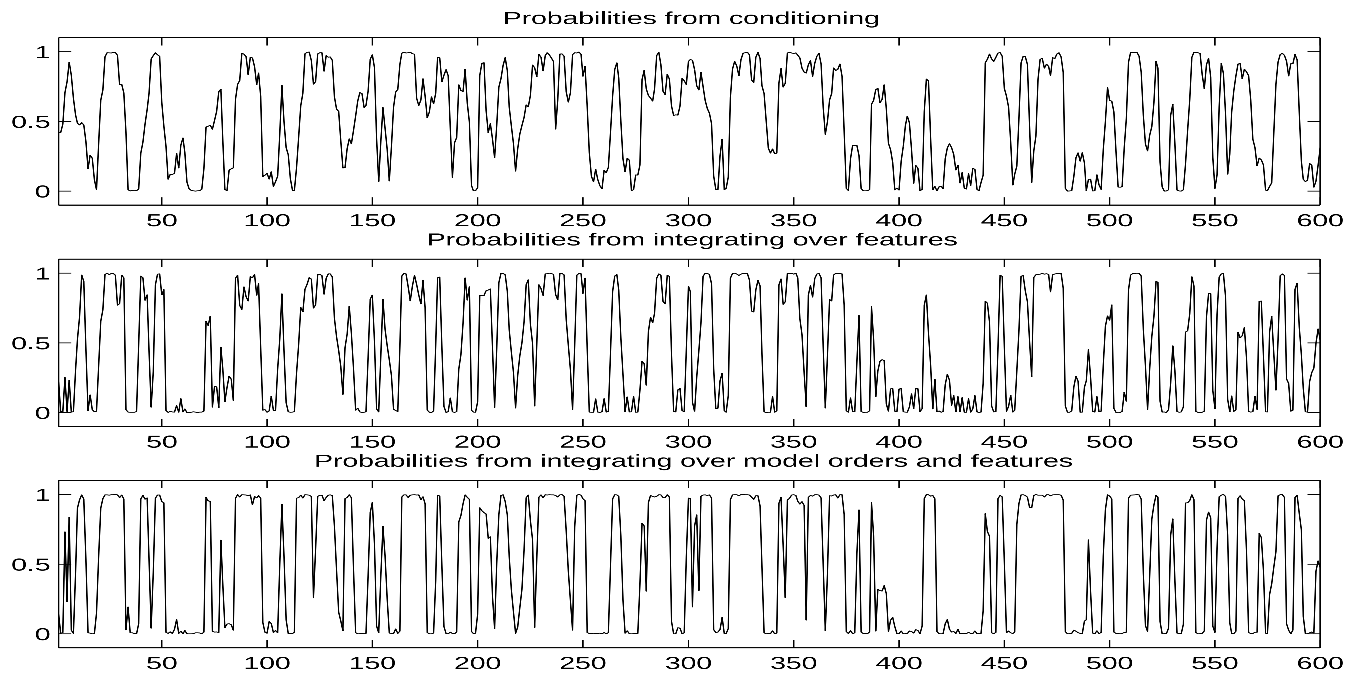
- Dellaportas and Stephens, "Bayesian analysis of errors-in-variables regression models" *Biometrics*, 51:1085-1095, 1995.

- Wright, "Bayesian approach to Neural-Network modeling with input uncertainty" *IEEE Trans. Neural Networks*, 10:1261-1270,1999.



A closer analysis of the probability plots suggests that we usually improve results. However, integration also introduces some mistakes.

Generalization results:



Generalization errors (differences highly significant!):

conditioning	marginalize features	full integration
24.7%	12.8% (1 vs. 2 $p \ll 0.01$)	9.5% (3 vs. 2 $p \ll 0.01$)