Discriminative Switching Linear Dynamical Systems applied to Physiological Condition Monitoring

Konstantinos Georgatzis
School of Informatics
University of Edinburgh
k.georgatzis@sms.ed.ac.uk

Christopher K. I. Williams
School of Informatics
University of Edinburgh
ckiw@inf.ed.ac.uk

Abstract

We present a Discriminative Switching Linear Dynamical System (DSLDS) applied to patient monitoring in Intensive Care Units (ICUs). Our approach is based on identifying the state-of-health of a patient given their observed vital signs using a discriminative classifier, and then inferring their underlying physiological values conditioned on this status. The work builds on the Factorial Switching Linear Dynamical System (FSLDS) (Quinn *et al.*, 2009) which has been previously used in a similar setting. The FSLDS is a generative model, whereas the DSLDS is a discriminative model. We demonstrate on two real-world datasets that the DSLDS is able to outperform the FSLDS in most cases of interest, and that an α-mixture of the two models achieves higher performance than either of the two models separately.

Condition monitoring of patients in intensive care units (ICUs) based on vital signs (e.g. heart rate, blood pressure) is of critical importance, as they can be subject to a number of serious physiological events such as bradycardia and hypotension. However, a variety of artifactual processes can "contaminate" the data, e.g. the taking of blood samples, performing suctions, recalibrating sensors, etc. These artifactual processes complicate the task of identifying the important physiological events and are the main source of false alarms in ICUs. Moreover, it is of interest to maintain beliefs about the true physiological values of a patient when these cannot be directly observed due to artifact. For example, it would be desirable to display the patient's estimated blood pressure, when the corresponding measuring device has been disconnected or is otherwise displaying artifactual values (as is the case during a blood sample event). Of course, this estimate should be clearly distinguishable from the raw data (e.g. by using a different display colour).

One approach to this problem is to build a latent variable model, using a number of discrete latent variables to model the physiological and artifactual events through time, and a linear dynamical system (LDS) conditional on these discrete variables to model the associated dynamics in the vital signs observations. This is the factorial switching LDS (or FSLDS) of Quinn *et al.* (2009). However, we have noticed that in building such systems it is necessary to construct quite detailed models of the artifactual events in order to capture them properly. This can be non-trivial since some of these events can be highly variable, which is hard to capture with a generative model. Despite this high variability, the vital signs can still contain informative features which could act as input to a discriminative model. Thus, if it is possible to build such a model that can fairly easily distinguish between the various events, then it would seem simpler and easier to make the discrete-state inference be discriminative, and use FSLDS-style inference for the continuous latent variables conditional on the inferred discrete state. We call this a *discriminative switching linear dynamical system* (DSLDS). In this paper we compare the FSLDS and DSLDS models on two ICU condition monitoring datasets. The results show that using the DSLDS gives increased performance in most cases of interest, and that an α-mixture of the two methods was able to achieve a higher performance than either of the two models separately.

To summarise, our goal is to build a model with increased performance for the following tasks:

- Identifying artifactual processes (e.g blood samples), which will reduce the high false alarm rate in ICUs and facilitate the task of identifying physiological processes.

- Identifying physiological processes which can be of critical importance (e.g bradycardias).

- Providing an estimate of a patient's true physiological values when these are obscured by artifact.

The structure of the remainder of the paper is as follows: in Section 1 we give a description of our proposed model and

compare its graphical structure and inference methods to those of the FSLDS, and briefly describe related work. In Section 2 we describe our experiments and provide results for the comparison between the DSLDS and the FSLDS. Finally, in Section 3 we conclude with general remarks about our proposed model and suggestions for future work.

1 MODEL DESCRIPTION

The graphical model of the FSLDS is depicted in Figure 1 (top). It operates on three different sets of variables: The observed variables, $\mathbf{y}_t \in \mathbb{R}^{d_y}$ represent the patient's vital signs obtained from the monitoring devices at time t, which act as the input to our model. The continuous latent variables, $\mathbf{x}_t \in \mathbb{R}^{d_x}$, track the evolution of the dynamics of a patient's underlying physiology. The discrete variable, s_t, represents the switch setting or regime which the patient is currently in (e.g. stable, a blood sample is being taken etc.). The switch variable can be factorised according to the cross-product of M factors, so that $s_t = f_t^1 \otimes f_t^2 \otimes ... \otimes f_t^M$. Each factor variable, f_t^m, is usually a binary vector indicating the presence or absence of a factor, but in general it can take on $L^{(m)}$ different values and $K = \prod_{m=1}^{M} L^{(m)}$ is the total number of possible configurations of the switch variable, s_t. Also, s_t depends explicitly on the previous time step, so that $p(s_t|s_{t-1}) = \prod_{m=1}^{M} p(f_t^m|f_{t-1}^m)$. Conditioned on a particular regime, the FSLDS is equivalent to an LDS. The FSLDS can be seen then as a collection of LDS's, where each LDS models the dynamics of a patient's underlying physiology under a particular regime, and can also be used to generate a patient's observed vital signs. An LDS provides a generative framework for modelling our belief over the state space, given observations.

We can alternatively adopt a discriminative view. We start by modelling $p(s_t|\mathbf{y}_{t-l:t+r})$ with a discriminative classifier, where (features of) observations from the previous l and future r time steps affect the belief of the model about s_t. The inclusion of r frames of future context is analogous to fixed-lag smoothing in an FSLDS (see e.g. Särkkä, 2013, sec. 10.5). We note that inclusion of future observations in the conditioning set means that the DSLDS will operate with a delay of r seconds, since an output of the model at time t can be produced only after time $t + r$. Provided that r is small enough ($r \leq 10$ in experiments), this delay is negligible compared to the increase in performance. The LDS can also be regarded from a similarly discriminative viewpoint which allows us to model $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}_t)$. This is similar to the Maximum Entropy Markov Model (MEMM) (McCallum *et al.*, 2000) with the difference that the latent variable is continuous rather than discrete. The main advantage of this discriminative view is that it allows for a rich number of (potentially highly correlated) features to be used without having to explicitly model their distribution or the interactions between them, as is the case in a generative model. A combination of these two discrimina-

tive viewpoints gives rise to the DSLDS graphical model in Figure 1 (bottom). The DSLDS, conditioned on s_t, can be seen then as a collection of MEMM's, where each MEMM in the DSLDS plays a role equivalent to that of each LDS in the FSLDS.

The DSLDS can be defined as

$$p(\mathbf{s}, \mathbf{x}|\mathbf{y}) = p(s_1|\mathbf{y}_1)p(\mathbf{x}_1|s_1, \mathbf{y}_1) \times$$
$$\prod_{t=2}^{T} p(s_t|\mathbf{y}_{t-l:t+r})p(\mathbf{x}_t|\mathbf{x}_{t-1}, s_t, \mathbf{y}_t) \quad . \quad (1)$$
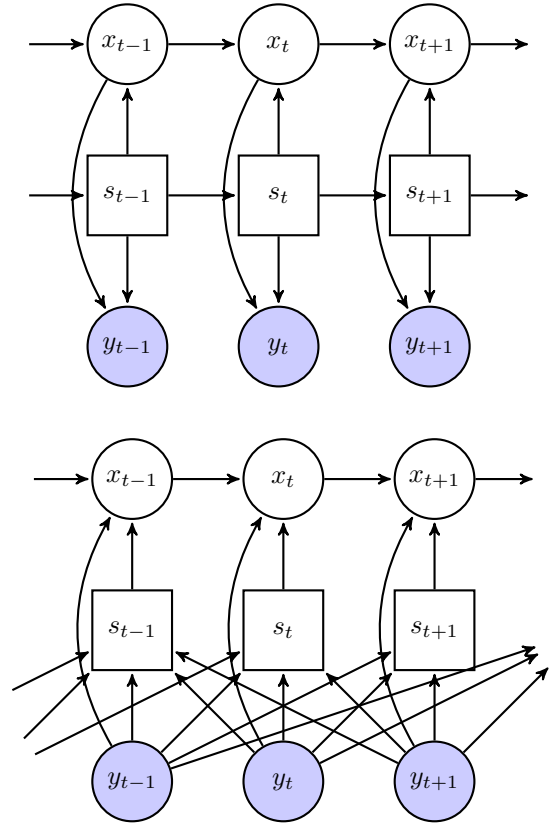

Figure 1: Graphical model of the FSLDS (top) and the DSLDS (bottom). The state-of-health and underlying physiological values of a patient are represented by s_t and $\mathbf{x}_t$ respectively. The shaded nodes correspond to the observed physiological values, $\mathbf{y}_t$. Note that in the case of the DSLDS the conditional probability $p(s_t|\mathbf{y}_{t-l:t+r})$ is modelled directly.

The simplest assumption we can make for the DSLDS is that $p(s_t|\mathbf{y}_{t-l:t+r})$ factorises, so that

$$p(s_t|\mathbf{y}_{t-l:t+r}) = \prod_{m=1}^{M} p(f_t^{(m)}|\mathbf{y}_{t-l:t+r}) \quad . \quad (2)$$

However, one could use a structured output model to predict the joint distribution of different factors.

1.1 Predicting s_t

Our belief about the state of health of a patient at time t is modelled by $p(s_t|\mathbf{y}_{t-l:t+r})$, the conditional probability of the switch variable given the observed vital signs. Following the factorisation of the switch variable in eq. 2, we model the conditional probability of each factor being active at time t given the observations with a probabilistic discriminative binary classifier, so that $p(f_t^{(i)} = 1|\mathbf{y}_{t-l:t+r}) = G(\phi(\mathbf{y}_{t-l:t+r}))$, where $G(\cdot)$ is a classifier-specific function, and $\phi(\mathbf{y}_{t-l:t+r})$ is the feature vector that acts as input to our model at each time step as described in Section 2.1. As is evident from Figure 1 (bottom) there is no explicit temporal dependence on the switch variable sequence. However, temporal continuity is implicitly incorporated in our model through the construction of the features.

1.1.1 An α-mixture of s_t

The DSLDS model can be seen as complementary to the FSLDS, and they can be run in parallel. One way of combining the two outputs is to maintain an α-mixture over s_t. If $p_g(s_t)$ and $p_d(s_t)$ are the outputs for the switch variable at time t from FSLDS and the DSLDS respectively, then their α-mixture is given by: $p_\alpha(s_t) = c\left(p_g(s_t)^{(1-\alpha)/2} + p_d(s_t)^{(1-\alpha)/2}\right)^{2/(1-\alpha)}$, where c is a normalisation constant which ensures that $p_\alpha(s_t)$ is a probability distribution. The family of α-mixtures then subsumes various known mixtures of distributions and defines a continuum across them via the α parameter. For example, for $\alpha = -1$ we retrieve the mixture of experts (with equally weighted experts) framework, while for $\alpha \to 1$, the formula yields $p_1(s_t) = c\sqrt{p_g(s_t)p_d(s_t)}$, rendering it equivalent to a product of experts viewpoint. In general, as α increases, the α-mixture assigns more weight to the smaller elements of the mixture (with $\alpha \to \infty$ giving $p_\infty(s_t) = \min\{p_g(s_t), p_d(s_t)\}$), while as α decreases, more weight is assigned to the larger elements (with $\alpha \to -\infty$ giving $p_{-\infty}(s_t) = \max\{p_g(s_t), p_d(s_t)\}$) A thorough treatment is given in Amari (2007).

1.2 Predicting $\mathbf{x}_t$

The model of the patient's physiology should capture the underlying temporal dynamics of their observed vital signs under their current health state. The idea is that the current latent continuous state of a patient should be dependent on (a) the latent continuous state at the previous time step, (b) the current state of health and (c) the current observed val-

ues. We model these assumptions as follows

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}, s_t, \mathbf{y}_t) \propto$$
$$\exp\{-\frac{1}{2}(\mathbf{x}_t - \mathbf{A}^{(s_t)}\mathbf{x}_{t-1})^\top(\mathbf{Q}^{(s_t)})^{-1}(\mathbf{x}_t - \mathbf{A}^{(s_t)}\mathbf{x}_{t-1})\} \times$$
$$\exp\{-\frac{1}{2}(\mathbf{C}^{(s_t)}\mathbf{x}_t - \mathbf{y}_t)^\top(\mathbf{R}^{(s_t)})^{-1}(\mathbf{C}^{(s_t)}\mathbf{x}_t - \mathbf{y}_t)\} \quad . \quad (3)$$

The first term on the RHS of eq. 3 is the **system model** for an LDS and captures the dynamics of a patient's latent physiology under state s_t. The second term can be seen as the discriminative counterpart of the **observation model** of an LDS. In our condition monitoring setting, the observed vital signs are considered to be noisy realisations of the true, latent physiology of a patient and thus, the observation model encodes our belief that $\mathbf{x}_t$ is a noisy version of $\mathbf{y}_t$. Under this assumption, $\mathbf{C}^{s_t}$ consists of 0/1 entries, which are set based on our knowledge of whether the observations $\mathbf{y}_t$ are artifactual or not under state s_t. In the FSLDS, the corresponding observation model encodes the belief that the generated $\mathbf{y}_t$ should be normally distributed around $\mathbf{x}_t$ with covariance $\mathbf{R}^{s_t}$, whereas in our discriminative version, the observation model encodes our belief that $\mathbf{x}_t$ should be normally distributed around $\mathbf{y}_t$ with covariance $\mathbf{R}^{s_t}$. The idea behind this model is that at each time step we update our belief about $\mathbf{x}_t$ conditioned on its previous value, $\mathbf{x}_{t-1}$, and the current observation, $\mathbf{y}_t$, under the current regime s_t. For example, under an artifactual process, the observed signals do not convey useful information about the underlying physiology of a patient. In that case, we drop the connection between $\mathbf{y}_t$ and $\mathbf{x}_t$ (for the artifact-affected channels) which translates into setting the respective entries of $\mathbf{C}^{s_t}$ to zero. Then, the latent state $\mathbf{x}_t$ evolves only under the influence of the appropriate system dynamics parameters $(\mathbf{A}^{(s_t)}, \mathbf{Q}^{(s_t)})$. Conversely, operation under a non-artifactual regime incorporates the information from the observed signals, effectively transforming the inferential process for $\mathbf{x}_t$ into a product of two "experts", one propagating probabilities from $\mathbf{x}_{t-1}$ and one from the current observations.

We note that the step of conditioning on the current regime s_t in order to predict $\mathbf{x}_t$ is required for our task, as we do not have training data for the $\mathbf{x}$-state. Otherwise, one could imagine building a simpler model such as a conditional random field (Lafferty *et al.*, 2001), to predict the $\mathbf{x}$-state directly from the observations. However, in our case, where only labels about the patient's regime are available, this is not possible.

1.3 Learning

We first describe learning in the general SLDS setting. The parameters that need to be learned are: $\{\mathbf{A}^s, \mathbf{Q}^s, \mathbf{C}^s, \mathbf{R}^s\}$. Given training data for each switch setting, these can be learned independently as LDS parameters for each configuration of s. Following Quinn *et al.* (2009) we use an

independent ARIMA model with added observation noise for each channel. Casting such a model into state space form is a standard procedure as described in Brockwell and Davis (2009, sec. 12.1), and amounts into reformulating the parameters of the ARIMA model into the parameters of a state-space model. Once the model is in state space form, $\mathbf{A}^s, \mathbf{Q}^s, \mathbf{C}^s, \mathbf{R}^s$ can be fit according to the maximum likelihood criterion by using numerical optimisation methods (like Newton-Raphson, Gauss-Newton), as presented in Shumway and Stoffer (2000, sec. 2.6) or expectation maximisation (EM) as presented in Ghahramani and Hinton (1996). We note that the vector ARMA (VARMA) representation is used, where for example a one-dimensional AR(p) process can be encoded as a $p + 1$-dimensional VAR(1) process by maintaining a latent state representation of the form $\mathbf{x}_t = [x_t \ x_{t-1} \ ... \ x_{t-p}]$.

In the DSLDS, the same set of parameters needs to be learned. As mentioned in Section 1.2, the assumptions for the DSLDS observation model constrain $\mathbf{C}^s$ to be a binary matrix, whose values are set so as to pick the most recent value $\mathbf{x}_t$ under the VARMA representation. For example, assuming that we are modelling one channel, under a physiological regime, as an AR(2) process, then $\mathbf{C}^s = [1 \ 0 \ 0]$. Under this constrained form of $\mathbf{C}^s$ we obtain the remaining parameters, $\mathbf{A}^s, \mathbf{Q}^s$ and $\mathbf{R}^s$, using the same learning process as the one already described for the case of a general SLDS.

The task of determining the order of the respective ARIMA models is less straightforward. We have followed a practical approach as suggested in Diggle (1990, sec. 6.2). The autocorrelation and partial autocorrelation function (ACF and PACF respectively) of the stationary data (if a time series is not stationary, we make it stationary by successive differencing) were examined to provide an initial estimate of the appropriate model order. A clear cut-off at lag q in the ACF plot is suggestive of an MA(q) process, while a clear cut-off at lag p in the PACF plot is suggestive of an AR(p) process. Clear cut-offs are rare in a real world application, in which case we looked for less clear tail-offs in the PACF and ACF plots. After establishing a small number of potential model orders suggested by these tail-offs, further exploration of the model order around these initial estimates was carried out by calculating the Akaike Information Criterion (AIC) score (Akaike, 1972) for each of these potential model orders, and finally the one with the smallest AIC value was chosen.

1.4 Inference

In this paper we are concerned with the task of computing the distribution $p(s_t, \mathbf{x}_t | \mathbf{y}_{1:t+r})$. According to our proposed model, $p(s_t | \mathbf{y}_{t-l:t+r})$ can be inferred at each time step via a classifier as described in Section 1.1. However, exact inference for $\mathbf{x}_t$ is still intractable. The same lim-

itation as in the case of a standard SLDS applies (Lerner and Parr, 2001): In order to maintain an exact belief over the posterior distribution of $\mathbf{x}_t$ we need to keep track of all the potential combinations of switch variable settings that could have lead us from $\mathbf{x}_{t-1}$ to $\mathbf{x}_t$, making inference scale exponentially with time. An approximation of this distribution can be maintained via the Gaussian Sum algorithm[1] (Alspach and Sorenson, 1972). The idea is that at each time step t we maintain an approximation of $p(\mathbf{x}_t | s_t, \mathbf{y}_{1:t+r})$ as a mixture of J Gaussians. Moving one time step forward will result in the posterior $p(\mathbf{x}_{t+1} | s_{t+1}, \mathbf{y}_{1:t+r+1})$ having KJ components, which are again collapsed to J components. In our experiments we use $J = 1$, which translates into matching moments (up to second order) of the distribution for each setting of s_t, as shown in Murphy (1998). Therefore inference in the DSLDS can be seen as a two-step process, where $p(s_t | \mathbf{y}_{t-l:t+r})$ is inferred by our discriminative classifier, and $p(\mathbf{x}_t | s_t, \mathbf{y}_{1:t+r})$ is inferred according to the Gaussian Sum algorithm.

1.5 Related work

In terms of methodology, our proposed model bears some similarities to the one used by Lu *et al.* (2009). However, their model was used to model spatial relationships and they were only concerned with a binary discrete latent space. In our case, we are concerned with modelling temporal structure and we have a richer and more complex discrete latent space. More importantly, in their work the distribution maintained over the continuous latent space is a single multivariate Gaussian, whereas in our model, as described in the previous section, the belief over the continuous latent space is modelled as a mixture of KJ Gaussians. This allows us to keep track of multiple modes about the belief over a patient's underlying physiology, since this is potentially affected by multiple factors.

In terms of application, our work is mostly similar to the one presented in Quinn *et al.* (2009). The same task of inferring artifactual and physiological processes was considered there. However a generative approach was taken there via the use of an FSLDS. In Lehman *et al.* (2014), a switching vector autoregressive model was used on minute-by-minute heart rate and blood pressure vital signs to provide inputs for a logistic regression classifier with the goal of patient outcome prediction. Also, Nemati *et al.* (2013) propose training a SLDS in a discriminative manner so as to optimize prediction of the **s** sequence given the observations, and apply this to identifying four postural categories under a controlled protocol. Stanculescu *et al.* (2014) use a hierarchical structure in the discrete space of an SLDS motivated by expert knowledge on modelling sepsis. In our work, we use a discriminative SLDS, capable of mod-

---

[1]The Gaussian Sum algorithm is also known as the Generalised Pseudo Bayesian (GPB) algorithm as mentioned in Murphy (1998).

elling both discrete and continuous latent states in a unified framework, applied to two challenging real-world datasets. It yields superior results for state-of-health identification, and maintains at the same time beliefs about a patient's underlying physiology.

2 EXPERIMENTS

In this section we describe experiments on two challenging datasets comprising of patients admitted to ICUs in two different hospitals, namely a neonatal ICU and an adult ICU. We emphasise that it is highly non-trivial to obtain annotations for medical datasets as it requires the very scarce resource of experienced clinicians. Indeed, for the adult ICU, the annotated data are the product of a one-year collaboration with that ICU. Physionet (Goldberger *et al.*, 2000), a freely available medical dataset, is not suitable for our task since the only available time-series annotations are a limited set of life threatening/terminal events, for which identification would not be of practical use in the ICU.

For both datasets, we evaluate the performance of the DSLDS compared to the FSLDS. We also report the performance of an α-mixture of the two models. Note that the FSLDS has been shown in Quinn *et al.* (2009) to achieve superior results compared to more basic models such as a factorial hidden Markov model (FHMM) for the task of condition monitoring in ICUs. We first provide a short description of the various features that were used as input to the state-of-health model as described in Section 1.1, followed by an outline of the main characteristics of the two datasets. We conclude this section by providing results on two tasks: a) inferring a patient's state of health and b) inferring a patient's underlying physiology in the presence of artifact corruption.

2.1 Features & Classifiers

As described in Section 1.1, the estimate of s_t is the output of a discriminative classifier. For both datasets, we found that using a random forest (Breiman, 2001) as our classification method yields the best performance. Suggestions for judicious selection of various tree-construction parameters can be found in Hastie *et al.* (2009, Ch. 15). The Gini index was used as the criterion for splitting nodes for each tree in the random forest. The output of the random forest for a new test point is an average of the predictions produced by each tree, where the prediction of each tree is the proportion of the observations that belong to the positive class in the leaf node in which the test point belongs to. Apart from their high performance, another appealing property of random forests is that they can handle missing observations via the construction of surrogate variables and splits within each decision tree as explained in Hastie *et al.* (2009, sec. 9.2.4).

We use a variety of features to capture interesting temporal structure between successive observations. At each time step, a sliding window of length $l + r + 1$ is computed. For some features we also divide the window into further sub-windows and extract additional features from them. More precisely, the full set of features that are being used are: (i) the observed, raw values of the previous l and future r time steps ($\mathbf{y}_{t-l:t+r}$); (ii) the slopes (calculated by least squares fitting) of segments of that sliding window that are obtained by dividing it in segments of length $(l + r + 1)/k$; (iii) an exponentially weighted moving average of this window of raw values (with a kernel of width smaller than $l + r + 1$); (iv) the minimum, median and maximum of the same segments; (v) the first order differences of the original window; and (vi) differences of the raw values between different channels.

2.2 Neonatal ICU

The first dataset is the one used in Quinn *et al.* (2009)[2]. It comprises 24-hour periods from fifteen neonates admitted to the ICU of the Edinburgh Royal Infirmary, with events of interest annotated by two clinical experts. These annotations include: i) blood sample events (BS), ii) periods during which an incubator is open (IO), iii) core temperature probe disconnections (TD), iv) bradycardias (BR), and v) periods that are clearly not stable but no further identification was made by the clinicians (X). These last cases can be collectively considered as a "none-of-the-above" factor, which is referred to as the X-factor by Quinn *et al.* (2009). More details about the events of interest can be found in the aforementioned work. We used the same parameters for the underlying physiology model as the ones used there.

2.3 Adult ICU

The second dataset comprises data collected from nine adults admitted to the neuro ICU of the Southern General Hospital in Glasgow. An average of 33-hour periods were collected from each of these patients, consisting of measurements recorded on a second-by-second basis for four different channels: heart rate (HR), systolic and diastolic blood pressure (BPsys, BPdia), and systolic intracranial pressure (ICPsys). These data were then annotated by a clinical expert. We give a brief description of the learning process for stability periods and modelled factors, which include blood samples, damped traces (DT), suction events (SC), and the X-factor.

Stable periods correspond to time periods when no annotation occurred from the experts, suggesting that the patient is in a stable condition. In Williams and Stanculescu (2011) it was found that in a similar setting a 15 minute period of stability provides an adequate amount of training data. We

---

[2]The dataset has been anonymised and is available at: *www.cit.mak.ac.ug/staff/jquinn/software.html*

use the same time interval for our experiments. We found that ARIMA(2,1,0) models were adequate for all channels.

An example of a **blood sample** is shown in Figure 4 (bottom). Changes in BPsys and BPdia can be modelled as a four-stage process: i) the blood is diverted to a syringe for blood sampling, which causes an artifactual ramp in the observed measurements. This is similar to the blood sample model described in Quinn *et al.* (2009) and we follow the same approach here. ii) A recalibration stage follows, causing measurements to drop to zero which can be modelled similarly to a dropout event as in Quinn *et al.* (2009). iii) BP measurements continue as a stable period for a brief period. iv) The blood sample is concluded with a flushing event for hygiene purposes which causes a sharp increase in measurements. This stage is modelled as an AR(3) process for both the BPsys and BPdia channels. A total number of 64 blood sample events have been annotated, with an average duration of 1.6 minutes.

During a **suction event**, a flexible catheter is inserted into the airway of the patient to remove secretions that have accumulated over time in their pulmonary system. This event is observed as a significant increase in the values of all observed channels. An AR(2) process models the HR channel, while AR(3) processes were used to model the remaining channels. A total number of 53 suction events have been annotated, with an average duration of 4.3 minutes.

A **damped trace**, an example of which is shown in Figure 4 (top), is usually observed due to blood residues being accumulated in the line used for measuring the blood pressure channel, which leads both BPsys and BPdia to converge to a similar mean value while at the same time the measurements exhibit high variability. Both channels were modelled with AR(3) processes. A total number of 32 damped trace events have been annotated, with an average duration of 14 minutes.

Except for the aforementioned factors which we explicitly model, there are a multitude of other factors present in our training data, corresponding to either known but not yet modelled factors (such as hygiene events, tachycardias etc.) or to unknown factors (clear abnormalities which however have not been identified by the clinicians). We collectively treat those events as unknown and model them according to the X-factor model proposed in Quinn *et al.* (2009). A total number of 278 X-factor events have been annotated, with an average duration of 7.5 minutes. Channels which are unaffected by an artifactual process (as shown in Table 1) are modelled as in the stable case. In every case the parameters of the x-state models were further optimised by EM.

Table 1: Channels affected by different processes for the adult ICU are marked by ●.

	HR	BPsys	BPdia	ICPsys
Blood sample		●	●	
Damped trace		●	●	
Suction	●	●	●	●
X-factor	●	●	●	●

Table 2: Comparison of DSLDS, FSLDS and α-mixture performance for the Neonatal ICU dataset. Optimal value of the α parameter is shown inside parenthesis.

AUC	BS	IO	TD	BR	X
DSLDS	0.98	0.83	0.90	0.94	0.57
FSLDS	0.92	0.87	0.88	0.85	0.66
α-mixture$^{(0.5)}$	0.98	0.89	0.93	0.92	0.67

2.4 Results

For both datasets we compare the performance of the DSLDS and the FSLDS for the task of inferring a patient's state of health. Having obtained estimates s_t for each factor and each time step, we proceed to calculate Receiver Operating Characteristic (ROC) curves for the classification of each factor. We also measure the performance of the models by reporting the Area under each ROC curve (AUC). Plots of the ROC curves, comparing the DSLDS, FSLDS, and an α-mixture of the two models, are shown in Figures 2 and 3.

In the case of the DSLDS, the features described in Section 2.1 involve a number of hyperparameters that need to be chosen. Fitting them with a standard cross-validation (CV) scheme when data are not abundant poses a non-negligible risk of overfitting. As is shown in Varma and Simon (2006), using CV to evaluate performance of a model when the model's hyperparameters have been themselves tuned using CV can lead to an optimistic bias of the estimate of the true performance. In that same work, a nested CV approach is shown to yield an almost unbiased estimate of the true performance, which we also follow in our experiments. In the outer loop the data are partitioned into P disjoint test sets. After choosing one of these partitions, the rest of the data are used in the inner loop in a standard CV setup to select the hyperparameters. The hyperparameters which yielded the highest performance (average cross-validated AUC across factors in our case) in the inner loop are then used to estimate the performance of the model on the partition (test set) in the outer loop. This process is repeated P times, once for each partition in the outer loop. For both datasets, we use leave-one-patient-out CV for the

inner loop and 3-fold CV for the outer loop. In the inner loop, we perform a grid search over hyperparameters in the following sets: a) number of trees of random forest classifiers in $\{10, 25, 50, 100, 200\}$; b) l in $\{4, 9, 14, 19, 29, 49\}$; c) r in $\{0, 5, 10\}$. The sub-segments lengths (for slope features) were always set to $\max\{5, (l + r + 1)/5\}$ and the kernel widths (for moving average features) were always set to $\max\{5, (l + r + 1)/5\}$.

In the case of the FSLDS, it is not necessary to follow the same procedure. Using the AIC score, as shown in Section 1.3, for choosing the orders of the ARIMA processes (which constitute the model's hyperparameters) avoids potential overfitting by penalising the model's likelihood as the parameters grow. We therefore use 3-fold CV to evaluate the FSLDS's performance.

To evaluate the α-mixture model, we have chosen the optimal α value as the one that maximises the average AUC across factors, via 3-fold CV. This also allowed us to explore the behaviour of the model as a function of α for both datasets.
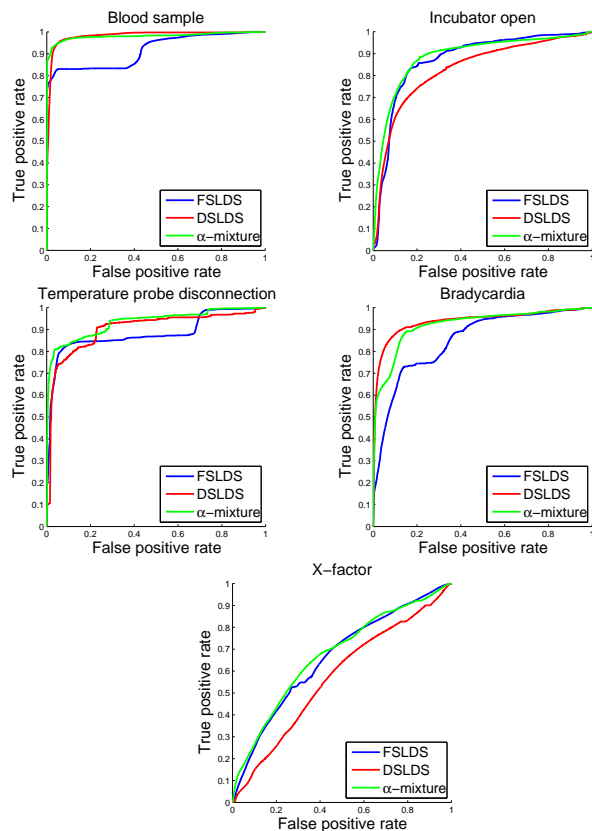

Figure 2: ROC curves per modelled factor in the case of the neonatal ICU.

2.4.1 Neonatal ICU

In the case of the neonatal ICU we compare the two models on the full set of annotated factors reported in Quinn *et al.* (2009). The results are shown in Table 2[3]. The DSLDS outperforms the FSLDS in three out of the four clinically identified factors. The difference in favour of the DSLDS is clear for bradycardias and blood samples, but less pronounced for core temperature disconnections. The FSLDS achieves slightly higher performance in the case of the incubator open factor, and clearly outperforms the DSLDS in the case of the X-factor. The FSLDS models the presence of outliers by the inclusion of an extra factor, which is essentially governed by the same parameters as stability with the only difference being that the system noise covariance is an inflated version of the respective covariance of the stability dynamics (for more details, see Quinn *et al.*, 2009). Such an approach has the potential to address the issue of outlier detection in a more general and thus more satisfactory way. In the case of the DSLDS, our approach is to collectively treat all abnormal events, other than the ones attributed to known factors, as an "X-class" and build a binary classifier to distinguish that class. As the training datapoints for this class are highly inhomogeneous in terms of shared discriminative features, and test points belonging to the X-class may not exhibit a high degree of similarity to the training set, it is not surprising that the DSLDS may perform rather poorly for the X-factor. However, by considering an α-mixture of the two models, we can combine the discriminative power of the DSLDS for known factors with the increased performance of the FSLDS for the X-factor, thus achieving a higher performance (bottom line of Table 2) compared to considering the two models separately. The behaviour of the α-mixture model as a function of α is shown in Figure 5 (top). The optimal α-mixture ($\alpha = 0.5$) yields the best average AUC across factors (in fact, $\alpha = 0.5$ yields optimal performance for each factor separately except bradycardia, where it is almost optimal) compared to all other considered α values and also outperforms the DSLDS and the FSLDS in all cases except for the bradycardia factor, where the DSLDS performs slightly better.

2.4.2 Adult ICU

In the case of the adult ICU, inferences for two example events are shown in Figure 4. In the top, a damped trace event is shown, which lasts for almost one hour before being resolved by a flushing event (spiking of both channels). The DSLDS accurately identifies the damped trace event,

---

[3]The FSLDS results were obtained using code provided by Quinn *et al.* (2009) with the same parameters as the ones mentioned there. The results are very close with the exception of the core temperature disconnection factor (for which the reported AUC in Quinn *et al.* (2009) was 0.79, while we obtained a value of 0.88), and the blood sample factor (for which the reported AUC in Quinn *et al.* (2009) was 0.96, while we obtained a value of 0.92).
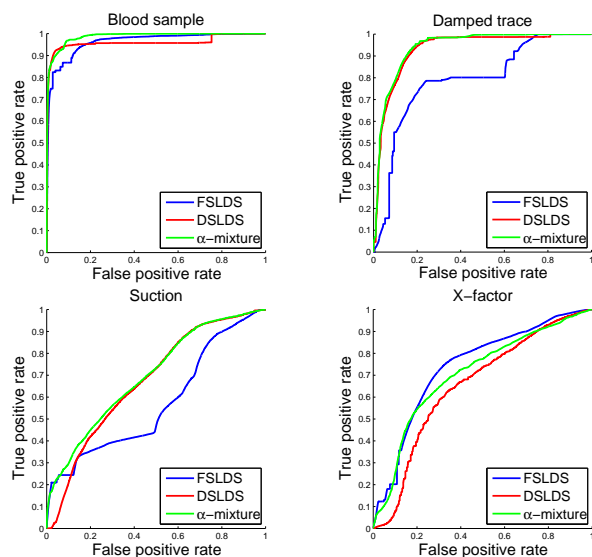
Figure 3: ROC curves per modelled factor in the case of the adult ICU.

Figure 4: Example of DSLDS and FSLDS inferences for a damped trace event (top) and a blood sample event (bottom).

Table 3: Comparison of DSLDS, FSLDS and α-mixture performance for the Adult ICU dataset. Optimal value of the α parameter is shown inside parenthesis.

AUC	BS	DT	SC	X
DSLDS	0.96	0.93	0.67	0.65
FSLDS	0.95	0.79	0.57	0.74
α-mixture$^{(0)}$	0.99	0.94	0.70	0.71

while the FSLDS fails totally to detect it, but hypothesises several incorrect blood sample events. In the bottom panel a blood sample event is shown, where the multiple stages are clearly visible. The event starts with two artifactual ramps, followed by a flushing, a zeroing, and finally with another flushing. This is slightly different than the description we have already given, but slight deviations from the standard protocol due to human error is to be expected. In this case, both models manage to capture the event in a generally satisfactory manner. Summary results are reported in Table 3. The DSLDS outperforms the FLSDS on all of the known factors. The damped trace and suction events particularly are characterised by high variability which is hard to capture with a generative process. However, simple discriminative features are able to capture them with higher accuracy. As was expected, the FSLDS achieves a higher AUC for the X-factor. Again, the optimal α-mixture ($\alpha = 0$) outperforms the DSLDS and the FSLDS in all cases except for the X-factor, where the FSLDS achieves a slightly higher AUC. Contrary to the neonatal ICU dataset, as shown in Figure 5 (bottom) there are alternative α values
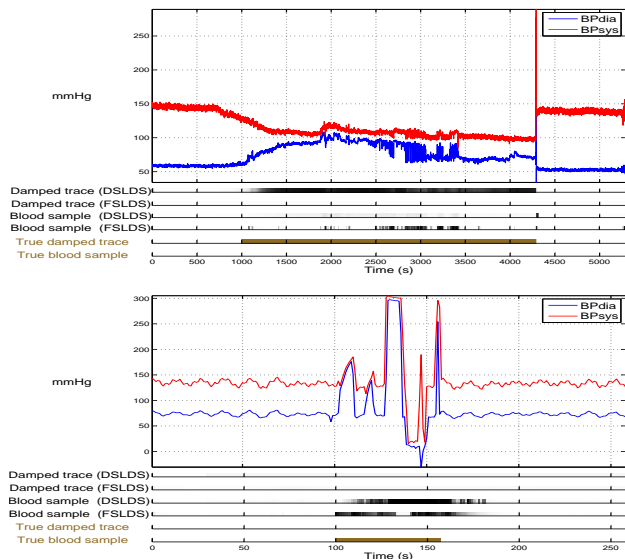
which can yield higher AUC across different factors. For example, an X-factor AUC value of 0.76 can be obtained by setting $\alpha = 5$. However, apart from the superior (on average) performance of the α-mixture, another appealing property is that α could be treated as a user-tunable parameter. In a practical setting, the model could be preset with the optimal α value, but a clinician could decide, for example, to make the model focus on maximising its predictive performance on the X-factor (or some important physiological factor like bradycardia) to the potential detriment of other factors. Then the model could adjust its α parameter in real-time based on training data results to maximise its performance on the desired factor.

2.4.3 Inference for x-state

Finally, Figure 6 shows the inferred distribution of underlying physiology during a blood sample taken from a neonate for both models. In both cases, estimates are propagated with increased uncertainty under the correctly inferred artifactual event. Note a small difference at the start of the event: The DSLDS partially identifies the event causing an increase in uncertainty, while the FSLDS (incorrectly) identifies this part as stable and thus its x-state update exhibits lower uncertainty. Maintaining an estimate of the underlying vital signs in the presence of artifacts can then be used for data imputation. Another use, which has been deemed important by our clinical experts, is that such an estimate can help doctors maintain an approximate view of a patient's underlying physiology during artifactual events that would otherwise completely obscure a patient's vital signs. This can be crucial during treatment of a patient un-
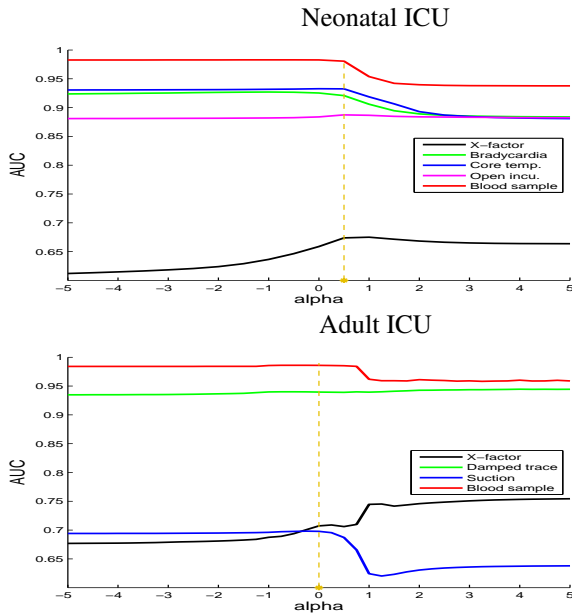
Figure 5: Performance of the α-mixture models as a function of α ($step = 0.25$) for the Adult ICU (top) and the neonatal ICU dataset (bottom). The asterisk marks the optimal value for α.

der critical conditions, such as the ones found in an ICU.

3 DISCUSSION

We have presented a discriminative approach for the very important application of patient monitoring in ICUs. We show that our new approach is able to outperform the previous generative approach used for the same task in most of the investigated cases. We also show that an α-mixture of the two approaches yields better results than either model separately. In our approach we have assumed that the prediction of the switching variable factorises over the state space. However, one could use a structured output model to predict the joint distribution of different factors. Finally, another issue is the lack of explicit temporal continuity in the s-chain. Implicitly, this is handled by the feature construction process. However, a future direction could be to establish a Markovian connection on the s-chain too and compare with our current approach.
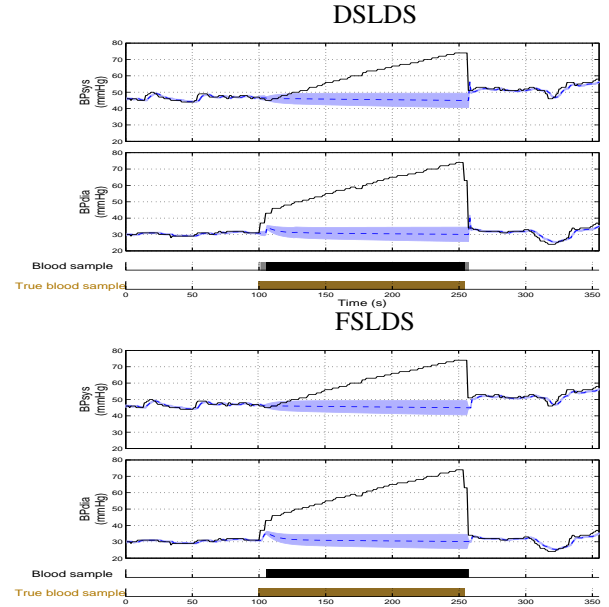
Figure 6: Example of the inferred underlying physiology in the presence of a blood sample in the case of the DSLDS (top) and the FSLDS (bottom). The solid line corresponds to the actual observations, while the estimated true physiology is plotted as a dashed line with the shaded area indicating two standard deviations.

References

Akaike, H. (1972). Information theory and an extension of the maximum likelihood principle. *2nd Int. Symp. Information Theory, Supp. to Problems of Control and Information Theory*, pages 267–281.

Alspach, D. L. and Sorenson, H. W. (1972). Nonlinear Bayesian Estimation Using Gaussian Sum Approximations. *Automatic Control, IEEE Transactions on*, **17**(4), 439–448.

Amari, S.-i. (2007). Integration of Stochastic Models by Minimizing α-Divergence. *Neural Computation*, **19**(10), 2780–2796.

Breiman, L. (2001). Random Forests. *Machine Learning*, **45**(1), 5–32.

Brockwell, P. J. and Davis, R. A. (2009). *Time Series: Theory and Methods*. Springer.

Diggle, P. (1990). *Time Series: A Biostatistical Introduction*. Oxford University Press.

Ghahramani, Z. and Hinton, G. E. (1996). Parameter Estimation for Linear Dynamical Systems. Technical report, Technical Report CRG-TR-96-2, University of Totronto, Dept. of Computer Science.

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. (2000). Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. *Circulation*, **101**(23), e215–e220.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.

Lafferty, J., McCallum, A., and Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning (ICML)*.

Lehman, L., Adams, R., Mayaud, L., Moody, G., Malhotra, A., Mark, R., and Nemati, S. (2014). A physiological time series dynamics-based approach to patient monitoring and outcome prediction. *Biomedical and Health Informatics*.

Lerner, U. and Parr, R. (2001). Inference in Hybrid Networks: Theoretical Limits and Practical Algorithms. In *Proceedings of the Seventeenth conference on Uncertainty in Artificial Intelligence (UAI)*, pages 310–318. Morgan Kaufmann Publishers Inc.

Lu, W.-L., Murphy, K. P., Little, J. J., Sheffer, A., and Fu, H. (2009). A Hybrid Conditional Random Field for Estimating the Underlying Ground Surface from Airborne LiDAR Data. *Geoscience and Remote Sensing, IEEE Transactions on*, **47**(8), 2913–2922.

McCallum, A., Freitag, D., and Pereira, F. C. (2000). Maximum Entropy Markov Models for Information Extraction and Segmentation. In *International Conference on Machine Learning (ICML)*, pages 591–598.

Murphy, K. P. (1998). Switching Kalman Filters. Technical report, U.C.Berkeley.

Nemati, S., Lehman, L.-W., and Adams, R. P. (2013). Learning outcome-discriminative dynamics in multivariate physiological cohort time series. In *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, pages 7104–7107. IEEE.

Quinn, J. A., Williams, C. K., and McIntosh, N. (2009). Factorial Switching Linear Dynamical Systems applied to Physiological Condition Monitoring. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **31**(9), 1537–1551.

Särkkä, S. (2013). *Bayesian Filtering and Smoothing*. Cambridge University Press.

Shumway, R. H. and Stoffer, D. S. (2000). *Time Series Analysis and Its Applications*. Springer New York.

Stanculescu, I., Williams, C. K., and Freer, Y. (2014). A Hierarchical Switching Linear Dynamical System Applied to the Detection of Sepsis in Neonatal Condition Monitoring. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 752–761. AUAI Press.

Varma, S. and Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, **7**(1), 91.

Williams, C. K. and Stanculescu, I. (2011). Automating the Calibration of a Neonatal Condition Monitoring System. In *Artificial Intelligence in Medicine*, pages 240–249. Springer.