Factorial Switching Kalman Filters for Condition Monitoring in Neonatal Intensive Care

Chris Williams

Institute for Adaptive and Neural Computation
School of Informatics, University of Edinburgh, UK

June 2007

Projects

- Neonatal Condition Monitoring
- Prediction with Gaussian Processes
- Visual object class recognition and localization
- Unsupervised learning of multiple objects from images
- Automated detection of spurious objects in astronomical catalogues
- Chorale harmonization (HMM Bach)
- Dynamic trees for image segmentation
- Generative Topographic Mapping (GTM)
- + Outlook

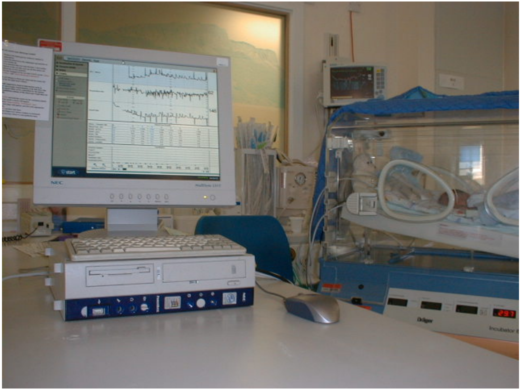
Machine Learning and Probabilistic Modelling

- **Supervised Learning**
  model $p(y|\mathbf{x})$: regression, classification, *etc*
- **Unsupervised Learning**
  model $p(\mathbf{x})$: not just clustering!
- **Reinforcement Learning**
  Markov decision processes, POMDPs, planning.

1. Premature Baby Monitoring

with John Quinn, Neil McIntosh

Why model this data?

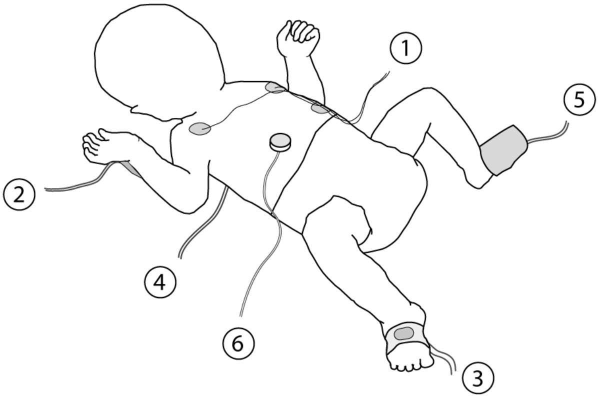


- Artifact corruption, leading to false alarms
- Our aim is to determine the baby's state of health despite these problems

Overview of Baby Monitoring

- Factors
- Factorial switching Kalman filter
- Inference
- Parameter estimation
- Results
- Modelling novel regimes

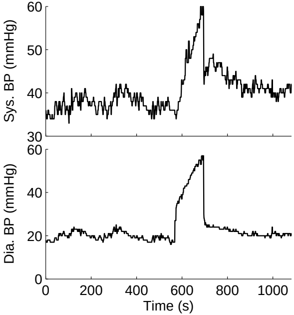
Probes



1. ECG, 2. arterial line, 3. pulse oximeter 4. core temperature, 5. peripheral temperature, 6. transcutaneous probe.

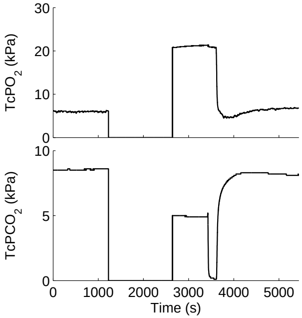
Factors affecting measurements

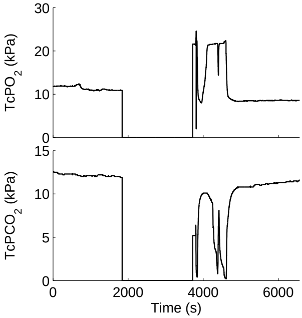
- The physiological **observations** are affected by different **factors**.
- Factors can be artifactual or physiological.
- An arterial blood sample (artifact):



Common factor examples

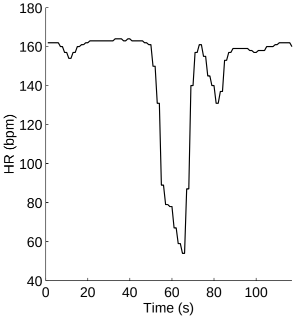
- Transcutaneous probe recalibration (artifact)

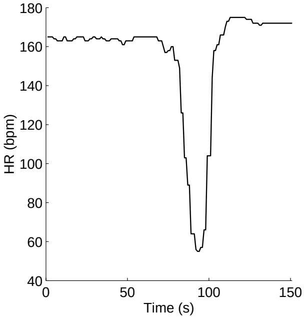




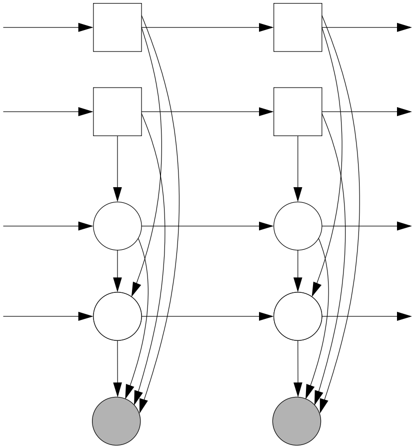
Common factor examples

- Bradycardia (physiological)





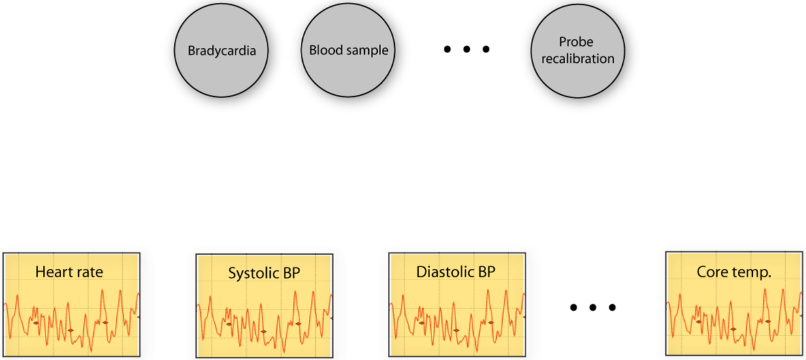
Factorial Switching Kalman Filter



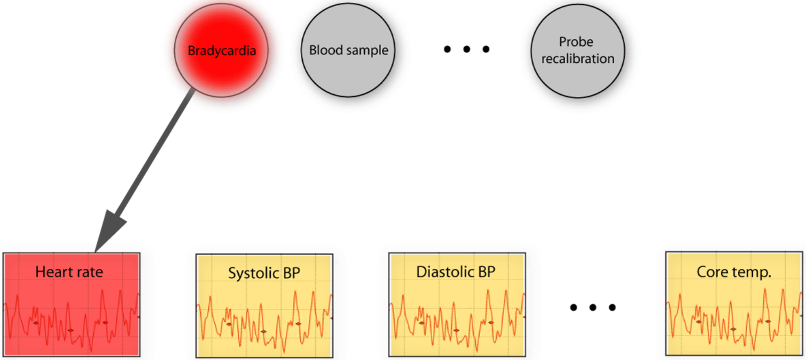
FSKF notation

- s_t is the switch variable, which indexes factor settings, e.g. ‘blood sample occurring **and** first stage of TCP recalibration’.
- $\mathbf{x}_t$ is the hidden continuous state at time t. This contains information on the true physiology of the baby, and on the levels of artifactual processes.
- $\mathbf{y}_{1:t}$ are the observations.

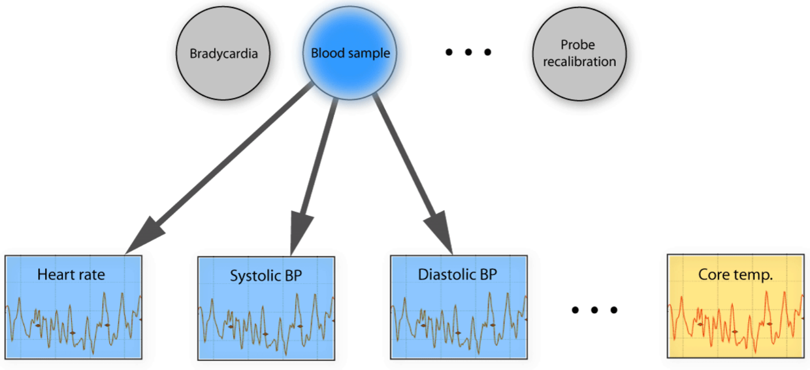
Factor interactions



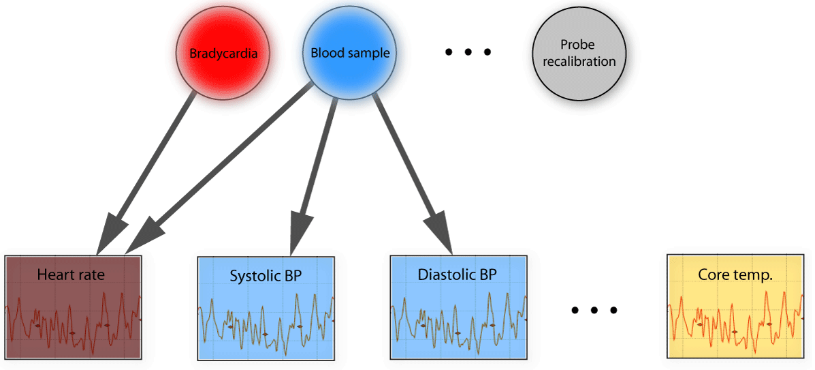
Factor interactions



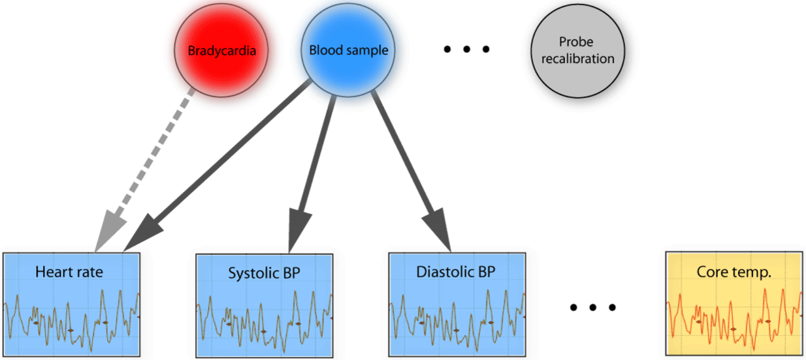
Factor interactions



Factor interactions



Factor interactions



Related work

- Switching linear dynamical models have been studied by many authors, e.g. Alspach and Sorenson (1972), Ghahramani and Hinton (1996).
- Applications include fault detection in mobile robots (de Freitas et al., 2004), speech recognition (Droppo and Acero, 2004), industrial monitoring (Morales-Menedez et al., 2002).
- A two-factor FSKF was used for speech recognition by Ma and Deng (2004). Factorised SKF also used for musical transcription (Cemgil et al., 2006).
- There has been previous work on condition monitoring in the ICU, though we are unaware of any studies that use a FSKF.

Kalman filtering

- Continuous hidden state affects some observations:

$$\begin{aligned} \mathbf{x}_t &\sim \mathcal{N}(\mathbf{A}\mathbf{x}_{t-1}, \mathbf{Q}) \\ \mathbf{y}_t &\sim \mathcal{N}(\mathbf{C}\mathbf{x}_t, \mathbf{R}) \end{aligned}$$

- Kalman filter equations can be used to work compute $p(\mathbf{x}_{1:t}|\mathbf{y}_{1:t})$
- Done iteratively by *predicting* and *updating*

Switching dynamics

- The switch variable s_t selects the dynamics for a particular combination of factor settings:

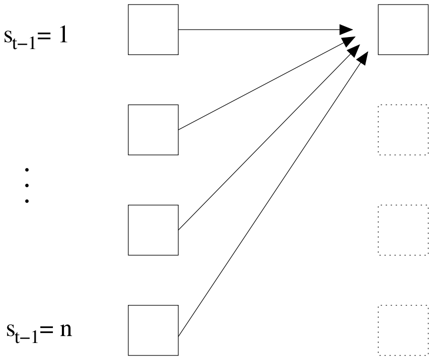
$$\begin{aligned} \mathbf{x}_t &\sim \mathcal{N}(\mathbf{A}^{(s_t)}\mathbf{x}_{t-1}, \mathbf{Q}^{(s_t)}) \\ \mathbf{y}_t &\sim \mathcal{N}(\mathbf{C}^{(s_t)}\mathbf{x}_t, \mathbf{R}^{(s_t)}) \end{aligned}$$

- For each setting of s_t, the Kalman filter equations give a predictive distribution for $\mathbf{x}_t$.

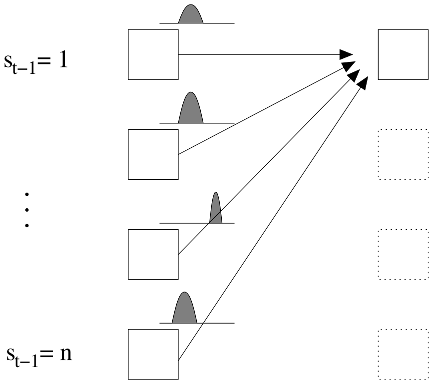
Inference

- For this application, we are interested in filtering, inferring $p(s_t, \mathbf{x}_t | \mathbf{y}_{1:t})$.
- Exact inference is intractable.
- Using two inference methods:
  - Gaussian Sum (Alspach and Sorenson, 1972), analytical approximation
  - Rao-Blackwellised particle filtering.

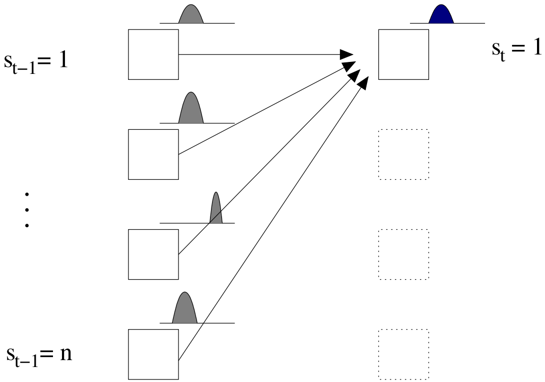
Gaussian Sum approximation



Gaussian Sum approximation



Gaussian Sum approximation

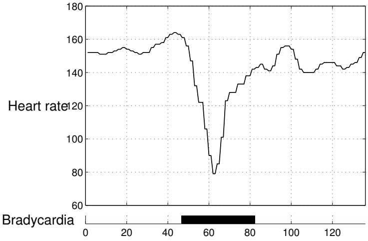


Parameter estimation

- We need to estimate a dynamical model for each continuous state variable for each setting of the factors
- We use AR/ARMA/ARIMA modelling, e.g. an AR(p) process

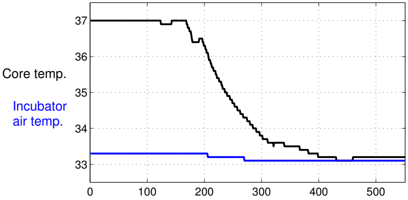
$$x_i(t) = \sum_{j=1}^{p} \alpha_{ij} x_i(t-j) + \epsilon_t$$

- Fortunately, annotated training data is available



- The hidden continuous state in this application is interpretable, and domain knowledge can be used to help parameterize the dynamical models for each factor.

Parameter estimation example



- For example, we know that the falling temperature measurements caused by a probe disconnection will follow an exponential decay
- Therefore we can model these dynamics as an AR(1) process, and set parameters by solving the Yule-Walker equations.

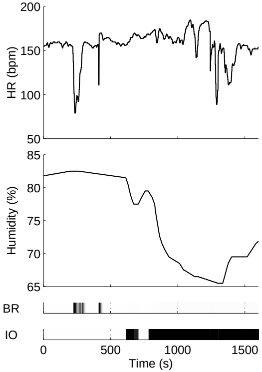
Learning stable physiological dynamics

- Each observation channel has different dynamics when the baby is 'stable' (self regulating) and no artifactual factors are active
- By analysing examples of stable data, dynamical models can be found for each channel with the Box-Jenkins approach and EM.
- For example, a hidden ARIMA(2,1,0) model is a good fit to baseline heart rate data.

Known factor classification demo

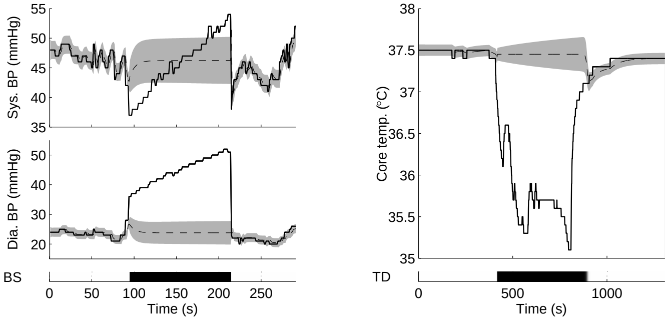
Inference results

- Inference of bradycardia and incubator open factors. Note that heart rate variation while incubator is open is attributed to handling of the baby (BR factor suppressed)



Inference results

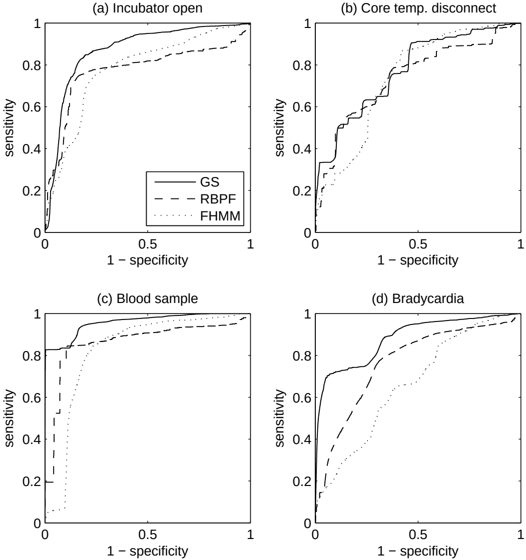
- Can examine variance variance of estimates of true physiology $\hat{\mathbf{x}}_t$, e.g. for blood sample (left) and temperature probe disconnection (right):



Quantitative Evaluation

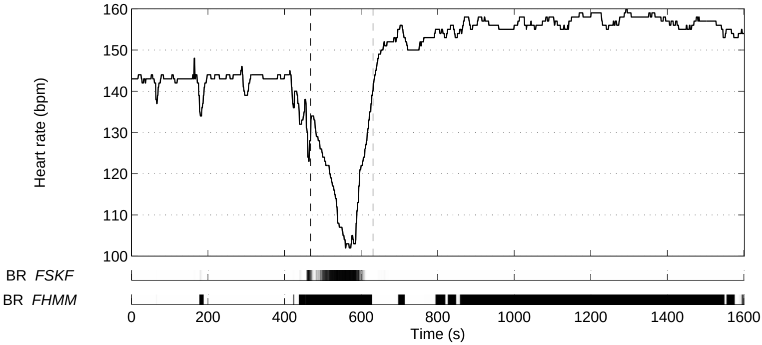
- 3-fold cross validation on 360 hours of monitoring data from 15 babies.
- FHMM has the same factor structure as the FSKF, with no hidden continuous state.

Inference type		Incu. open	Core temp.	Blood sample	Bradycardia
GS	AUC	0.87	0.77	0.96	0.88
	EER	0.17	0.34	0.14	0.25
RBPF	AUC	0.77	0.74	0.86	0.77
	EER	0.23	0.32	0.15	0.28
FHMM	AUC	0.78	0.74	0.82	0.66
	EER	0.25	0.32	0.20	0.37


(a) Incubator open
(b) Core temp. disconnect
(c) Blood sample
(d) Bradycardia
sensitivity
1 − specificity
0
0.2
0.4
0.6
0.8
1
0.5
GS
RBPF
FHMM

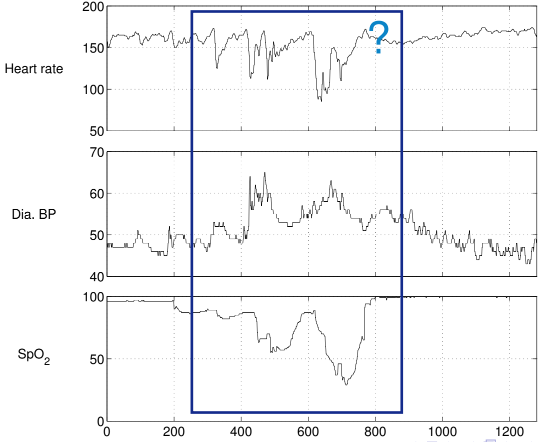
Comparison with FHMM model

- FSKF can handle drift in baseline levels:



Novel dynamics

- There are many other factors influencing the data: drugs, sepsis, neurological problems...



Known Unknowns

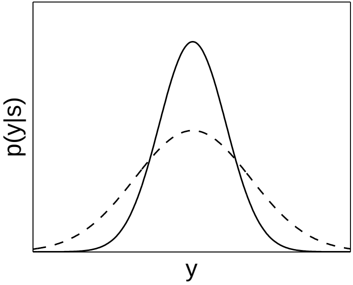
- Add a factor to represent abnormal dynamics

Known Unknowns

- Add a factor to represent abnormal dynamics

X-factor for static 1-D data

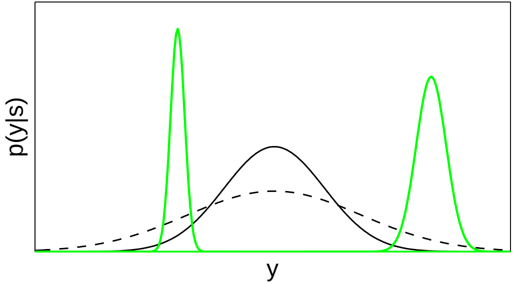
- For static data, we can use a model $\mathcal{M}_*$ representing 'abnormal' data points.



- The high-variance model wins when the data is not well explained by the original model

X-factor with known factors

- The X-factor can be applied to the static data in conjunction with known factors (green):



X-factor for dynamic data

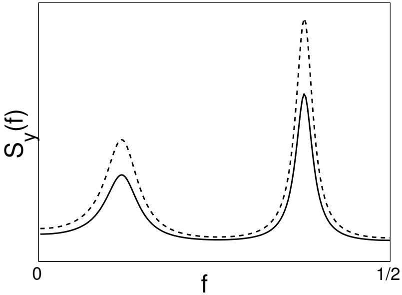
$$
\begin{aligned}
\mathbf{x}_t &\sim \mathcal{N}(\mathbf{A}\mathbf{x}_{t-1}, \mathbf{Q}) \\
\mathbf{y}_t &\sim \mathcal{N}(\mathbf{C}\mathbf{x}_t, \mathbf{R})
\end{aligned}
$$

- Can construct an 'abnormal' dynamic regime analogously:

Normal dynamics: $\{\mathbf{A}, \mathbf{Q}, \mathbf{C}, \mathbf{R}\}$

X-factor dynamics: $\{\mathbf{A}, \xi\mathbf{Q}, \mathbf{C}, \mathbf{R}\}, \quad \xi > 1.$

Spectral view of the X-factor

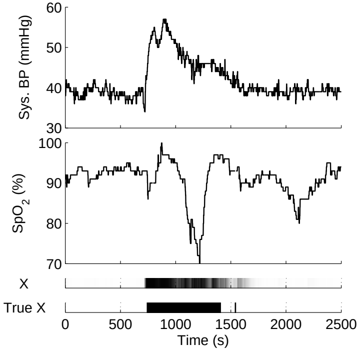


- Plot shows the spectrum of a hidden AR(5) process, and accompanying X-factor
- More power at every frequency
- Dynamical analogue of the static 1-D case

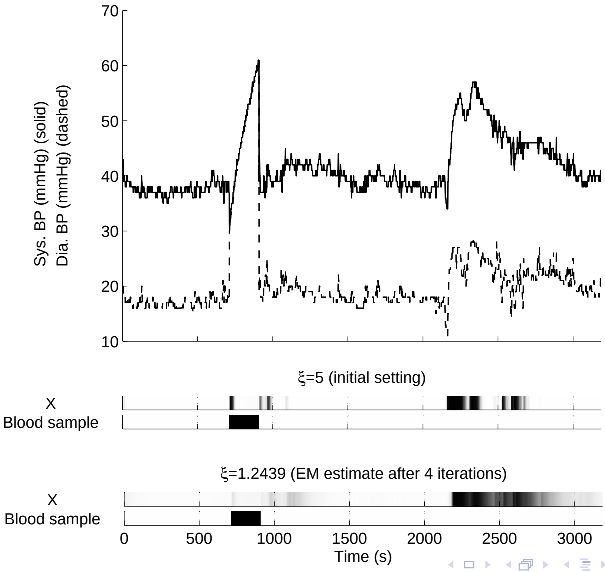
X-factor demo

More inference results

- Classification of periods of clinically significant cardiovascular disturbance:



EM for novel regimes

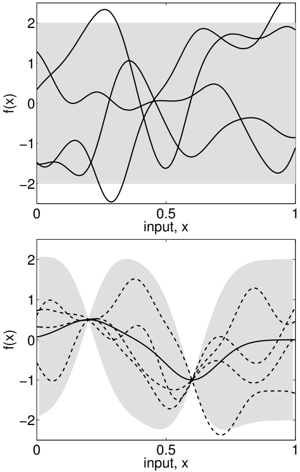


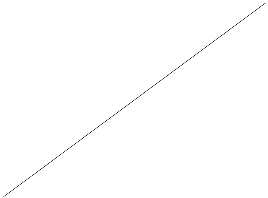
Summary

- FSKF successfully applied to complex physiological monitoring data
- FSKF can be applied more generally to condition monitoring problems
- Interpretable structure
- Knowledge engineering used to parameterize dynamic models
- Allows monitoring of known and novel dynamics (supervised and unsupervised learning)

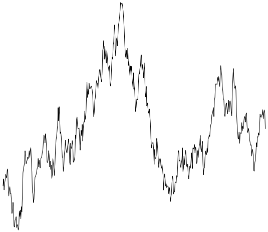
2. Gaussian Processes

- A non-parametric Bayesian prior over functions
- Mean function $\mathbb{E}[f(\mathbf{x})]$, set $= 0$
- Covariance function $\mathbb{E}[f(\mathbf{x})f(\mathbf{x}')] = k(\mathbf{x}, \mathbf{x}')$
- Although GPs are infinite-dimensional objects, prediction from a finite dataset is $O(n^3)$

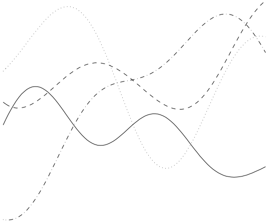




$$k(x, x') = \sigma_0^2 + \sigma_1^2 x x'$$



$$k(x, x') = \exp -|x - x'|$$



$$k(x, x') = \exp -(x - x')^2$$

Gaussian Process Regression

Dataset $\mathcal{D} = (\mathbf{x}_i, y_i)_{i=1}^n$, Gaussian likelihood $p(y_i|f_i) \sim N(0, \sigma^2)$

$$\bar{f}(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}, \mathbf{x}_i)$$

where

$$\boldsymbol{\alpha} = (K + \sigma^2 I)^{-1}\mathbf{y}$$

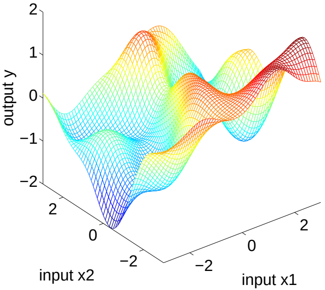
$$\mathrm{var}(f(\mathbf{x})) = k(\mathbf{x}, \mathbf{x}) - \mathbf{k}^T(\mathbf{x})(K + \sigma^2 I)^{-1}\mathbf{k}(\mathbf{x})$$

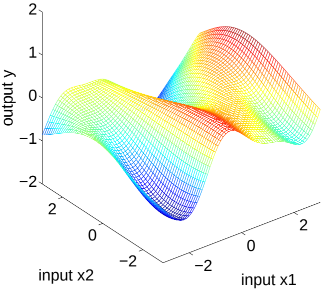
in time $O(n^3)$, with $\mathbf{k}(\mathbf{x}) = (k(\mathbf{x}, \mathbf{x}_1), \ldots, \mathbf{k}(\mathbf{x}, \mathbf{x}_n))^T$

Automatic Relevance Determination

$$k_{SE}(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{x}')^\top M(\mathbf{x}_p - \mathbf{x}_q)\right)$$

- Isotropic $M = \ell^{-2} I$
- ARD: $M = \mathrm{diag}(\ell_1^{-2}, \ell_2^{-2}, \ldots, \ell_D^{-2})$





Dealing with hyperparameters

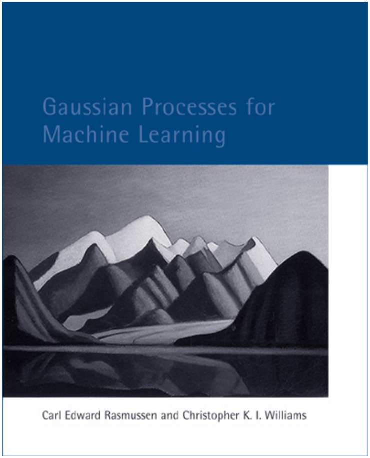
- Marginal likelihood $p(\mathbf{y}|X, \boldsymbol{\theta})$
- For the regression case

$$\log p(\mathbf{y}|X, \boldsymbol{\theta}) = -\frac{1}{2}\mathbf{y}^T(K + \sigma^2 I)^{-1}\mathbf{y} - \frac{1}{2}|K + \sigma^2 I| - \frac{n}{2}\log 2\pi$$

- Optimize by gradient descent (etc) on objective function
- Can also use LOO-CV: $\sum_{i=1}^{n} \log p(y_i|\mathbf{y}_{-i}, X, \boldsymbol{\theta})$
- Note that SVMs do not generally have good methods for kernel selection

- Classification: binary, multiclass, e.g. handwritten digit classification
- SVMs (Vapnik, 1995): non-probabilistic, use “kernel trick” and quadratic programming
- Regularization framework (Tikhonov and Arsenin, 1977; Poggio and Girosi, 1990); MAP rather than fully probabilistic
- Challenges:
  - Design of kernels
  - Approximation methods for large datasets

Carl Edward Rasmussen and Chris Williams, MIT Press, 2006



3. Outlook

Scaling Learning Algorithms towards AI (Bengio and LeCun, 2007; sec. 2)

3. Outlook

Scaling Learning Algorithms towards AI (Bengio and LeCun, 2007; sec. 2)

- The *AI-set*: those tasks involved in intelligent behaviour, e.g. visual perception, auditory perception, planning, control ...

3. Outlook

Scaling Learning Algorithms towards AI (Bengio and LeCun, 2007; sec. 2)

- The *AI-set*: those tasks involved in intelligent behaviour, e.g. visual perception, auditory perception, planning, control ...
- For successful learning we need priors over functions

3. Outlook

Scaling Learning Algorithms towards AI (Bengio and LeCun, 2007; sec. 2)

- The *AI-set*: those tasks involved in intelligent behaviour, e.g. visual perception, auditory perception, planning, control ...
- For successful learning we need priors over functions
- Prior knowledge can be embedded by specifying:
  - Data representation (pre-processing, feature extraction)
  - Architecture of the machine
  - Loss function and regularizer

3. Outlook

Scaling Learning Algorithms towards AI (Bengio and LeCun, 2007; sec. 2)

- The *AI-set*: those tasks involved in intelligent behaviour, e.g. visual perception, auditory perception, planning, control ...
- For successful learning we need priors over functions
- Prior knowledge can be embedded by specifying:
  - Data representation (pre-processing, feature extraction)
  - Architecture of the machine
  - Loss function and regularizer
- Shallow vs Deep architectures

Three strategies

Three strategies

- *Defeatism*: No good parameterization of the AI-set is currently available. Therefore do careful hand-design of pre-processing, architecture and regularizer for each task.

Three strategies

- *Defeatism*: No good parameterization of the AI-set is currently available. Therefore do careful hand-design of pre-processing, architecture and regularizer for each task.
- *Denial*: Kernel machines (or indeed nearest neighbour methods) can approximate any function: why would we need anything else? The issue is that they can *efficiently* represent only a small subset of functions.

Three strategies

- *Defeatism*: No good parameterization of the AI-set is currently available. Therefore do careful hand-design of pre-processing, architecture and regularizer for each task.
- *Denial*: Kernel machines (or indeed nearest neighbour methods) can approximate any function: why would we need anything else? The issue is that they can *efficiently* represent only a small subset of functions.
- *Optimism*: "Let's look for learning models that can be applied to the largest possible subset of the AI-set, while requiring the smallest possible amount of hand-coded knowledge for each specific task in the AI-set."

- Bengio and LeCun emphasize that the “main challenge is to design learning algorithms that can discover representations of the data that compactly describe regularities in it.”
- They argue that such representations will need multiple levels of composition of simpler functions
- Note that learning such representations will be facilitated by multi-task learning
- AI as involving learning, representation and inference

References

- Factorial Switching Kalman Filters for Condition Monitoring in Neonatal Intensive Care.
  Christopher K. I. Williams, John Quinn, Neil McIntosh. In Advances in Neural Information Processing Systems 18, MIT Press (2006)
- Known Unknowns: Novelty Detection in Condition Monitoring. John A. Quinn, Christopher K. I. Williams. Proc 3rd Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA 2007)
- Both available from
  `http://www.dai.ed.ac.uk/homes/ckiw/online_pubs.html`
- Scaling Learning Algorithms towards AI. Y. Bengio and Y. LeCun, to appear in Large Scale Kernel Machines eds. L. Bottou, O. Chapelle, D. DeCoste, J. Weston, MIT Press (2007), see `http://www.iro.umontreal.ca/~bengioy/`

Models for stable physiology

- A fitted model can be verified by comparing real physiological data against a sample from that model, e.g. for heart rate: