Factorial Switching Linear Dynamical Systems applied to Physiological Condition Monitoring

John A. Quinn, Christopher K.I. Williams, Neil McIntosh

Abstract— **Condition monitoring often involves the analysis of systems with hidden factors that switch between different modes of operation in some way. Given a sequence of observations, the task is to infer the filtering distribution of the switch setting at each time step. In this paper we present factorial switching linear dynamical systems as a general framework for handling such problems. We show how domain knowledge and learning can be successfully combined in this framework, and introduce a new factor (the "X-factor") for dealing with unmodelled variation.**

We demonstrate the flexibility of this type of model by applying it to the problem of monitoring the condition of a premature baby receiving intensive care. The state of health of a baby cannot be observed directly, but different underlying factors are associated with particular patterns of physiological measurements and artifacts. We have explicit knowledge of common factors and use the X-factor to model novel patterns which are clinically significant but have unknown cause. Experimental results are given which show the developed methods to be effective on typical intensive care unit monitoring data.

Index Terms— **Condition monitoring, switching linear dynamical system, switching Kalman filter, novelty detection, intensive care.**

I. Introduction

CONDITION monitoring often involves the analysis of systems with hidden factors that "switch" between different modes of operation and collectively determine the observed data. Given just the monitoring data, we are interested in recovering the state of the factors that gave rise to it. In real-world data (from medicine, robotic control or finance, for example) it may be the case that there are a very large number of possible factors, and that we only have explicit knowledge of commonly occurring ones.

We consider the use of factorial switching linear dynamical systems (FSLDS) for this kind of problem. The switch setting is determined by a number of discrete factors. When conditioned on the factor settings, the FSLDS is equivalent to a linear dynamical system (LDS), which has two types of variables that we call *observations* and *state*. The state denotes continuous-valued quantities; we can use this to model the "true" values relating to different aspects of the system being monitored (see below). The observations are those readings obtained from the monitoring equipment, and in general might be subject to corruption by artifact and sensor noise.

Unfactorised switching linear dynamical systems (SLDS) [1], [2] have been used previously in applications such as detecting faults in mobile robots [3], monitoring industrial processes [4], [5], modelling human motion [6], [7], modelling financial series [8], modelling creatinine levels in patients with kidney transplants [9] and speech recognition [10]. Factorial switching linear dynamical systems have also been used recently for speech recognition [11] and musical transcription [12]. Their use was also reported in our conference papers [13], [14], which this paper extends.

A common feature of previous FSLDS work is a lack of diversity in the factors. Some of this work has a small number of factors; in [11] there are two, one modelling vocal-tract resonance and one modelling measurable acoustics. Factors can also be higher in number but similar in nature, as in [12] where each represents a different note in a polyphonic transcription. All previous work additionally assumes that there are a fixed number of factors which are known in advance. In this paper we build on the success of previous (F)SLDS applications by focusing on cases where the factors may be numerous and of a diverse nature. We also consider the possibility that we do not have explicit knowledge of all factors governing the data.

We demonstrate the flexibility of the FSLDS model in this context by applying it to the problem of monitoring the condition of a premature baby receiving intensive care. The state of health of a baby cannot be observed directly, but different states of health are associated with particular patterns of measurements, e.g. in the heart rate, blood pressure and temperature. In this case the factors can be both physiological (such as a spontaneous slowing of the heart) or artifactual (such as a probe disconnection), and are potentially so numerous that it would be impractical to explicitly model them all. We exploit known structure between the factors and observation channels, e.g. so that only a subset of factors influence a given channel.

The main contributions of this paper are as follows:

- We show that it is often impractical to model all possible factors affecting the observations. To deal with this situation we introduce an "X-factor" to handle unmodelled variation.
- We demonstrate how to exploit knowledge of the structure of how the various latent factors interact so as to reduce the amount of training data needed for the system. A combination of domain knowledge engineering and learning is used to produce an effective solution.
- We demonstrate that the FSLDS framework can be applied effectively to the important real-world problem of neonatal condition monitoring.

We describe the model in section II, discussing learning, verification, and relative merits compared to alternative models for condition monitoring. In section III we consider the case where monitoring data is influenced by factors we do not know about in advance, possibly in conjunction with other factors which we do know about. Inference in the model is discussed in section IV. In section V we show how the FSLDS can be applied to the neonatal condition monitoring. We give experimental results for this application in section VI, and draw conclusions in section VII. Demonstration code is available [15].

Author JQ is with the Department of Computer Science, Makerere University, Uganda. Author CW is with the School of Informatics, University of Edinburgh, UK, and author NM is with the Department of Child Life and Health, University of Edinburgh, UK. *Emails: jquinn@cit.mak.ac.ug, [c.k.i.williams, neil.mcintosh]@ed.ac.uk.*

II. Model description

We first review the SLDS before generalising to the factorial case. In such models, the hidden switch setting s_t affects the hidden continuous state $\mathbf{x}_t$ and the observations $\mathbf{y}_t$. Conditional on a particular switch setting, the model is equivalent to a linear Gaussian state-space model (Kalman filter). The switch setting evolves according to the transition probabilities $p(s_t|s_{t-1})$, and for a given setting of s_t the hidden continuous state and the observations are related by:

$$\mathbf{x}_t \sim \mathcal{N}\left(\mathbf{A}^{(s_t)}\mathbf{x}_{t-1} + \mathbf{d}^{(s_t)}, \mathbf{Q}^{(s_t)}\right) \tag{1}$$

$$\mathbf{y}_t \sim \mathcal{N}\left(\mathbf{C}^{(s_t)}\mathbf{x}_t, \mathbf{R}^{(s_t)}\right) \tag{2}$$

where $\mathbf{x} \in \mathbb{R}^{d_x}$ and $\mathbf{y} \in \mathbb{R}^{d_y}$. Here $\mathbf{A}^{(s_t)}$ is a square system matrix, $\mathbf{d}^{(s_t)}$ is a drift vector, $\mathbf{C}^{(s_t)}$ is the state-observations matrix, and $\mathbf{Q}^{(s_t)}$ and $\mathbf{R}^{(s_t)}$ are noise covariance matrices. Note that in this formulation, all dynamical parameters can be switched between regimes. Similar models referred to in the above literature sometimes switch only the state dynamics $\{\mathbf{A}, \mathbf{Q}\}$, or the observation dynamics $\{\mathbf{C}, \mathbf{R}\}$.

It is possible to factorise the switch variable, so that M factors $f_t^{(1)} \dots f_t^{(M)}$ affect the observations $\mathbf{y}_t$. The factor $f^{(m)}$ can take on $L^{(m)}$ different values. The state space is the cross product of the factor variables,

$$s_t = f_t^{(1)} \otimes \dots \otimes f_t^{(M)} \tag{3}$$

with $K = \prod_{m=1}^{M} L^{(m)}$ being the number of settings that s_t can take on. The value of $f_t^{(m)}$ depends on $f_{t-1}^{(m)}$. The factors are a priori independent, so that

$$p(s_t|s_{t-1}) = \prod_{m=1}^{M} p\left(f_t^{(m)}|f_{t-1}^{(m)}\right) \ . \tag{4}$$

Notice that the factors are not, in general, a posteriori independent. The joint distribution of the model is

$$p(s_{1:T}, \mathbf{x}_{1:T}, \mathbf{y}_{1:T}) = p(s_1)p(\mathbf{x}_1)p(\mathbf{y}_1|\mathbf{x}_1, s_1) \ . \\ \prod_{t=2}^{T} p(s_t|s_{t-1})p(\mathbf{x}_t|\mathbf{x}_{t-1}, s_t)p(\mathbf{y}_t|\mathbf{x}_t, s_t) \tag{5}$$

where $s_{1:T}$ denotes the sequence $s_1, s_2, \dots, s_T$ and similarly for $\mathbf{x}_{1:T}$ and $\mathbf{y}_{1:T}$. $p(\mathbf{x}_t|\mathbf{x}_{t-1}, s_t)$ is defined in eq (1), $p(\mathbf{y}_t|\mathbf{x}_t, s_t)$ in eq (2) and $p(s_t|s_{t-1})$ in eq (4). By considering the factored nature of the switch setting, we have an observation term of the form $p(\mathbf{y}_t|\mathbf{x}_t, f_t^{(1)}, \dots, f_t^{(M)})$. This can be parameterised in different ways. In this work, we specify conditional independencies between particular components of the observation $\mathbf{y}_t$ given the factor settings. This is explained further in sections II-B and V-E. Although we make use of prior factored dynamics in eq (4) in this work, it is very simple to generalize the model so that this no longer holds. The inference algorithms described in section IV can still be applied. However, the separate factors are crucial in structuring the system dynamics and observations model.

A. Learning

In a condition monitoring problem, it is assumed that we are able to interpret at least some of the regimes in the data; otherwise we would be less likely to have an interest in monitoring them. We can therefore usually expect to obtain some labelled training data $\{\mathbf{y}_{1:T}, s_{1:T}\}$. When available, this data greatly simplifies the learning process, because determining the switch setting in the (F)SLDS makes the model equivalent to a linear dynamical system, therefore making the process of parameter estimation a standard system identification problem.

Given training data with known switch settings, the learning process is therefore broken down into the training of a set of LDS models—one per switch setting. We might choose a particular parameterisation, such as an autoregressive (AR) model of order p hidden by observation noise and fit parameters accordingly [16]. Expectation maximisation can be useful in this setting to improve parameter settings given an initialisation [17]. We describe particular methods used for parameter estimation in the physiological monitoring application in section V which incorporate both of these ideas. Note that if labellings for the training data were not available, it would still be possible to learn the full switching model directly using EM [11] or variational learning [18].

When labelled training data is available, estimates of the factor transition probabilities are given by

$$P(f_t^{(m)} = j|f_{t-1}^{(m)} = i) = \frac{n_{ij} + \zeta}{\sum_{k=1}^{M} n_{ik} + \zeta} \ , \tag{6}$$

where n_{ij} is the number of transitions from factor setting i to setting j in the training data. The constant terms ζ (set to $\zeta = 1$ in the experiments described later in the paper) are added to stop any of the transition probabilities being zero or very small.

Some verification of the learned model is possible by clamping the switch setting to a certain value and studying the resulting LDS. One simple but effective test is to draw a sample sequence and check by eye whether it resembles the dynamics of training data which is known to follow the same regime. Some insight into the quality of the parameter settings can also be gained by considering estimation of the hidden state $\mathbf{x}$ in the LDS. The Kalman filter equations yield both an *innovation sequence*, $\tilde{\mathbf{y}}_{1:T}$ (the difference between the predicted and actual observations), and a specification of the covariance of the innovations under ideal conditions. An illuminating test is therefore to compare the actual and ideal properties of the innovation sequence when applied to training data. In particular, the innovations $\tilde{\mathbf{y}}_t$ should come from a Gaussian distribution with zero mean and a specific covariance, and should be uncorrelated in time. We find in practice that such tests are highly significant when training (F)SLDS models for condition monitoring. For more details about verification in linear dynamical systems, see [19, §5.5].

B. Learning the factorial model

The previous discussion assumes that we train the model conditioned on each switch setting independently, and then combine parameters. Where there are many factors this implies a great quantity of training data is needed. In practice, however, this requirement can be mitigated.

Where there are several measurement channels it may be found that some factors "overwrite" others. For example, if we are monitoring the physiological condition of a patient, we might have two factors: *heart problem* and *probe disconnection*. If there is a heart problem and the probe is disconnected, then we would see the same measurements as though only the probe was disconnected (that is, a sequence of zeros). It is often possible to specify an ordering of factors such that some overwrite measurement channels of others in this way. The significance of this is that examples of every combination of factors do not need to be found

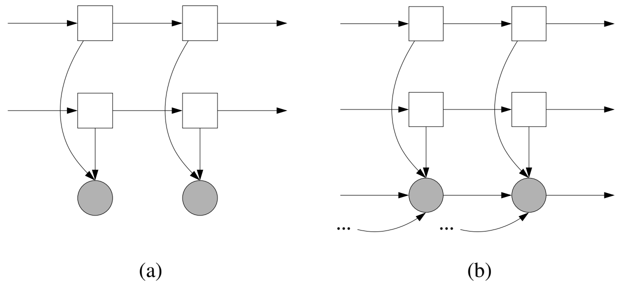


Fig. 1. Graphical representations of different factorial models, with $M = 2$ factors. Squares are discrete values, circles are continuous and shaded nodes are observed. (a) The Factorial HMM, (b) the Factorial AR-HMM, in which each observation depends on previous values.

in order to train the factorial model. The factors can be trained independently, and then combined together by reasoning about which channels are overwritten for each combination. Details for the physiological monitoring application are given in section V-E.

C. Comparison with other switching models for condition monitoring

We have assumed the existence of a discrete switch variable which indexes different modes of operation. In our formulation, the problem of condition monitoring is essentially to infer the value of this switch variable over time from new data. We are particularly interested in the class of models in which there are first-order Markovian transitions between the switch settings at consecutive time steps. Given the switch setting is is possible to characterise the different dynamic regimes on other ways, yielding alternative models for condition monitoring. In this section, we first review the hidden Markov model (HMM) and autoregressive hidden Markov model (AR-HMM), and then discuss their advantages and disadvantages for condition monitoring with respect to the (F)SLDS.

A simple model for a single regime is the Gaussian distribution on $\mathbf{y}_t$. When this is conditioned on a discrete, first-order Markovian switching variable, we obtain an instance of a HMM. This model can therefore be used for condition monitoring when the levels and variability of different measurement channels are significant (though note that in general the HMM can use any reasonable distribution on $\mathbf{y}_t$).

Autoregressive (AR) models are a common choice for modelling stationary time series. Conditioning an AR model on a Markovian switching variable we obtain an autoregressive hidden Markov model (AR-HMM), also known as a switching AR model—see e.g. [20]. This provides a model for conditions in which observations might be expected to oscillate or decay, for example. During inference, the model can only confidently switch into a regime if the last p observations have been generated under that regime; there will be a loss of accuracy if any of the measurement channels have dropped out in that period, for example, or another artifactual process has affected any of the readings.

The general condition monitoring problem involves independent factors which affect a system. In both of these models the switch variable can be factorised, giving the factorial HMM [21] and the factorial AR-HMM respectively. The graphical models for these two constructions are shown in Figure 1.

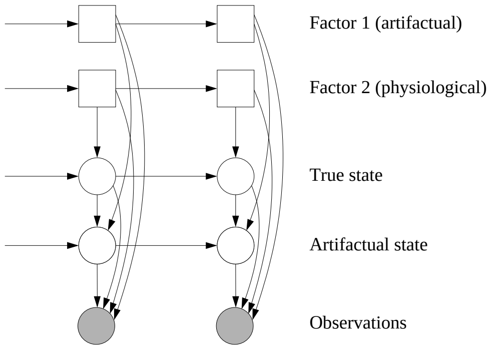


Fig. 2. Factorial switching linear dynamical system for physiological condition monitoring, with $M = 2$ factors as an example. The state is split up into two sets of variables, containing estimates of the 'true' physiology and of the levels of artifactual processes.

By characterising each regime as a linear Gaussian state-space model we obtain the (F)SLDS. The SLDS can be thought of as a "hybrid" model, having both discrete switch settings as in the HMM and continuous hidden state as in a linear dynamical system. The FSLDS is similar, though with the discrete switch setting structure of the factorial HMM. Note, however, that observations in the FHMM [21] are generated through an additive process in which each factor makes a contribution. The mechanisms used to generate observations under different factor settings can in general be more complex and nonlinear than this, as in the the overwriting mechanism explained in section II-B.

(F)SLDS models have a number of representational advantages for condition monitoring. First, we can have many dimensions of hidden state for each observed dimension. This allows us to deal with situations in which different elements affect the observations. For example, consider again the case where some observations are corrupted by artifact, e.g. where there is a fault with the monitoring equipment and measurements temporarily drop out to zero. With extra dimensions in the hidden state, we have the potential to keep track of how the "real" signal might be evolving. While the physiology is unobserved in this way, the discrete switch settings can evolve according to prior dynamics—the most desirable strategy when there is no evidence.

In the physiological monitoring case, for example, we can construct detailed representations of the causes underlying observations. For instance, the state can be split into two groups of continuous latent variables, those representing the "true" physiology and those representing the levels associated with different artifactual processes. Similarly, factors can be physiological or artifactual processes. Physiological factors can affect any state variable, whereas artifactual processes affect only artifactual state. This formulation of the model for physiological condition monitoring is illustrated in Figure 2. More specific details of the model structure in this application are given in section V.

The (F)SLDS also gives us the ability to represent different sources of uncertainty in the system. We can explicitly specify the intra-class variability in the dynamics using the parameter $\mathbf{Q}$ and the measurement noise using the parameter $\mathbf{R}$. There is no way to make this distinction in either of the other models, which have only one noise term per regime.

The application-specific details in section V provide further examples of how this flexibility can be utilised in practise. However, this flexibility in the FSLDS is obtained at the cost of greater complexity, particularly in terms of computing inferences, as we examine in section IV.

III. Novel conditions

So far we have assumed that the monitoring data contains a limited number of regimes, for which labelled training data is available. In real-world monitoring applications, however, there is often such a great number of potential dynamical regimes that it might be impractical to model them all, or we might never have comprehensive knowledge of them. It can therefore be useful to include a factor in the condition monitoring model which represents all "unusual cases".

In this section we present a method for modelling previously unseen dynamics as an extra factor in the model, referred to as the "X-factor". This represents all dynamics which are not normal and which also do not correspond to any of the known regimes. A sequence of data can only be said to have novelty relative to some reference, so the model is learnt taking into account the parameters of the normal regime. The inclusion of this factor in the model has two potential benefits. First, it is useful to know when novel regimes are encountered, e.g. in order to raise an alarm. Second, the X-factor provides a measure of confidence for the system. That is, when a regime is confidently classified as "none of the above", we know that there is some structure in the data which is lacking in the model.

A. The X-factor

First consider a case in which we have independent, one-dimensional observations which normally follow a Gaussian distribution. If we expect that there will also occasionally be spurious observations which come from a different distribution, then a natural way to model them is by using a wider Gaussian with the same mean. Observations close to the mean retain a high likelihood under the original Gaussian distribution, while outliers are claimed by the new model.

The same principle can be applied when there are a number of known distributions, so that the model is conditionally Gaussian, $\mathbf{y}|s \sim \mathcal{N}\left(\mu^{(s)}, \Sigma^{(s)}\right)$. For condition monitoring we are interested in problems where we assume that the possible settings of s represent a "normal" mode and a number of known additional modes. We assume here that the normal regime is indexed by $s = 1$, and the additional known modes by $s = 2, \ldots, K$. In this static case, we can construct a new model, indexed by $s = *$, for unexpected data points by inflating the covariance of the normal mode, so that

$$\Sigma^{(*)} = \xi\Sigma^{(1)}, \qquad \mu^{(*)} = \mu^{(1)} , \tag{7}$$

where normally $\xi > 1$. We refer to this type of construction for unexpected observations as an "X-factor". The parameter ξ determines how far outside the normal range new data points have to fall before they are considered "not normal".

The likelihood functions for a normal class and a corresponding X-factor are shown in Figure 3(a). Clearly, data points that are far away from the normal range are more likely to be classified as belonging to the X-factor. For condition monitoring this can be used in conjunction with a number of known classes, as shown in 3(b). Here, the X-factor has the highest likelihood for regions which are far away from any known modes, as well as far away from normality.

We can generalise this approach to dynamic novelty detection by adding a new factor to a trained factorial switching linear dynamical model, by inflating the system noise covariance of the normal dynamics

$$\begin{aligned} \mathbf{Q}^{(*)} &= \xi\mathbf{Q}^{(1)} , && (8) \\ \left\{\mathbf{A}^{(*)}, \mathbf{C}^{(*)}, \mathbf{R}^{(*)}, \mathbf{d}^{(*)}\right\} &= \left\{\mathbf{A}^{(1)}, \mathbf{C}^{(1)}, \mathbf{R}^{(1)}, \mathbf{d}^{(1)}\right\} && (9) \end{aligned}$$

In the LDS, any sequence of $\mathbf{x}$'s is jointly Gaussian. Consider the case where the state is a scalar variable; the eigenfunctions are sinusoids and the eigenvalues are given by the power spectrum. Increasing the system noise has the effect of increasing the power at all frequencies in the state sequence (see for example Figure 3(c)). Hence we have a dynamical analogue of the static construction given above.

A similar model for changes in dynamics is mentioned by West and Harrison [22, p. 458 and §12.4], who suggest it as the parameterisation of an extra state in the unfactorised SLDS for modelling large jumps in the $\mathbf{x}$-process, and suggest setting $\xi = 100$. Their analysis in §12.4.4 shows that this is used to model single-time-step level changes, and not (as we are doing) sustained periods of abnormality. We find a much smaller value $\xi = 1.2$ to be effective for our task (larger values of ξ mean that an observation sequence must deviate further from normal dynamics to be claimed by the X-factor). A different generative model for the X-factor in principle would be white noise, but we find in practice that this model is too dissimilar to the real signal and is not effective.

Note that the nature of the measurement noise, and hence the value of the parameter $\mathbf{R}^{(s)}$, is assumed to be the same for both the normal regime and for the X-factor. Care needs to be taken that the known factor dynamics do not have a very high variance compared to the normal dynamics. It is clear from Figure 3(b) that the X-factor will not be effective if any of the factors are wider than normality. This can be ascertained by examining the spectra of the different model dynamics.

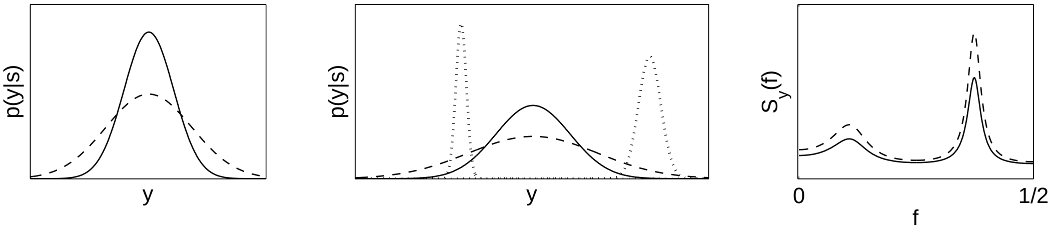
B. Interaction with other factors

It was described in section II-B how factors in a factorial model overwrite different dimensions in the hidden state. As the X-factor operates on every state dimension, there are two possibilities for combining it with other known factors: either it can overwrite everything, or it can be overwritten by everything (except normality). For this application the latter approach is more sensible, so that for example if there is a period of unusual dynamics and an ECG probe dropout then the dropout dynamics generate the heart rate observations and the X-factor generates all other channels.

C. Learning

Unlike the factors for which we have an interpretation, we do not assume that labelled training data is available for learning X-factor dynamics. We therefore consider a partial labelling of the training data $\mathbf{y}_{1:T}$, comprising of annotations for known factors and for some representative quantity of normal dynamics. The remainder of the training data is unlabelled, giving us a semi-supervised learning problem.

To apply the expectation-maximisation algorithm to the X-factor within a SLDS (non-factorised switch setting), the M-step



(a) (b) (c)

Fig. 3. (a) Class conditional likelihoods in a static 1D model, for the normal class (solid) and the X-factor (dashed). (b) Likelihoods of the normal class and X-factor in conjunction with other known, abnormal regimes (shown dotted). (c) The power spectral density of a latent AR(5) process with white observation noise (solid), and that of a corresponding X-factor process (dashed).

update to ξ is given by

$$\tilde{\xi} = \frac{1}{\sum_{t=2}^{T} p(s_t = *|\mathbf{y}, \theta_{\text{old}})} \cdot \sum_{t=2}^{T} (\hat{\mathbf{x}}_t - \mathbf{A}^{(1)}\hat{\mathbf{x}}_{t-1})^\top {\mathbf{Q}^{(1)}}^{-1} (\hat{\mathbf{x}}_t - \mathbf{A}^{(1)}\hat{\mathbf{x}}_{t-1}) p(s_t = *|\mathbf{y}, \theta_{\text{old}}) \tag{10}$$

where $s_t = *$ indexes the X-factor switch setting at time t and $\hat{x}_t$ is the mean of the inferred state distribution (we describe strategies for calculating this in section IV). The parameters $\mathbf{A}^{(1)}$ and $\mathbf{Q}^{(1)}$ are the system matrix and system noise covariance matrix respectively for the normal dynamical regime. Intuitively, this update expression calculates a Z-score, considering the covariance of novel points and the covariance of the normal regime. Every point is considered, and is weighted by the probability of having been generated by the X-factor regime. Note that (10) does not explicitly constrain $\tilde{\xi}$ to be greater than 1, but with appropriate initialisation it is unlikely to violate this condition.

The factorial case is a little more complicated due to the possibility that different combinations of factors can overwrite different channels. For example, if a bradycardia is occurring in conjunction with some other, unknown regime, then the heart rate dynamics are already well explained and should not be taken into account when re-estimating the X-factor parameter ξ.

A derivation of (10) and an extension to the factorial case is given in [23, §C.4].

D. Relation to work in novelty detection

There is a large body of work on statistical approaches to novelty detection, reviewed in [24]. In general the goal is to learn the density of training data and to raise an alarm for new data points which fall in low density areas. In a time-series context this involves modelling the next observation $p(\mathbf{y}_{t+1}|\mathbf{y}_{1:t})$ based on the earlier observations, and detecting observations that have low probability. This method is used, for example, by Ma and Perkins [25]. Such approaches define a model of normality, and look for deviations from it, e.g. by setting a threshold.

A somewhat different take is to define a broad 'outlier' distribution as well as normality, and carry out probabilistic inference to assign patterns to the normal or outlier components. For time-series data this approach was followed by Smyth [26], who considered the use of an unknown state when using a HMM for condition monitoring. This uses a similar idea to ours but in a simpler context, as in his work there is no factorial state structure and no explicit temporal model.

IV. INFERENCE

In this application we are interested in filtering, but the time taken to calculate the exact filtering distribution $p(s_t, \mathbf{x}_t|\mathbf{y}_{1:t})$ in the switching linear Gaussian state-space model scales exponentially with t, making it intractable. This is because the probabilities of having moved between every possible combination of switch settings in times $t-1$ and t are needed to calculate the posterior at time t. Hence the number of Gaussians needed to represent the posterior exactly at each time step increases by a factor of K, the number of cross-product switch settings. The intractability of inference in this model is rigorously demonstrated in [27], which also concentrates on a fault diagnosis setting.

Various approximation schemes are possible to make inference tractable, and we concentrate on two: use of a Gaussian sum approximation, and Rao-Blackwellised particle filtering.

A Gaussian Sum approximation [1] can be used to reduce the time required for inference. At each time step we maintain an approximation of $p(\mathbf{x}_t|s_t, \mathbf{y}_{1:t})$ as a mixture of K Gaussians. Calculating the Kalman updates and likelihoods for every possible setting of s_{t+1} will result in the posterior $p(\mathbf{x}_{t+1}|s_{t+1}, \mathbf{y}_{1:t+1})$ having K^2 mixture components, which can be collapsed back into K components by matching means and variances of the distribution for each setting of s_t, as described in [28].

Rao-Blackwellised particle filtering (RBPF) [29] is another technique for approximate inference, which exploits the conditionally linear dynamical structure of the model to try to select particles close to the modes of the true filtering distribution. A number of particles are propagated through each time step, each with a switch state s_t and an estimate of the mean and variance of $\mathbf{x}_t$. A value for the switch state s_{t+1} is obtained for each particle by sampling from the transition probabilities, after which Kalman updates are performed and a likelihood value can be calculated. Based on this likelihood, particles can be either discarded or multiplied. Because Kalman updates are not calculated for every possible setting of s_{t+1}, this method can give a significant increase in speed when there are many factors. The fewer particles used, the greater the trade-off of speed against accuracy, as it becomes less likely that the particles can collectively track all modes of the true posterior distribution. RBPF has been shown to be successful in condition monitoring problems with switching linear dynamics, for example in fault detection in mobile robots [3].

In condition monitoring we sometimes want to treat zero measurements specially, as missing values. An obvious way to do this is to have a set of dropout factors, one for each measurement channel, which have zeros in the observation matrix $\mathbf{C}$ to indicate the quantity not being observed. We can effectively calculate these on the fly, by checking at each step in the inference routine for the presence of a zero in each measurement. When this occurs, the corresponding column of $\mathbf{C}^{(i)}$ is set to zero for all i.

We can also exploit the knowledge that the factor settings in a given application might tend to change slowly relative to the frequency of the measurements. Within the factorial model, it is possible to constrain the transitions so that only one factor can change its setting at each time step. Using the Gaussian sum approximation, this speeds up inference from order $O(K^2)$ per time step to $O(K \log K)$. We use this approximation in the experiments described in section VI.

V. Application to neonatal condition monitoring

We now turn our attention to the application of monitoring the condition of a premature baby receiving intensive care. Babies born three or four months prematurely in their first week *post partum* are kept in a closely regulated environment, with measurements of the heart rate, blood pressure, temperature and so on taken every second. An experienced clinician can make inferences about a baby's condition based on these signals, though this task is complicated by the fact that the observations depend not just on the state of a baby's physiology but also on the operation of the monitoring equipment. There is observation noise due to inaccuracies in the probes, and some operations can cause the measurements to become corrupted with artifact.

Much of the time babies can be expected to be in a "normal" state, where a degree of homeostasis is maintained and measurements are stable. In specific situations, characteristic patterns can appear which indicate particular conditions or pathologies. Some patterns are common and can be easily recognised, whereas at other times there might be periods of unusual physiological variation to which it is difficult to attribute a cause.

In this section, we first review previous work in intensive care unit (ICU) monitoring, then summarise the measurement channels which are to be analysed in this particular application. Constructing the model involves a combination of learning and domain knowledge. We first characterise the normal dynamics of the measurements, and then learn factor dynamics one by one to obtain the full factorial model.

A. Relation to previous work on ICU monitoring

We briefly review some relevant work in the specific area of intensive care unit monitoring. This work broadly fits into two categories. One approach is based on using domain knowledge to formulate high-level representations of particular patterns or situations, then to find suitable abstractions of the data in order to apply some matching rules. In this type of work, the goal is to describe what is happening, and sometimes to suggest what to do next; an *interpretation* is put on the data. Different schemes for heuristic description of patterns have been used, see for example [30]–[32].

By contrast, another body of work is based on making inferences of a statistical nature from monitoring data using time series analysis techniques. The goal in this case is to use the methodology of time series analysis to obtain informative *descriptions* of the data, which offer insight into the underlying processes. Notably, a switching linear dynamical system was used in [9] in order to identify statistically significant changes in liver function. Parametric models such as AR processes have been used to identify significant changes (e.g. level changes or slope changes) in physiological dynamics [33], [34]. Other work in this category has looked at finding segmentations of physiological monitoring data, e.g. finding segments which are approximately linear [35], [36].

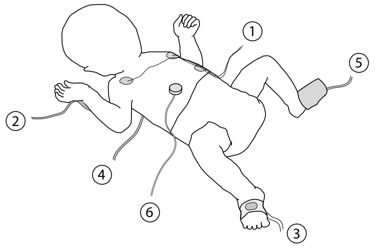


Fig. 4. Probes used to collect vital signs data from an infant in intensive care. 1) Three-lead ECG, 2) arterial line (connected to blood pressure transducer), 3) pulse oximeter, 4) core temperature probe (underneath shoulder blades), 5) peripheral temperature probe, 6) transcutaneous probe.

The first of these bodies of work uses expert knowledge, but captures it using a series of ad-hoc frameworks. The second uses established statistical techniques, but in general without incorporating the same level of expert insight and interpretation. The work described in this paper is motivated by the idea that these two approaches are not mutually exclusive, and uses extensive knowledge engineering within a principled (probabilistic) time series analysis framework.

B. Measurement channels

We now briefly describe the observations which are to be used in this application. A number of probes, illustrated in Figure 4, continuously collect physiological data from each baby. The resulting data channels are listed in Table I. *Heart rate* is obtained either from the ECG unit or blood pressure sensor. The latter also derives *systolic* and *diastolic blood pressure* measurements (the arterial pressure when the heart is contracting and relaxing, respectively). A transcutaneous probe, sited on the chest, measures the *partial pressures* of oxygen (Tc$\mathrm{PO_2}$) and carbon dioxide (Tc$\mathrm{PCO_2}$) in the blood[1]. A pulse oximeter, attached to the foot, measures the saturation of oxygen in arterial blood—a related but different quantity to transcutaneous $\mathrm{O_2}$. The *core temperature* and *peripheral temperature* are measured by two probes, one of which is placed under the baby's back (or under the chest if the baby is prone) and the other attached to a foot. In addition, environmental measurements (*ambient temperature* and *humidity*) are collected directly from the incubator. The probes used to collect these measurements are illustrated in Figure 4. All these measurements are taken once per second. All the data channels are applied without preprocessing to the model, with the exception of incubator humidity. It is necessary to apply a form of smoothing to this data channel because of measurement quantisation; the measurements change gradually relative to the measurement accuracy in this case, resulting in a "stepped" signal which causes problems during learning and inference.

[1] Various gases are dissolved in the bloodstream, and the partial pressure is used to quantify the amount of each. It is the amount of pressure that a particular gas would exert on a container if it was present without the other gases.

TABLE I
PHYSIOLOGICAL MEASUREMENT CHANNELS

Channel name	Label
Core body temperature (°C)	Core temp.
Diastolic blood pressure (mmHg)	Dia. Bp
Heart rate (bpm)	HR
Peripheral body temperature (°C)	Periph. temp.
Saturation of oxygen in pulse (%)	$\mathrm{SpO_2}$
Systolic blood pressure (mmHg)	Sys. Bp
Transcutaneous partial pressure of $\mathrm{CO_2}$ (kPa)	$\mathrm{TcPCO_2}$
Transcutaneous partial pressure of $\mathrm{O_2}$ (kPa)	$\mathrm{TcPO_2}$

C. Learning normal dynamics

In training the FSLDS model for this application, we first learn the "normal" dynamics for a baby. Much of the time, infants in intensive care are in a stable condition. Because infants with a low gestational age are usually asleep and motionless, there tends to be low variability in their vital signs when in a stable condition. The physiological systems underlying the observation channels are too complicated to model explicitly, being governed by complex interactions between a number of different sub-systems including the central nervous system. Instead, the approach adopted here is to try to find relatively simple models that are statistically compelling.

The approach used here for fitting linear Gaussian state-space models to each observation channel is first illustrated with heart rate observations, which are generally the least stable and most difficult to model of the observed channels. We then go on to show how this approach is adapted to model the other observed channels. Our resulting joint model is univariate in each observation channel, so that $\mathbf{A}$ and $\mathbf{Q}$ have a block diagonal structure. This makes it easy to add or remove channels from the overall model, and to specify the dependence of the state and channel dynamics on various factors.

1) Normal heart rate dynamics: Looking at examples of normal heart rate dynamics as in the top left and right panels of Figure 5, it can be observed first of all that the measurements tend to fluctuate around a slowly drifting baseline. This motivates the use of a model with two hidden components: the signal x_t, and the baseline b_t. These components are therefore used to represent the true heart rate, without observation noise. The dynamics can be formulated using autoregressive (AR) processes, such that an AR(p_1) signal varies around an AR(p_2) baseline, as given by the following equations:

$$x_t - b_t \quad \sim \quad \mathcal{N}\left(\sum_{k=1}^{p_1} \alpha_k (x_{t-k} - b_{t-k}), \eta_1\right) \ , \qquad (11)$$

$$b_t \quad \sim \quad \mathcal{N}\left(\sum_{k=1}^{p_2} \beta_k b_{t-k}, \eta_2\right) \ , \qquad (12)$$

where η_1, η_2 are noise variances. For example, an AR(2) signal with AR(2) baseline has the following state-space representation:

$$\mathbf{x}_t = \begin{bmatrix} x_t \\ x_{t-1} \\ b_t \\ b_{t-1} \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} \alpha_1 & \alpha_2 & 1-\alpha_1 & -\alpha_2 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & \beta_1 & \beta_2 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad (13)$$

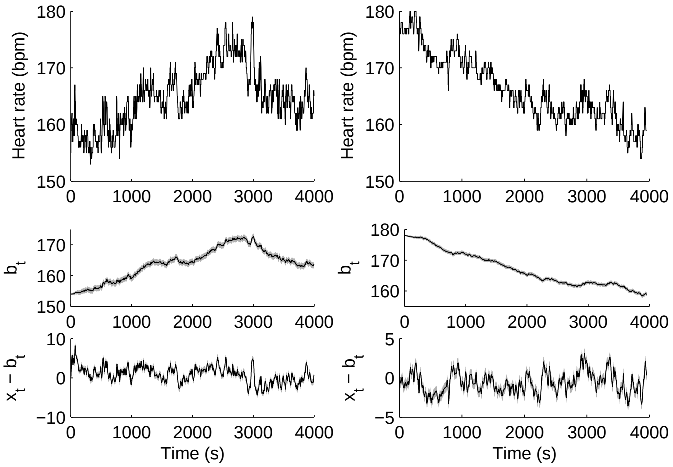


Fig. 5. In these two examples, HR measurements (in the top left and top right panels) are varying quickly within normal ranges. The estimates of the underlying signal (bottom left and bottom right panels) are split into a smooth baseline process and zero-mean high frequency component.

$$\mathbf{Q} = \begin{bmatrix} \eta_1 + \eta_2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \eta_2 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{C} = [1\ 0\ 0\ 0] \ . \qquad (14)$$

It is straightforward to adjust this construction for different values of p_1 and p_2. The measurements are therefore generally taken to be made up of a baseline with low frequency components and a signal with high frequency components. We begin training this model with a heuristic initialisation, in which we take sequences of training data and remove high frequency components by applying a symmetric 300-point moving average filter. The resulting signal is taken to be the low frequency baseline. The residual between the original sequences and the moving-averaged sequences are taken to contain both stationary high frequency hemodynamics as well as measurement noise. These two signals can be analysed according to standard methods and modelled as AR or integrated AR processes (specific cases of autoregressive integrated moving average (ARIMA) processes [37]) of arbitrary order. Heart rate sequences were found to be well modelled by an AR(2) signal varying around an ARIMA(1,1,0) baseline. An ARIMA model is a compelling choice for the baseline, because with a low noise term it produces a smooth drift[2]. Having found this initial setting of the model parameters, EM updates are then applied [17]. This has been found to be particularly useful for refining the estimates of the noise terms $\mathbf{Q}$ and $\mathbf{R}$.

Examples of the heart rate model being applied as a Kalman filter to heart rate sequences are shown in Figure 5. The top panels show sequences of noisy heart rate observations, and the lower panel shows estimates of the high frequency and low frequency components of the heart rate.

2) Other channels : Most of the remaining observation channels are modelled according to the same principle. Heart rate,

[2] The ARIMA(1,1,0) model has the form $(X_t - \beta X_{t-1}) = \alpha_1(X_{t-1} - \beta X_{t-2}) + Z_t$ where $\beta = 1$ and $Z_t \sim N(0, \sigma_Z^2)$. This can be expressed in un-differenced form as a non-stationary AR(2) model. In our implementation we set $\beta = 0.999$ and with $|\alpha_1| < 1$ we obtain a stable AR(2) process, which helps to avoid problems with numerical instability. This slight damping makes the baseline mean-reverting, so that the resulting signal is stationary. This has desirable convergence properties for dropout modelling.

systolic and diastolic blood pressures have the same structure—an AR(2) signal with an ARIMA(1,1,0) baseline. Transcutaneous O_2 and CO_2 are well modelled by an AR(2) signal with AR(1) baseline. All temperature measurements are modelled with an AR(1) signal and AR(1) baseline. Oxygen saturation and incubator humidity do not have a changing baseline, and are both sufficiently well modelled by AR(1) processes.

D. Learning dynamics under known factors

Having built a model for normal dynamics, in which the baby is stable and the monitoring equipment is operating correctly, we are in a position to consider different types of deviations from this regime, in which different factors can "overwrite" the model parameters. In this section, we show how the dynamics can be trained for the cases in which we have interpretable factor patterns (and can therefore obtain training data).

1) Drop-outs : Probe dropouts, which cause the observations on a given channel or set of channels to go to zero, are simple to model in this framework by taking normal dynamics and changing the appropriate entry in the observation matrix $\mathbf{C}$ to zero. This indicates that the relevant underlying physiology is entirely unobserved. In this way, the estimates of the underlying physiology are unaffected. Normal dynamics continue to update the estimates of the true physiology, but without being updated by the observations. The Kalman gain is always zero, so that the new observations have no weight upon the estimates. Uncertainty therefore increases until reaching a stable state.

2) Temperature probe disconnection : When a temperature probe becomes disconnected, artifactual measurements are received which reflect the transition of the probe from thermal equilibrium with the baby's body to equilibrium with the air in the incubator. The decay rate should be the same for each disconnection, since the same type of probe is used which therefore has the same thermal inertia. This gives a way of telling whether the probe is cooling according to Newton's laws of cooling or whether the baby is getting colder, for which there is no reason to assume the same type of dynamics. An exponential decay model (equivalent to an AR(1) process) for the artifactual temperature measurements is fitted using the Yule-Walker equations. During normal dynamics (temperature probe correctly applied), the artifactual temperature state is tied to the physiological temperature state. See Figure 6(b) for an example.

3) Blood sampling : An arterial blood sample might be taken every few hours from each baby. This involves diverting blood from the arterial line containing the pressure sensor, causing heart rate readings to cease. Throughout the operation a saline pump acts against the sensor, causing an artifactual ramp in the blood pressure measurements. The slope of the ramp is not always the same, as the rate at which saline is pumped can vary. See Fig. 7(b) for an example.

The average gradient of these artifactual ramps can be learnt for all blood samples, and used as a constant linear drift term. A state-space is then formulated which has a random walk on the differences of the data with a small noise term. In this way, the average drift is used as an initial guess, and the integrated random walk term can alter this guess to converge with the data.

4) Opening of the incubator : Incubator humidity and temperature are closely regulated, so that with all incubator portals shut the ambient humidity and temperature readings normally have low variance. When a portal is opened there is a significant drop in these readings. These drops can be modelled as an AR(1) decay, where the level to which these measurements drop is unknown but cannot be lower than the humidity and temperature of the room.

TABLE II
PARTIAL ORDERING OF FACTORS OVERWRITING PHYSIOLOGICAL CHANNELS. FACTORS HIGHER ON THE LIST OVERWRITE THE CHANNELS ON LOWER FACTORS.

	Heart rate	Sys BP	Dia BP	$TcPO_2$	$TcPCO_2$	SpO_2	Core temp.	Periph. temp.
Dropouts	■	■	■	■	■	■	■	■
Blood sample	■	■	■					
Temp. disconnection							■	
Incubator open	■	■	■			■		
TCP recalibration				■	■			
Bradycardia	■							
X-factor	■	■	■	■	■	■	■	■
Normal	■	■	■	■	■	■	■	■

The opening of the incubator implies that an intervention to the baby is taking place. This can be expected to have some kind of physiological effect, normally an increase of variance on the cardiovascular channels and a slight decrease in peripheral temperature due to the influx of room air in the incubator. Parameters can then be set by repeating the process for training normal dynamics on data which was obtained during handling episodes. In practice, this tends to result in physiological dynamics that are similar to the normal dynamics but with a larger system noise term. The signficant change in incubator humidity dynamics distinguishes this factor from the X-factor.

5) Bradycardia : Bradycardia is a slowing of the heart rate, and brief episodes are common for premature infants. It can have many causes, some benign and some serious. Bradycardic drops and subsequent rises in heart rate were found to be adequately modelled by retraining the ARIMA(1,1,0) model for baseline heart rate dynamics. The high frequency heart rate dynamics are kept the same as for the stable heart rate regime. As for the normal regime, this model learnt in terms of hidden ARIMA processes was used as an initial setting and updated with three iterations of EM.

6) Transcutaneous probe recalibration : Transcutaneous probes (TCPs) need to be recalibrated every few hours, and the resulting artifactual patterns have a number of distinct stages. First there is the application of a calibration solution to the probe, then the removal of this solution so that the probe gives a reading in room air, then the reapplication of the probe to the baby. After this final step, the levels of the measurements decay to the true physiological levels. The constant levels of the first stage do not require any dynamics to model; only a mean and a variance need to be specified. The other two stages are modelled as exponential decays.

E. Learning the factorial model

In section II-B we discussed the possibility of specifying a partial ordering of factors, such that some can overwrite particular observation channels of others. This is developed from earlier work in [38]. Table II shows the ordering of factors in this application, for example where the 'Incubator open' factor overwrites the normal blood pressure dynamics, but is itself overwritten by the 'Blood sample' factor (that is, if these two factors occur simultaneously, the blood pressure dynamics are entirely governed by the blood sample factor). Knowing this structure substantially reduces the amount of training data required. We simply learn LDS parameters for individual factor settings and then combine them accordingly [23, §5.10].

VI. Experiments

This section describes experiments used to evaluate the model for condition monitoring. Experiments done to evaluate the classification of known patterns are described in section VI-A, while section VI-B describes experiments done to evaluate the X-factor. Other than the X-factor, we consider here the incubator open/handling of baby factor (denoted 'IO'), the blood sample factor (denoted 'BS'), the bradycardia factor (denoted 'BR') and the temperature probe disconnection factor (denoted 'TD'). We demonstrate the operation of the transcutaneous probe recalibration factor (denoted 'TR'), but do not evaluate it quantitatively due to a scarcity of training data. We also have a dropout factor for each observation channel, but handle these implicitly in the inference routine (see section IV).

Some conventions in plotting the results of these experiments are adopted throughout this section. Horizontal bars below time-series plots indicate the posterior probability of a particular factor being active, with other factors in the model marginalised out. White and black indicate probabilities of zero and one respectively[3]. In general the plots show a subset of the observation channels and posteriors from a particular model—this is indicated in the text.

24-hour periods of monitoring data were obtained from fifteen premature infants in the intensive care unit at Edinburgh Royal Infirmary. The babies were between 24 and 29 weeks gestation (around 3-4 months premature), and all in around their first week *post partum*.

Each of the fifteen 24-hour periods was annotated by two clinical experts. At or near the start of each period, a 30 minute section of normality was marked, indicating an example of that baby's current baseline dynamics. Each of the known common physiological and artifactual patterns were also marked up.

Finally, it was noted where there were any periods of data in which there were clinically significant changes from the baseline dynamics not caused by any of the known patterns. While the previous annotations were made collaboratively, the two annotators marked up this 'Abnormal (other)' category independently. The software package TSNet [39] was used to record these annotations, and the recorded intervals were then exported into Matlab. The number of intervals for each category, as well as the total and average durations, are shown in Table III. The figures for the 'Abnormal' category were obtained by combining the two annotations, so that the total duration is the number of points which either annotator thought to be in this category, and the number of incidences was calculated by merging overlapping intervals in the two annotations (two overlapping intervals are counted as a single incidence).

TABLE III
Number of incidences of different factors, and total time for which each factor was annotated as being active in the training data (total duration of training data $15 \times 24 = 360$ hours).

Factor	Incidences	Total duration	Average duration
Incubator open	690	41 hours	3.5 mins
Abnormal (other)	605	32 hours	3.2 mins
Bradycardia	272	161 mins	35 secs
Blood sample	91	253 mins	2.8 mins
Temp. disconnection	87	572 mins	6.6 mins
TCP recalibration	11	69 mins	6.3 mins

TABLE IV
Inference results on three CV-folds of the evaluation data.

		Incu. open	Core temp.	Blood sample	Brady.
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

The rest of this section shows the results of performing inference on this data and comparing it to the gold standard annotations provided by the clinical experts.

A. Evaluation of known factors

In order to maximise the amount of test data and reduce the possibility of bias, evaluation was done with three-fold cross validation. The fifteen 24-hour data periods were split into three groups of five (grouped in order of the date at which each baby first arrived in the NICU). Three tests were therefore done for each model, in each case testing on five babies and training on the remaining ten, and summary statistics were obtained by averaging over the three runs. From each 24-hour period, a 30 minute section near the start containing only normal dynamics was reserved for calibration (learning normal dynamics according to section V-C). Testing was therefore conducted on the remaining $23\frac{1}{2}$ hour periods.

The quality of the inferences made were evaluated using area under the receiver operating characteristic curve (AUC) and equal error rates (EER)[4]. These statistics are a useful summary of performance when there are disparities in the numbers of points of each class.

Summary statistics for three types of models are given in Table IV, and the corresponding ROC curves are shown in Figure 8. Four factors are considered (incubator open, temperature probe disconnection, bradycardia and blood sample). Inferences are made for the set of factors with a factorial switching linear dynamical model, first with the Gaussian sum approximation, and then with Rao-Blackwellised particle filtering. The number of particles was set so that inference time was the same as for the Gaussian sum approximate inference, in this case $N = 71$.

[3] A convenient property of the models evaluated here, from the perspective of visualisation, is that the factor posteriors tend be close to zero or one. This is partly due to the fact that the discrete transition prior $p(s_t|s_{t-1})$ is usually heavily weighted towards staying in the same switch setting (long dwell times).

[4] EER is the error rate for the threshold setting at which the false positive rate is equal to the false negative rate. This a useful statistic when the number of true positives and negatives are unequal. We give error rates, so smaller numbers are better (some authors give 1 - EER).

For comparison, the same set of factors was inferred with the FHMM model, in which training was carried out using maximum likelihood estimation. The performance of the FHMM is a useful comparison because it has similar structure to the FSKF but with no hidden continuous dynamics. For all factors, the effect of adding the continuous latent dynamics is to improve performance, as can be seen by comparing the FHMM performance to the two FSKF models. RBPF inferences tend to be less accurate than those made with the Gaussian-sum approximation. This is at least partly due to the inability of the model to sample effectively from all the latent space when there is a high number of switch settings, and in this case the number of possible switch settings (16) is significant relative to the number of particles (71). Increasing the number of particles improves the inferences somewhat, though even when the number of particles in RBPF is doubled, we find that AUC only increases by 2-3%, well below the Gaussian sum results [23, §7.2.2].

It can be seen that core temperature probe disconnection is in general the most difficult factor to infer, partly because very long periods of disconnection are eventually misclassified by the model as being normal.

Specific examples of the operation of these models are now given. Figures 6-9 show inferences of switch settings made with the FSKF with Gaussian sum approximation (denoted ‘GS’ in Table IV). In each case the switch settings have been accurately inferred. Figure 6 shows examples of transcutaneous probe recalibration, correctly classified in conjunction with a blood sample and a core temperature probe disconnection. Note that in 6(b) the recalibration and disconnection begin at around the same time, as a nurse has handled the baby in order to access the transcutaneous probe, causing the temperature probe to become detached.

Figure 7 shows inference of bradycardia, blood sampling, and handling of the baby. Note in 7(a) that it has been possible to recognise the disturbance of heart rate at $t = 800$ as being caused by handling of the baby, distinguished from the bradycardia earlier where there is no evidence of the incubator having been entered.

For the blood sample and temperature probe disconnection factors, the measurement data bears no relation to the actual physiology, and the model should update the estimated distribution of the true physiology in these situations accordingly. Figure 9 contains examples of the inferred distribution of true physiology in data periods in which these two artifacts occur. In each case, once the artifactual pattern has been detected, the physiological estimates remain constant or decay towards a mean. As time passes since the last reliable observation, the variance of the estimates increases towards a steady state.

B. Novelty detection

In practice, neonatal monitoring data exhibits many unusual patterns. The number of potential unusual patterns is in fact so great that it would be impractical to explicitly include every possibility in a model. Examples include rare dynamical regimes caused by sepsis, neurological problems, or the administration of drugs, even a change of linen or the flash of a camera. Experiments were done to evaluate the ability of the X-factor to represent novel physiological and artifactual dynamics. Preliminary trials (including EM estimation) showed $\xi = 1.2$ to be a suitable setting.

Three-fold cross validation was again used to analyse the inferences of different models with different sets of factors. The first model considered contained only the X-factor, the two switch settings therefore being ‘normal’ or ‘abnormal’. The intention with this construction was for it to place probability mass for the X-factor on any period in which anything non-normal was happening. As the X-factor here stands in for any known or unknown pattern, the ground truth for this model is the conjunction of all the annotated intervals of every type—known factors and ‘abnormal’ periods. Another four models are considered, in which the known factors are added to the model one by one. So, for the second model the ‘Incubator Open’ factor is added and the corresponding intervals are removed from the ground truth for the X-factor. The factors are added in reverse order of total duration in Table III. In the fifth set of factors each known factor has ground truth given by the corresponding annotation, and the X-factor has ground truth given by the ‘Abnormal (other)’ annotation. Examining the performance of these different models and particular examples of operation gives some insight into the operation of the X-factor, both on its own and in conjunction with the other factors.

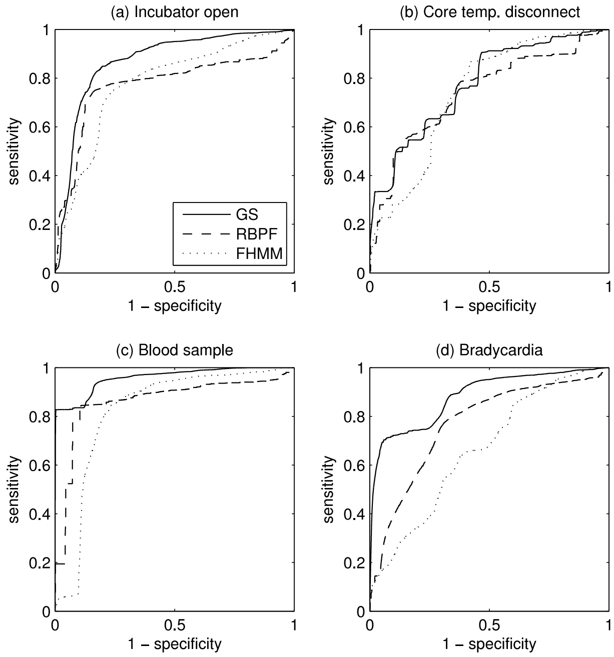
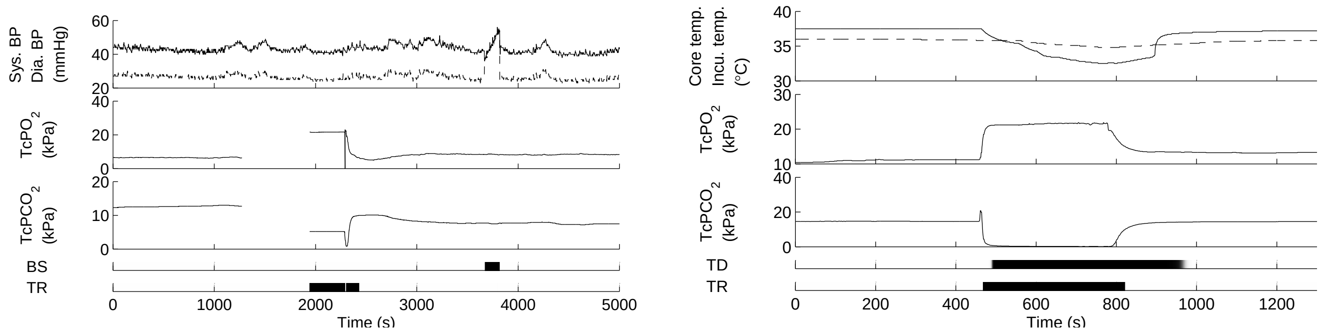


Fig. 8. ROC curves for classification of four known factors.

Summary statistics are shown in Table V, where the models above are numbered 1-5. Only approximate Gaussian sum inference was considered here. The performance in classifying the presence of known factors is almost the same as for when the X-factor was not included (model ‘GS’ in Table IV), only minor variations in AUC and EER being evident. For each of the five models, the X-factor inferences had a rough correlation to the annotations.

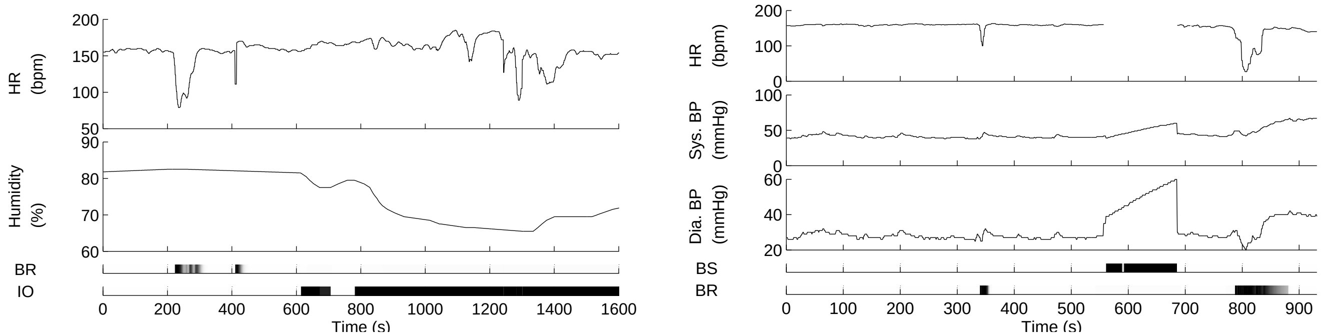
Examples of the operation of the X-factor are shown in Figures 10-12, beginning with inferences from model 5 in which the full set of factors is present with the X-factor. Figure 10 shows two examples of inferred switch settings under this model for periods in which there are isolated physiological disturbances. Both the posteriors for the X-factor and the gold standard intervals for the ‘Abnormal (other)’ category are shown. The physiological disturbances in both panels are cardiovascular and have clearly observable effects on the blood pressure and oxygen saturation measurements.



(a)

(b)

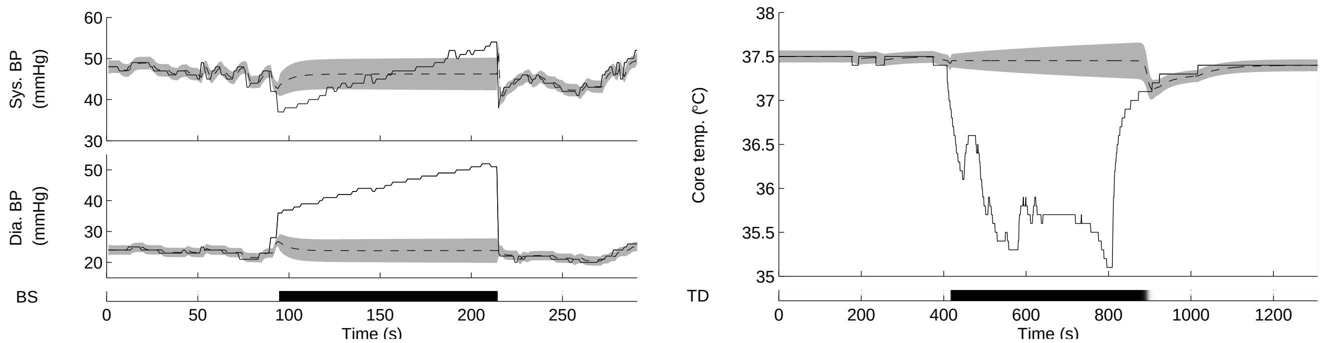
Fig. 6. Inferred distributions of switch settings for two situations involving recalibration of the transcutaneous probe. BS denotes a blood sample, TR denotes a recalibration, and TD denotes a core temperature probe disconnection. In panel (a) the recalibration is preceeded by a dropout, followed by a blood sample. Diastolic BP is shown as a dashed line which lies below the systolic BP plot. Transcutaneous readings drop out at around $t = 1200$ before the recalibration. In panel (b), the solid line shows the core temperature and the dashed line shows incubator temperature. A core temperature probe disconnection is identified correctly, as well as the recalibration. Temperature measurements can occasionally drop below the incubator temperature if the probe is near to the portals; this is accounted for in the model by the system noise term $\mathbf{Q}$.



(a)

(b)

Fig. 7. Inferred distributions of switch settings for two further situations in which there are effects due to multiple known factors. In panel (a) there are incidences of bradycardia, after which the incubator is entered. There is disturbance of heart rate during the period of handling, which is correctly taken to be associated with the handling and not an example of spontaneous bradycardia. In panel (b), bradycardia and blood samples are correctly inferred. During the blood sample, heart rate measurements (supplied by the blood pressure sensor) are interrupted.



(a)

(b)

Fig. 9. Inferred distributions of the true physiological state during artifactual corruption of measurements. Panel (a) shows correct inference of the duration of a blood sample, and panel (b) shows correct inference of a temperature probe disconnection. Measurements are plotted as a solid line, and estimates $\hat{\mathbf{x}}_t$ relating to true physiology are plotted as a dashed line with the gray shading indicating two standard deviations. In each case, during the period in which measurements are corrupted the estimates of the true physiology are propagated with increased uncertainty.

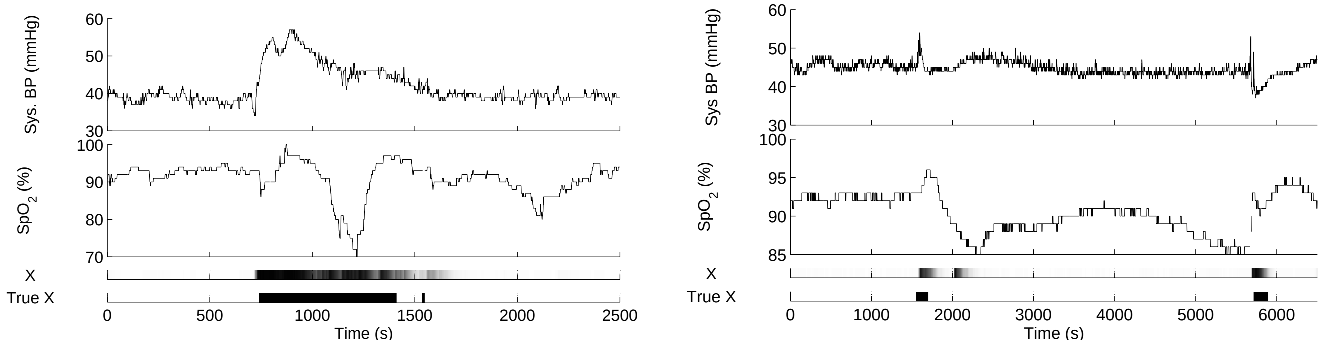


Fig. 10. Inferred switch settings for the X-factor, during periods of cardiovascular disturbance, compared to the gold standard annotations.

TABLE V
SUMMARY STATISTICS FOR THE QUALITY OF X-FACTOR INFERENCES, FOR MODELS 1-5. SEE MAIN TEXT FOR DETAILS.

		X-factor	Incu. open	Core temp.	B. sample	Brady.
1	AUC	.72	-	-	-	-
	EER	.33	-	-	-	-
2	AUC	.74	.87	-	-	-
	EER	.32	.17	-	-	-
3	AUC	.71	.87	.78	-	-
	EER	.35	.18	.28	-	-
4	AUC	.70	.87	.78	.96	-
	EER	.36	.18	.28	.14	-
5	AUC	.69	.87	.79	.96	.88
	EER	.36	.18	.28	.14	.25

In Figure 10 (left), the X-factor is triggered by a sudden, prolonged increase in blood pressure and a desaturation, in broad agreement with the ground truth annotation. In Fig. 10 (right) there are two spikes in BP and shifts in saturation which are picked up by the X-factor, also mainly in agreement with the annotation. A minor turning point in the two channels was also picked up at around $t = 2000$, which was not considered significant in the gold standard (a false positive).

Effects of introducing known factors to model (1) are shown in Figure 11. In panel (a), there are two occurrences of spontaneous bradycardia, HR making a transient drop to around 100bpm. The X-factor alone in model (1) picks up this variation. Looking at the inferences from model (5) for the same period, it can be seen that the bradycardia factor provides a better match for the variation, and probability mass shifts correctly: the X-factor is now inactive. In panel (b), a similar effect occurs for a period in which a blood sample occurs. The X-factor picks up the change in dynamics when on its own, and when all factors are present in model (5) the probability mass shifts correctly to the blood sample factor. The blood sample factor is a superior description of the variation, incorporating the knowledge that the true physiology is not being observed, and so able to handle the discontinuity at $t = 900$ effectively.

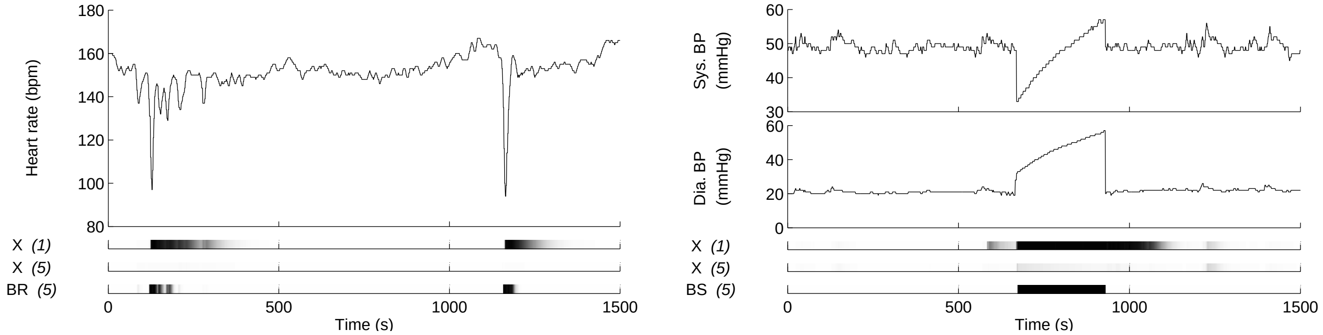
Figure 12 shows examples of inferred switch settings from model (5) in which there are occurrences of both known and unknown types of variation. In Fig. 12(a) a bradycardia occurs in the middle of a period of elevated blood pressure and a deep drop in saturation. The bradycardia factor is active for a period which corresponds closely to the ground truth. The X-factor picks up the presence of a change dynamics at about the right time, but its onset is delayed when compared to the ground truth interval. This again highlights a difficulty with filtered inference, since at time just over 1000 it is difficult to tell that this is the beginning of a significant change in dynamics without the benefit of hindsight. In panel (b) a blood sample is correctly picked up by the blood sample factor, while a later period of physiological disturbance on the same measurement channels is correctly picked up by the X-factor. Panel (c) shows another example of the bradycardia factor operating with the X-factor, where this time the onset of the first bradycardia is before the onset of the X-factor. The X-factor picks up a desaturation, a common pattern which is already familiar from panel (a). In panel (d), an interaction between the X-factor and the 'Incubator open' factor can be seen. From time 270 to 1000 the incubator has been opened, and all variation including the spike in HR at $t = 420$ are attributed to handling of the baby. Once the incubator appears to have been closed, further physiological disturbance is no longer explained as an effect of handling and is picked up by the X-factor.

VII. DISCUSSION

This paper has presented a general framework for inferring hidden factors from monitoring data, and has shown its successful application to the significant real-world task of monitoring the condition of a premature infant receiving intensive care. We have shown how knowledge engineering and learning can be successfully combined in this framework. Our formulation of an additional factor (the "X-factor") allows the model to handle novel dynamics. Experimental demonstration has shown that these methods are effective when applied to genuine monitoring data.

There are a number of directions in which this work could be continued. The set of known factors presented here is limited, and more could usefully be added to the model given training data. Deep oxygen desaturations, as seen in Figure 12(a) and (c), are currently handled by the X-factor, but are a clear and significant pattern that could be usefully learnt as a new factor. Desaturation is usually followed by a bradycardia, since lack of oxygen to the heart will slow it. We would therefore want to change the bradycardia factor dynamics to be a priori dependent on the new desaturation factor (a departure from eq. (4)). Additional factors could include other common patterns such as hypotension, hypertension, hypothermia and pyrexia, as well as serious conditions such as pneumothorax or intraventricular haemorrhage.

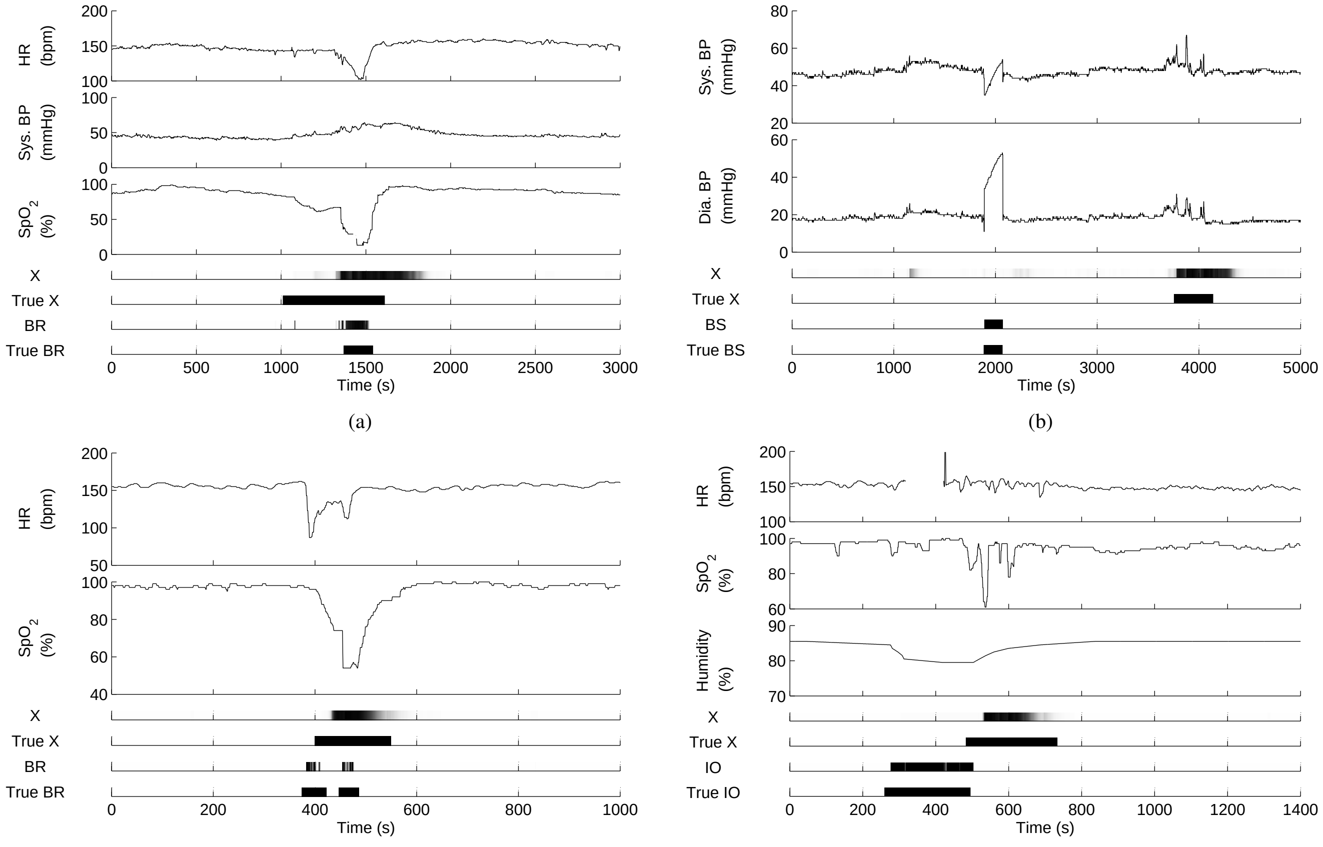
Bradycardia is often associated with a compensatory rise in blood pressure, as seen in Figure 12(a). Incorporating this effect



(a)

(b)

Fig. 11. Inferred switch settings for the X-factor, and for known patterns for models (1) and (5) in Table V. Model (1) contains the X-factor only, whereas model (5) includes the X-factor and all known factors. Panel (a) shows two instances of bradycardia, (b) shows a blood sample.



(a)

(b)

(c)

(d)

Fig. 12. Inferred switch settings for the X-factor, in regions where other factors are active. In panel (a) a bradycardia occurs in conjunction with a rise in blood pressure and deep desaturation. The X-factor is triggered around the right region but is late compared to ground truth. In panel (b), unusual BP variation is correctly classified as being due to a blood sample, followed by variation of unkown cause. Panel (c) shows bradycardia with a desaturation picked up by the X-factor, and (d) shows the X-factor picking up disturbance after the incubator has been entered.

into the model would help to stop the elevation in BP (which we have an explanation for) being claimed by the X-factor. To do this, we would introduce a new factor governing the BP observations which is a priori dependent on the bradycardia factor.

The experiments with the X-factor have shown that there are a significant number of non-normal regimes in the data which have not yet been formally analysed. Future work might therefore look at learning what different regimes are claimed by the X-factor. This could be cast as an unsupervised or semi-supervised learning problem within the model.

The FSLDS with novelty detection is a general model for condition monitoring in multivariate time series, and could potentially be applied in many other domains. Also, we have only considered filtered inference in this paper, being interested in real-time diagnosis. An interesting extension would be to consider fixed-lag smoothing [40] in such problems.

ACKNOWLEDGMENTS

We thank Jim Hunter for modifying the Time Series Workbench software for use in this research, and to Birgit Wefers for supplying additional annotation of the data. We also thank the anonymous referees and the Associate Editor for their comments which have helped improve the manuscript. Author JQ was funded by the premature baby charity BLISS. The work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views.

REFERENCES

[1] D. Alspach and H. Sorenson, "Nonlinear Bayesian Estimation using Gaussian Sum Approximation," *IEEE Trans. Autom. Control*, vol. 17, pp. 439–447, 1972.

[2] R. Shumway and D. Stoffer, "Dynamic Linear Models with Switching," *J. Am. Statistical Assoc.*, vol. 86, pp. 763–769, 1991.

[3] N. de Freitas, R. Dearden, F. Hutter, R. Morales-Menedez, J. Mutch, and D. Poole, "Diagnosis by a waiter and a Mars explorer," *Proc. IEEE*, vol. 92, no. 3, 2004.

[4] R. Morales-Menedez, N. de Freitas, and D. Poole, "Real-Time Monitoring of Complex Industrial Processes with Particle Filters," in *Advances in Neural Information Processing Systems 15*, S. Becker, S. Thrun, and K. Obermayer, Eds. MIT Press, 2002.

[5] U. Lerner, R. Parr, D. Koller, and G. Biswas, "Bayesian fault detection and diagnosis in dynamic systems," in *AAAI*, 2000, pp. 531–537.

[6] V. Pavlović, J. Rehg, and J. MacCormick, "Learning Switching Linear Models of Human Motion," in *Advances in Neural Information Processing Systems 13*, T. Leen, T. Dietterich, and V. Tresp, Eds. MIT Press, 2000.

[7] Li, Y. and Wang, T. and Shum, H.-Y., "Motion Texture: A Two-Level Statistical Model for Character Motion Synthesis," in *SIGGRAPH*, 2002, pp. 465–472.

[8] M. Azzouzi and I. Nabney, "Modelling Financial Time Series with Switching State Space Models," *Proceedings of the IEEE/IAFE Conference on Computational Intelligence for Financial Engineering*, pp. 240–249, 1999.

[9] A. Smith and M. West, "Monitoring Renal Transplants: An Application of the Multiprocess Kalman Filter," *Biometrics*, vol. 39, pp. 867–878, 1983.

[10] J. Droppo and A. Acero, "Noise Robust Speech Recognition with a Switching Linear Dynamic Model," in *Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing*, 2004.

[11] J. Ma and L. Deng, "A mixed level switching dynamic system for continuous speech recognition," *Computer Speech and Language*, vol. 18, pp. 49–65, 2004.

[12] A. Cemgil, H. Kappen, and D. Barber, "A Generative Model for Music Transcription," *IEEE Trans. Speech Audio Process.*, vol. 14, no. 2, pp. 679–694, 2006.

[13] C. Williams, J. Quinn, and N. McIntosh, "Factorial Switching Kalman Filters for Condition Monitoring in Neonatal Intensive Care," in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. Platt, Eds. MIT Press, 2006.

[14] J. Quinn and C. Williams, "Known Unknowns: Novelty Detection in Condition Monitoring," in *Proc 3rd Iberian Conference on Pattern Recognition and Image Analysis*, J. Martí, J.-M. Benedí, A. M. Mendonça, and J. Serrat, Eds. Springer, 2007.

[15] J. Quinn, "Neonatal condition monitoring demonstration code," http://cit.ac.ug/jquinn/software.html, 2008.

[16] K. Tsien, "Dynamic Bayesian networks: representation, inference and learning," Ph.D. dissertation, University of California, Berkeley, 2002.

[17] Z. Ghahramani and G. Hinton, "Parameter Estimation for Linear Dynamical Systems," Department of Computer Science, University of Toronto, Tech. Rep., 1996.

[18] ——, "Variational learning for switching state-space models," *Neural Computation*, vol. 12, no. 4, pp. 963–996, 1998.

[19] J. Candy, *Model-Based Signal Processing*. Wiley-IEEE Press, 2005.

[20] P. C. Woodland, "Hidden Markov Models using Vector Linear Prediction and Discriminative Output Distributions," in *Proceedings of 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. I. IEEE, 1992, pp. 509–512.

[21] Z. Ghahramani and M. Jordan, "Factorial Hidden Markov Models," *Machine Learning*, vol. 29, pp. 245–273, 1997.

[22] M. West and J. Harrison, *Bayesian Forecasting and Dynamic Models*. Springer, 1999.

[23] J. Quinn, "Bayesian Condition Monitoring in Neonatal Intensive Care," Ph.D. dissertation, University of Edinburgh, http://www.era.lib.ed.ac.uk/handle/1842/1645, 2007.

[24] M. Markou and S. Singh, "Novelty detection: a review - part 1: statistical approaches," *Signal Processing*, vol. 83, pp. 2481–2497, 2003.

[25] J. Ma and S. Perkins, "Online Novelty Detection on Temporal Sequences," *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 613–618, 2003.

[26] P. Smyth, "Markov monitoring with unknown states," *IEEE Journal on Selected Areas in Communications*, vol. 12(9), pp. 1600–1612, 1994.

[27] U. Lerner and R. Parr, "Inference in Hybrid Networks: Theoretical Limits and Practical Algorithms," in *Proceedings of the 17th Annual Conference on Uncertainty in Artificial Intelligence*, 2001, pp. 310–318.

[28] K. Murphy, "Switching Kalman filters," U.C. Berkeley, Tech. Rep., 1998.

[29] K. Murphy and S. Russell, "Rao-Blackwellised particle filtering for dynamic Bayesian networks," in *Sequential Monte Carlo in Practice*, A. Doucet, N. de Freitas, and N. Gordon, Eds. Springer-Verlag, 2001.

[30] J. Hunter and N. McIntosh, "Knowledge-Based Event Detection in Complex Time Series Data," in *Proceedings of the Joint European Conference on Artificial Intelligence in Medicine and Medical Decision Making*, W. Horn, Y. Shahar, G. Lindberg, S. Andreassen, and J. Wyatt, Eds., 1999.

[31] S. Miksch, Horn, C. Popow, and F. Paky, "Utilizing Temporal Data Abstraction for Data Validation and Therapy Planning for Artificially Ventilated Newborn Infants," *Artificial Intelligence in Medicine*, vol. 8, no. 6, pp. 543–576, 1996.

[32] I. J. Haimowitz, P. P. Le, and I. S. Kohane, "Clinical monitoring using regression based trend templates," *Artificial Intelligence in Medicine*, vol. 7, no. 6, pp. 473–496, 1995.

[33] M. Imhoff, M. Bauer, U. Gather, and D. Löhlein, "Statistical pattern detection in univariate time series of intensive care in-line monitoring data," *Intensive Care Med*, vol. 24, pp. 1305–1314, 1998.

[34] S. Hoare and P. Beatty, "Automatic artifact identification in anaesthesia patient record keeping: a comparison of techniques," *Medical Engineering and Physics*, vol. 22, pp. 547–553, 2000.

[35] S. Charbonnier, G. Becq, and G. Biot, "On-Line Segmentation Algorithm for Continuously Monitored Data in Intensive Care Units," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 3, pp. 484–492, 2004.

[36] R. Kennedy, "A modified Trigg's Tracking Variable as an 'advisory' alarm during anaesthesia," *Journal of Clinical Monitoring and Computing*, vol. 12, pp. 197–204, 1995.

[37] R. Shumway and D. Stoffer, *Time Series Analysis and Its Applications*. Springer-Verlag, 2000.

[38] A. Spengler, "Neonatal baby monitoring," Master's thesis, School of Informatics, University of Edinburgh, 2003.

[39] J. Hunter, "TSNet A Distributed Architecture for Time Series Analysis," in *Intelligent Data Analysis in bioMedicine and Pharmacology (IDAMAP 2006)*, N. Peek and C. Combi, Eds., 2006, pp. 85–92.

[40] D. Barber and B. Mesot, "A Novel Gaussian Sum Smoother for Approximate Inference in Switching Linear Dynamical Systems," in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. Platt, Eds. MIT Press, 2006.