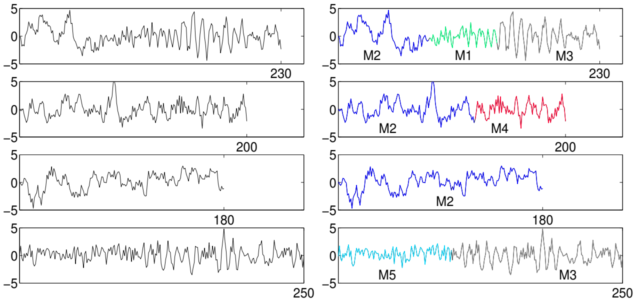
A Bayesian Approach to Switching Linear Gaussian State-Space Models for Unsupervised Time-Series Segmentation

Silvia Chiappa
Max-Planck Institute for Biological Cybernetics,
Spemannstraße 38, 72076 Tübingen, Germany
silvia.chiappa@tuebingen.mpg.de

Abstract

Time-series segmentation in the fully unsupervised scenario in which the number of segment-types is a priori unknown is a fundamental problem in many applications. We propose a Bayesian approach to a segmentation model based on the switching linear Gaussian state-space model that enforces a sparse parametrization, such as to use only a small number of a priori available different dynamics to explain the data. This enables us to estimate the number of segment-types within the model, in contrast to previous non-Bayesian approaches where training and comparing several separate models was required. As the resulting model is computationally intractable, we introduce a variational approximation where a reformulation of the problem enables the use of efficient inference algorithms.



(a) (b)

Figure 1. (a) Four uni-dimensional time-series. (b) Segmentation into five underlying dynamical regimes.

1. Introduction

This paper introduces a model for segmenting a set of time-series in the fully unsupervised scenario where the number of segment-types is a priori unknown. As an example, consider the four uni-dimensional unsegmented time-series plotted in Fig. 1 (a). Our aim is to discover the dynamical structure underlying these time-series. An analysis of the data reveals that there are five different underlying dynamical regimes. As shown in Fig. 1 (b), the first time-series goes through three dynamical regimes (M2, M1 and M3). Similarly, the other time-series exhibit changes in their dynamical properties through time. More generally, given a collection of *multi-dimensional* time-series, we are interested in discovering the set of underlying dynamical regimes, and identifying for each series which regime operates at any particular time. This may be viewed as *unsupervised segmentation* of time-series, with an automatic discovery of the number of segment-types.

The approach that we take is to consider a generative probabilistic temporal model of the observations, namely a switching Linear Gaussian State-Space Model (LGSSM) [6] where each underlying dynamical regime is modeled by a different set of parameters. Segmentation is then be performed based on an application of Bayes' rule to infer which set of parameters was most likely to have generated the observations at any particular time. This approach has successfully been used in several application domains such as finance, speech processing, modeling of human motion and medicine [1, 10, 11, 9].

In our fully unsupervised scenario, the underlying number of different parameter sets is not known in advance and needs to be estimated. A common approach to solve this model structure selection problem is to train a separate model for each possible structure, and then select the model that optimizes a trade-off between accuracy and complexity, as measured by the Bayesian Information Criterion for example. The drawback of this approach is that training many separate models may lead to a large computational overhead.

A computationally less expensive alternative for penalizing model complexity is offered by Bayesian approaches, where the model parameters are treated as random variables and integrated out to yield the marginal likelihood of the data. The parameter prior distributions can be chosen such as to enforce a sparse representation, i.e., to select a subset of the available parameters that explains the data well by making the remaining parameters inactive. As a result, structure selection can be achieved within the model without the need to compare several models.

A Bayesian approach to the segmentation model based on the switching LGSSM poses considerable challenges due to intractability issues. We describe how these issues can be addressed using a variational approximation in which the problem is formulated such that efficient inference methods can be used[1].

The rest of the paper is organized as follows. In Section 2 we introduce the basic segmentation model, while in Section 3 we describe our Bayesian approach and discuss a variational approximation method to deal with intractability issues. We then give some demonstrations in Section 4 and draw some conclusions in Section 5.

2. Time-Series Segmentation

We start our exposition with the simpler case of segmenting a single multi-dimensional time-series $v_{1:T} \equiv \{v_1, \dots, v_T\}$, with $v_t \in \Re^V$. To model the situation that a possible different dynamical system operates at any particular time, we consider a switching Linear Gaussian State-Space Model (LGSSM). This model assumes that $v_{1:T}$ is generated from a latent Markovian linear dynamical system on hidden states $h_{1:T}$, with $h_t \in \Re^H$, according to[2]

$$\begin{aligned} v_t &= B^{z_t} h_t + \eta_t^{v,z_t}, & \eta_t^{v,z_t} &\sim \mathcal{N}\left(\mathbf{0}_V, \Sigma_V^{z_t}\right) \\ h_t &= A^{z_t} h_{t-1} + \eta_t^{h,z_t}, & \eta_t^{h,z_t} &\sim \mathcal{N}\left(\mathbf{0}_H, \Sigma_H^{z_t}\right), \end{aligned}$$

where $z_t \in 1, \dots, K$ is an indicator variable that controls which of K sets of hidden dynamics and emission parameters is used at time t. To simplify the notation, unlike the more standard assumption of a Markovian dynamics, we assume that the indicator variables fully factorize, i.e., $p(z_{1:T}) = \prod_t p(z_t)$, and that $p(z_t)$ is time independent and parameterized by a vector π such that $p(z_t = k|\pi) = \pi_k$.

The joint density[3] $p(z_{1:T}, h_{1:T}, v_{1:T}|\Theta^{1:K}, \pi)$ admits the following factorization

$$\prod_{t=1}^{T} p(v_t|z_t, h_t, \Theta^{1:K}) p(h_t|z_t, h_{t-1}, \Theta^{1:K}) p(z_t|\pi),$$

where $\Theta^{1:K} = \{A^{1:K}, B^{1:K}, \Sigma_H^{1:K}, \Sigma_V^{1:K}, \mu^{1:K}, \Sigma^{1:K}\}$ and

$$\begin{aligned} p(v_t|z_t = k, h_t, \Theta^{1:K}) &= \mathcal{N}\left(B^k h_t, \Sigma_V^k\right) \\ p(h_t|z_t = k, h_{t-1}, \Theta^{1:K}) &= \mathcal{N}\left(A^k h_{t-1}, \Sigma_H^k\right). \end{aligned}$$

Unlike the standard LGSSM with fixed parameters over time, performing inference in the switching LGSSM is intractable, as for example the filtered state estimate $p(h_t|v_{1:t}, \Theta^{1:K})$ is a mixture of Gaussians with an exponential explosion of mixtures with time. To deal with this problem, several approximation methods have been introduced in the past. A particularly efficient approximation based on a novel form of Gaussian sum smoother has recently been proposed in [2].

For the case of N time-series[4] $v_{1:T}^{1:N}$, we introduce a set of indicator variables $z_{1:T}^{1:N}$ $(p(z_{1:T}^{1:N}) = \prod_{n,t} p(z_t^n))$, where each indicator $z_t^n \in \{1, \dots, K\}$ denotes which hidden dynamics and emission parameters generated observation v_t^n. We then form the likelihood of all observations

$$p\left(v_{1:T}^{1:N}|\Theta^{1:K}, \pi\right) = \prod_{n=1}^{N} p\left(v_{1:T}^{n}|\Theta^{1:K}, \pi\right). \qquad (1)$$

Instead of a standard approach were the optimal values of $\Theta^{1:K}$ and π are learned by maximizing the likelihood, here we take a Bayesian approach in which the parameters are treated as random variables and integrated out from Eq. (1).

3. Bayesian Approach

In our Bayesian approach we define prior distributions $p(\Theta^{1:K}|\hat{\Theta}^{1:K})$ and $p(\pi|\gamma)$, where $\hat{\Theta}^{1:K}$ and γ are the associated hyperparameters. We then form the marginal likelihood $p(v_{1:T}^{1:N}|\hat{\Theta}^{1:K}, \gamma)$ as

$$\int_{\Theta^{1:K}, \pi} p(v_{1:T}^{1:N}|\Theta^{1:K}, \pi) p(\Theta^{1:K}|\hat{\Theta}^{1:K}) p(\pi|\gamma). \qquad (2)$$

For the hyperparameters we either take a type-II maximum likelihood approach [8], where the optimal values are found by maximizing Eq. (2), or define additional prior distributions and integrate them out from Eq. (2). Assignment to a certain dynamical regime is then performed by computing $\arg\max_k p(z_t^n = k|v_{1:T}^{1:N}, \hat{\Theta}^{1:K}, \gamma)$.

Prior Distributions on $\Theta^{1:K}$

As prior distributions for $\Theta^{1:K}$ we define zero-mean Gaussians on the elements of A^k and on the columns of B^k as follows[5]

[1] A similar approximation approach to a constrained non-fully Bayesian switching LGSSM has independently been introduced in [9] in the context of supervised speech processing.

[2] $\mathcal{N}(m, S)$ denotes a Gaussian with mean m and covariance S, while $\mathbf{0}_X$ denotes an X-dimensional zero vector. $h_1 \sim \mathcal{N}\left(\mu^{z_1}, \Sigma^{z_1}\right)$.

[3] $X^{1:K}$ is a shorthand for $X^1, \dots, X^K$.

[4] To simplify the notation, we assume that all time-series are of equal length T.

[5] We omit the dependency on k. $\left[X^{-1}\right]_{ii}$ denotes the ii-th element of the matrix X^{-1}, while X_j denotes the j-th column of the matrix X. The dependency of the priors on Σ_H and Σ_V is chosen to render the implementation feasible.

$$p\left(A|\alpha,\Sigma_H^{-1}\right) = \prod_{i,j=1}^{H} \frac{\alpha_{ij}^{1/2}}{\sqrt{2\pi\left[\Sigma_H\right]_{ii}}} e^{-\frac{\alpha_{ij}}{2}\left[\Sigma_H^{-1}\right]_{ii} A_{ij}^2}$$

$$p\left(B|\beta,\Sigma_V^{-1}\right) = \prod_{j=1}^{H} \frac{\beta_j^{V/2}}{\sqrt{|2\pi\Sigma_V|}} e^{-\frac{\beta_j}{2} B_j^{\mathsf{T}}\Sigma_V^{-1} B_j},$$

where α^k and β^k are hyperparameters. The use of type-II maximum likelihood with this type of priors has the effect of penalizing complex models and gives rise to a sparse parametrization. More specifically, during learning some α_{ij}^k and β_j^k get close to infinity, whereby (the posterior distribution of) A_{ij}^k and B_j^k get close to zero. As an alternative approach to type-II maximum likelihood, we can define Gamma distributions on α_{ij}^k and β_j^k. When the hyperparameters of the Gamma distributions are set to zero these two approaches are equivalent, whilst for other values the latter approach penalizes model complexity less severely. A discussion on this pruning effect can be found in [12][6].

For modeling general or diagonal inverse covariances Σ_H^{-1}, Σ_V^{-1}, and Σ^{-1} we use Wishart or Gamma distributions respectively, while we define a zero-mean Gaussian prior for μ.

These choices for the prior distributions render our Bayesian treatment feasible (see [5] for more details).

Prior Distribution on π

As prior for π, we define a symmetric Dirichlet distribution

$$p(\pi|\gamma) = \frac{\Gamma(\gamma)}{\Gamma(\gamma/K)^K} \prod_{k=1}^{K} \pi_k^{\gamma/K-1},$$

where $\Gamma(\cdot)$ is the Gamma function. This distribution is conjugate to the multinomial, which greatly simplifies our Bayesian treatment of the model[7].

Model Intractability

The joint distribution of all observations $p(v_{1:T}^{1:N}|\hat{\Theta}^{1:K},\gamma)$ is given by

$$\int_{\Theta^{1:K},\pi} p(\pi|\gamma)\prod_k p(\Theta^k|\hat{\Theta}^k)\sum_{z_{1:T}^{1:N}}\prod_{n,t} p(v_t^n|z_t^n,\Theta^{1:K})p(z_t^n|\pi).$$

Due to the integration over $\Theta^{1:K},\pi$ and dynamics switching this distribution is intractable. To deal with the first intractability issue we use a variational approximation method, where the biggest challenge is to perform inference on the hidden state and indicator variables. In [3], we showed how to achieve the same task in the Bayesian LGSSM by reformulating the problem such that any inference method developed for the (non-Bayesian) LGSSM could be used. A similar strategy enables us to perform the required inference in this model by employing any approximate inference method developed for the (non-Bayesian) switching LGSSM. This automatically provides a solution to the second intractability problem.

Variational Approximation

In our variational approximation, we introduce a new distribution q such that[8]

$$q\left(h_{1:T}^{1:N}|z_{1:T}^{1:N},\Theta^{1:K}\right) = q\left(h_{1:T}^{1:N}|z_{1:T}^{1:N}\right)$$
$$q\left(z_{1:T}^{1:N},\Theta^{1:K},\pi\right) = q\left(z_{1:T}^{1:N}\right)q\left(\Theta^{1:K},\pi\right).$$

That is, we assume that the posterior distribution of the hidden states are decoupled from $\Theta^{1:K}$ given $z_{1:T}^{1:N}$, and furthermore that $z_{1:T}^{1:N}$ are decoupled from the $\Theta^{1:K}$ and π. Nevertheless, the dependence of the hidden states on the indicators is retained. The aim is to find the q that is closest to the original distribution p. This can be achieved by minimizing the KL divergence between $q(h_{1:T}^{1:N},z_{1:T}^{1:N},\Theta^{1:K},\pi)$ and $p(z_{1:T}^{1:N},h_{1:T}^{1:N},\Theta^{1:K},\pi|v_{1:T}^{1:N},\hat{\Theta}^{1:K},\gamma)$, which is equivalent to maximizing a tractable lower bound on the log-likelihood $\log p(v_{1:T}^{1:N}|\hat{\Theta}^{1:K}\!\!,\gamma) \geq \mathcal{F}(\hat{\Theta}^{1:K},\gamma,q)$. We thus proceed by iteratively maximizing the lower bound with respect to the q distributions for fixed hyperparameters $\hat{\Theta}^{1:K}$ and γ and vice-versa until no further improvement is found. Observation v_t^n is then placed in the most likely dynamical regime by computing $\arg\max_k q(z_t^n = k)$. The resulting recursive updates are given in the Appendix.

Inference on the Hidden State and Indicator Variables

The final observation assignment to the most likely dynamical regime and the updates for the parameter distributions require the non-trivial estimation of $q(z_t^n)$, $\langle h_t^n\rangle_{q(h_t^n|z_t^n)}$ and $\left\langle h_t^n h_{t-1}^n\right\rangle_{q(h_{t-1:t}^n|z_t^n)}$, where $\langle\cdot\rangle_q$ denotes expectation with respect to q. To solve this task we reformulate the optimal $q(h_{1:T}^n,z_{1:T}^n)$ as proportional to the joint distribution of a (non-Bayesian) switching LGSSM. This allows us to use the efficient approximate inference method developed for the switching LGSSM in [2]. The details of this approach are given in the Appendix.

Automatic Selection of Number of Segment-Types

As we have seen above, our prior distributions enforce pruning of elements of $A^{1:K}$ and $B^{1:K}$. In particular for certain k all elements of A^k and B^k are pruned out from the model, such that the k-th dynamical regime becomes inactive ($q(z_t^n = k) = 0$ for all t,n). This means that, even

[6] In our exposition we take a type-II maximum likelihood approach. The case of Gamma priors is discussed in [4].

[7] For Markovian indicator variables with $p(z_t^n = j|z_{t-1}^n = i,\pi^i) = \pi_j^i$, $p(z_1^n = j|\pi) = \pi_j$, we can define a Dirichlet distribution for each vector π^i and for π (details are given in [4]).

[8] We omit conditioning on the observations and hyperparameters.

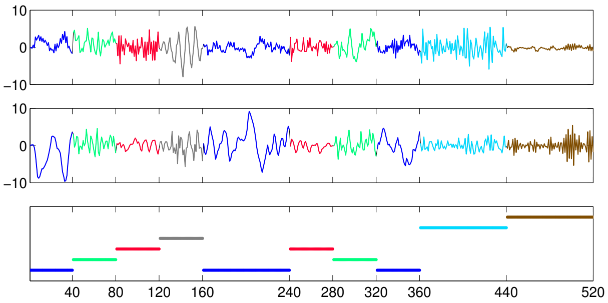
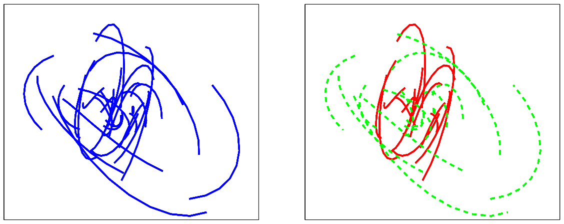


Figure 2. Segmentation of a two-dimensional time-series generated by a switching LGSSM with six dynamical regimes. Our model correctly identifies the underlying regimes.



(a) Unclustered Trajectories (b) Clustered Trajectories

Figure 3. (a) Thirty two-dimensional time-series resulting from the dynamics of two different LGSSMs. (b) Underlying clustering correctly identified by our model.

if we initialize the model with a fixed number of dynamical regimes K, our Bayesian approach ensures that the unnecessary regimes are pruned out during training. Thus the selection of the number of segment-types is automatically performed within the model.

4. Demonstrations

In this section we present some examples on how our model performs on artificial data and on sequences generated by video recordings of human movements. As we expect the time-series to be in the same dynamical regime for a certain interval of time, we force the value of z_t^n to be constant for 10 contiguous time-steps (imposing such a constraint in the model is straightforward). We also present an example of the extreme situation where each time-series is forced to be in the same dynamical regime for all time-steps ($z_t^n = z^n$). In this case inference in the hidden variables of the model can be performed exactly, since there is no longer switching in the dynamics, and our segmentation method reduces to a mixture model for more standard time-series clustering [13, 7], assigning time-series to the same cluster if they display similar dynamics.

Artificial Time-Series Segmentation

We first consider the task of segmenting a *single* multi-dimensional time-series into different dynamical regimes. As an illustrative example, we generated a sequence of length $T = 520$, using a switching LGSSM with output dimensionality $V = 2$, hidden state dimensionality $H = 3$, and $K = 6$ dynamical regimes. The time-series and its segmentation are plotted Fig. 2. We trained our model with $K = 10, H = 6$ and initial random parameters. Thanks to the Bayesian approach enforcing sparsity, the model pruned out four dynamical regimes and correctly segmented the time-series into the six underlying regimes with deterministic posterior ($q(z_t^n = k) = 0/1$).

As a demonstration for the more general problem of segmenting *several* multi-dimensional trajectories, we generated eight sequences of length $T = 300$, using a switching LGSSM with $V = 2$, $H = 4$ and five dynamical regimes (M1-M5). The segmentation is given on the right. We trained our segmentation model with $K = 10$, $H = 7$ and initial random parameters. The model found five active dynamical regimes and segmented the time-series in agreement with the underlying segmentation.

Time-Series

M1	M2	M3		M5	M1
M1	M2	M3		M5	
M1	M2	M1	M2	M3	M2
M1	M2	M1	M2	M3	M2
M1	M4			M3	M4
M1	M4			M3	M4
M1	M4			M5	
M1	M4			M5	M3

Time

Artificial Time-Series Clustering

As last example on artificially generated data, we consider a more standard time-series clustering task where each time-series is in a single dynamical regime for all time-steps. This can be explicitly imposed in our model with the constraint $z_t^n = z^n$.

We generated 30 time-series of dimension $V = 2$ and length $T = 10$, using a LGSSM with two different sets of hidden dynamics and emission parameters and hidden state dimensionality $H = 4$. The parameters were chosen so that all time-series looked visually dissimilar, see Fig. 3 (a). We trained our model with six clusters (sets of hidden dynamics and emission parameters), $H = 10$ and initial random parameters. The model pruned out four clusters and perfectly grouped the data into the two underlying clusters (see Fig. 3 (b)).

More examples on standard time-series clustering, including the case of missing observations, are given in [5].

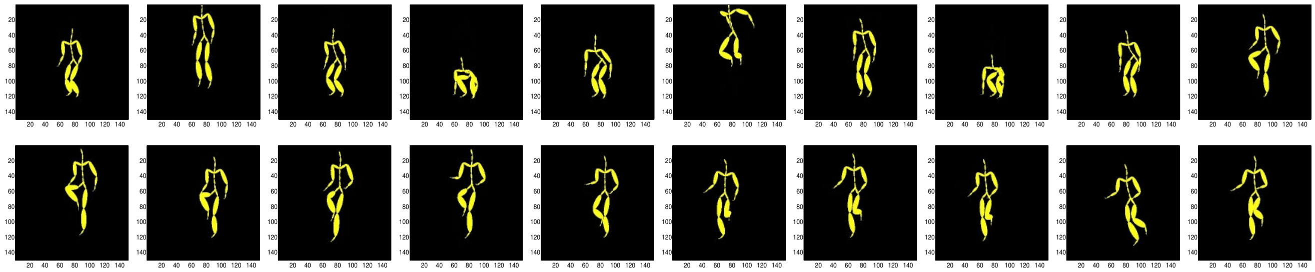


Figure 4. From the top, left to right: sequence of movements for each of the following actions: low jumping up and down (images 1-3), high jumping up and down (images 4-8), hopping on the left foot on the spot and moving forward and backward (images 10-15), hopping on the right foot on the spot and moving laterally (images 16-20). Image 9 represents a transition movement.

Human Body Movement Time-Series Segmentation

In this section, we show an application of our model to the video recordings of human movements. Our aim is to see whether the model can identify dynamically similar segments of motions.

The data analyzed is part of the CMU Graphics Lab Motion Capture Database, and contains recordings from a person performing several repetitions of the following actions for approximatively 17 seconds[9]: (1) low jumping up and down, (2) high jumping up and down, (3) hopping on the left foot, initially on the spot and later moving forward/backward, (4) hopping on the right foot, initially on the spot and later moving laterally. In Fig. 4 we show a sequence of movements extracted from the generated video for each of these actions.

Motions are captured using markers positioned at several places on the body, giving rise to a 62-dimensional time-series of length 495. We selected the 24 dimensions corresponding to the markers positioned on the middle and lower parts of the body, disregarding the ones on the head, neck and arms. The time-series is displayed Fig. 5.

We trained our model with $K = 15$ dynamical regimes, hidden state dimensionality $H = 7$ and initial random parameters. The model found six active regimes. In Fig. 5/row 13 we show the segmentation obtained by our model, while in Fig. 5/row 14 we show the following manual segmentation performed by analyzing the video: first/seventh segments (1-102/446-495): low jumping up and down, second segment (103-212): high jumping up and down, third/fourth segments (213-280/281-347): hopping on the left foot on the spot and moving forward/backward respectively, fifth/sixth segments (348-374/375-445): hopping on the right foot on the spot and moving laterally respectively. Notice that, due to its manual nature, this segmentation is not precise around the boundaries. Furthermore, intermediate movements performed to switch from one action to the next are incorrectly assigned to one of the four actions (see for example image 9 in Fig. 5).

Our model identified the four main actions, but did not discriminate different ways of performing hopping on one foot (on the spot vs moving). This seems reasonable since the video reveals a small difference in the respective body movements. On the other hand, our model identified different dynamical regimes underlying the action high jumping up and down, which, as we can see from Fig. 4, indeed requires very different movements of the legs. Finally, our model considered the last part of the first manual segment as a different dynamical regime. An analysis of the video reveals that, at that time, the person performs a different movement to switch action.

5. Conclusions

We introduced a generative probabilistic temporal model for segmenting a set of time-series when the number of segment-types is a priori unknown. The model assumes that the time-series are generated by a switching Linear Gaussian State-Space Model where a different set of parameters represents each underlying dynamical regime. A Bayesian treatment of the model enforces to obtain a sparse parametrization, such that only a small number of a priori available dynamical systems is used to explain the data. To deal with the resulting model intractability issues we described a variational approximation, designed to retain many of the statistical dependencies in the model and to enable the use of efficient inference algorithms on the hidden variables.

Appendix

Below we discuss the updates obtained by maximizing the lower bound on the log-likelihood. For space reasons,

[9] The .avi file can be downloaded at http://mocap.cs.cmu.edu, Subject #49, Trial #2. We disregarded the initial and last part of the movie where the subject is not moving or performs a non well defined action.

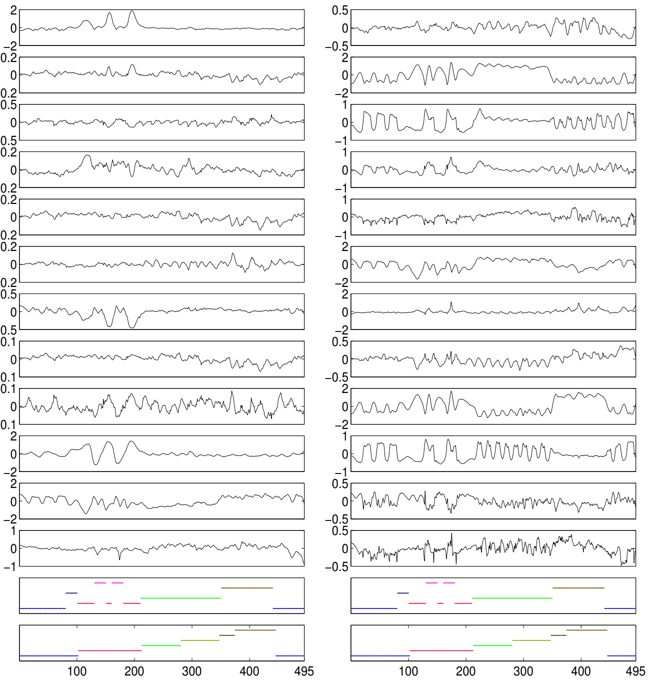


Figure 5. Top Rows 1-12: 24-dimensional time-series generated by video recording a person performing four different actions. Bottom Row 13: Segmentation into different dynamical regimes obtained by our model. Bottom Row 14: Segmentation into different actions and sub-actions performed manually by visual inspection of the video.

we consider only diagonal noise covariances and omit the updates for $\mu^{1:K}$ and $\Sigma^{1:K}$ distributions and hyperparameters. The missing updates and the case of Gamma priors on $\alpha^{1:K}, \beta^{1:K}$ and Markovian indicators can be found in [4].

Hidden State and Indicator Updates

The hidden state and indicator distribution update for each time-series is given by

$$q\left(h^n_{1:T}, z^n_{1:T}\right) \propto e^{\langle \log p(z^n_{1:T}|\pi)\rangle_{q(\pi)}} \qquad (3)$$
$$e^{\left\langle \log p\left(v^n_1, h^n_1|\Theta^{z^n_1}\right)\right\rangle_{q(\Theta^{z^n_1})} + \sum_{t=2}^{T}\left\langle \log p\left(v^n_t, h^n_t|h^n_{t-1}, \Theta^{z^n_t}\right)\right\rangle_{q(\Theta^{z^n_t})}}.$$

The final observation assignment to the most likely dynamical regime requires inferring $q(z^n_t)$. Furthermore, the parameter distribution updates require inferring the posterior averages $\langle h^n_t \rangle$ and $\langle h^n_t h^n_{t-1} \rangle$ (see below).

To address this non-trivial inference problem, the idea is to rewrite Eq. (3) as proportional the joint distribution of a switching LGSSM $\tilde{q}(h^n_{1:T}, \tilde{v}^n_{1:T}|z^n_{1:T}, \tilde{\Theta}^{1:K})\tilde{q}(z^n_{1:T}|\tilde{\pi})$, where $\tilde{v}^n_{1:T}$ and $\tilde{\Theta}^{1:K}, \tilde{\pi}$ are new defined observations and parameters. Once done that, we can perform inference using the algorithm proposed in [2] for the switching LGSSM, which has the advantage of being a numerical stable and accurate approximation.

To reformulate Eq. (3), we first notice that thanks to the factorization

$$e^{\langle \log p(z^n_{1:T}|\pi)\rangle_{q(\pi)}} = e^{\sum_{t=1}^{T}\langle \log p(z^n_t|\pi)\rangle_{q(\pi)}}$$

we can define a new distribution

$$\tilde{q}(z^n_t = j|\tilde{\pi}) \propto e^{\langle \log p(z^n_t = j|\pi)\rangle_{q(\pi)}} = e^{\psi(\tilde{\pi}_j) - \psi(\sum_{i=1}^{K}\tilde{\pi}_i)},$$

where $\psi(\cdot)$ is the derivative of the logarithm of the Gamma function and $\tilde{\pi}_j$ are the hyperparameters of the Dirichlet distribution $q(\pi)$ (see below).

We then reexpress the second row in Eq. (3) as proportional to the conditional density of a switching LGSSM $\tilde{q}(h^n_{1:T}, \tilde{v}^n_{1:T}|z^n_{1:T}, \tilde{\Theta}^{1:K})$. More specifically, we rewrite[10] $\left\langle (v^n_t - Bh^n_t)^{\mathsf{T}}\Sigma_V^{-1}(v^n_t - Bh^n_t)\right\rangle_{q(B,\Sigma_V^{-1})}$ in Eq. (3) as

$$\underbrace{\left(v^n_t - \langle B\rangle h^n_t\right)^{\mathsf{T}}\left\langle \Sigma_V^{-1}\right\rangle\left(v^n_t - \langle B\rangle h^n_t\right)}_{mean} + \underbrace{(h^n_t)^{\mathsf{T}} S_B h^n_t}_{fluctuation},$$

where $S_B \equiv \left\langle B^{\mathsf{T}}\Sigma_V^{-1}B\right\rangle - \langle B\rangle^{\mathsf{T}}\left\langle \Sigma_V^{-1}\right\rangle\langle B\rangle$, and similarly for the part of the exponent in Eq. (3) containing $p(h^n_t|h^n_{t-1}, \Theta^{z^n_t})$. The mean terms represent the contribution of a standard switching LGSSM with parameters replaced by their average values. The key observation is to consider the extra 'fluctuation' terms as having been generated from fictitious zero-valued observations, by defining[11]

$$\tilde{v}^n_t \equiv vert\left(v^n_t, \mathbf{0}_H, \mathbf{0}_H\right), \qquad \tilde{B} \equiv vert\left(\langle B\rangle, U_A, U_B\right),$$

where U_B is the Cholesky decomposition of S_B, so that $U_B^{\mathsf{T}}U_B = S_B$ (similarly, U_A is the Cholesky decomposition of S_A). The equivalent switching LGSSM is then completed by specifying $\tilde{A} \equiv \langle A\rangle, \tilde{\Sigma}_H \equiv \left\langle \Sigma_H^{-1}\right\rangle^{-1}, \tilde{\Sigma}_V \equiv bdg(\left\langle \Sigma_V^{-1}\right\rangle^{-1}, I_H, I_H), \tilde{\mu} \equiv \langle \mu\rangle, \tilde{\Sigma} \equiv \left\langle \Sigma^{-1}\right\rangle^{-1}$.

Up to negligible constants, the joint distribution of this augmented switching LGSSM has the same form as the rhs of Eq. (3).

Parameter Updates

The parameter distribution updates are given by

$$q(\Theta^k) \propto p(\Theta^k|\hat{\Theta}^k) e^{\sum_{n=1}^{N} q(z^n_1 = k)\left\langle \log p\left(v^n_1, h^n_1|\Theta^k\right)\right\rangle_{q(h^n_1|z^n_1 = k)}}$$
$$e^{\sum_{n=1}^{N}\sum_{t=2}^{T} q(z^n_t = k)\left\langle \log p\left(v^n_t, h^n_t|h^n_{t-1}, \Theta^k\right)\right\rangle_{q(h^n_{t-1:t}|z^n_t = k)}}$$

$$q(\pi) \propto p(\pi) e^{\left\langle \log p\left(z^{1:N}_{1:T}|\pi\right)\right\rangle_{q(z^{1:N}_{1:T})}}.$$

[10] To simplify the notation, we omit the dependency of the parameters on the indicator value. This is also done in the description of the parameter distribution updates.

[11] $vert\left(X_1, \ldots, X_n\right)$ denotes the vertical concatenation of $X_1, \ldots, X_n$, while $bdg\left(X_1, \ldots, X_n\right)$ denotes the block-diagonal matrix with blocks $X_1, \ldots, X_n$.

The detailed updates are given below.

Determining $q(vc\,(A)\,|\Sigma_H^{-1})$
Let $vr\,(X)$ denote the vector formed by stacking the rows of the matrix X. The optimal $q(vr\,(A)\,|\Sigma_H^{-1})$ is Gaussian with mean and covariance given by

$$\begin{aligned} & ver\Big(\big([N_A]_{1'}\, H_{1A}^{-1}\big)^{\mathsf{T}}, \ldots, \big([N_A]_{H'}\, H_{HA}^{-1}\big)^{\mathsf{T}}\Big) \\ & bdg\left([\Sigma_H]_{11}\, H_{1A}^{-1}, \ldots, [\Sigma_H]_{HH}\, H_{HA}^{-1}\right), \end{aligned}$$

where $X_{i'}$ denotes the i-th row of the matrix X and

$$\begin{aligned} N_A &\equiv \sum_{n=1}^{N}\sum_{t=2}^{T} q(z_t^n)\left\langle h_{t-1}^n\,(h_t^n)^{\mathsf{T}}\right\rangle_{q(h_{t-1:t}^n|z_t^n)} \\ H_{iA} &\equiv \sum_{n=1}^{N}\sum_{t=2}^{T} q(z_t^n)\left\langle h_{t-1}^n\,\left(h_{t-1}^n\right)^{\mathsf{T}}\right\rangle_{q(h_{t-1}^n|z_t^n)} + dg\,(\alpha_{i'}), \end{aligned}$$

where $dg\,(X)$ is the diagonal matrix with the elements of the vector X on the diagonal.

Determining $q(vc\,(B)\,|\Sigma_V^{-1})$
Let $vc\,(X)$ denote the vector formed by stacking the columns of the matrix X. Then $q(vc\,(B)\,|\Sigma_V^{-1}) \sim \mathcal{N}\left(vc\left(N_B H_B^{-1}\right), H_B^{-1}\otimes\Sigma_V\right)$ where $\otimes$ is the Kronecker product and

$$\begin{aligned} N_B &\equiv \sum_{n=1}^{N}\sum_{t=1}^{T} q(z_t^n)\, v_t^n\,\langle h_t^n\rangle_{q(h_t^n|z_t^n)}^{\mathsf{T}} \\ H_B &\equiv \sum_{n=1}^{N}\sum_{t=1}^{T} q(z_t^n)\left\langle h_t^n (h_t^n)^{\mathsf{T}}\right\rangle_{q(h_t^n|z_t^n)} + dg\,(\beta)\,. \end{aligned}$$

Determining $q(\Sigma_H^{-1})$
If we constraint Σ_H^{-1} to be diagonal with elements τ_i following a Gamma prior $\mathcal{G}(a_1, a_2) \propto \tau_i^{a_1-1}e^{-a_2\tau_i}$, then $q(\tau_i)$ is Gamma distributed with parameters

$$\begin{aligned} & a_1 + \frac{1}{2}\sum_{n=1}^{N}\sum_{t=2}^{T} q(z_t^n) \\ & a_2 + \frac{1}{2}\left(\sum_{n=1}^{N}\sum_{t=2}^{T} q(z_t^n)\left\langle [h_t]_i^2\right\rangle - [N_A]_{i'}\, H_{iA}^{-1}\,[N_A]_{i'}^{\mathsf{T}}\right). \end{aligned}$$

Determining $q(\Sigma_V^{-1})$
If we constraint Σ_V^{-1} to be diagonal with elements $\rho_i \sim \mathcal{G}(b_1, b_2)$, then $q(\rho_i)$ is Gamma distributed with parameters

$$\begin{aligned} & b_1 + \frac{1}{2}\sum_{n=1}^{N}\sum_{t=1}^{T} q(z_t^n) \\ & b_2 + \frac{1}{2}\left(\sum_{n=1}^{N}\sum_{t=1}^{T} q(z_t^n)\,[v_t^n]_i^2 - \left[N_B H_B^{-1} N_B^{\mathsf{T}}\right]_{ii}\right). \end{aligned}$$

Determining $q(\pi)$
The optimal $q(\pi)$ is a Dirichlet distribution with parameters $\tilde{\pi}_k = \gamma/k + \sum_{n=1}^{N}\sum_{t=1}^{T} q\,(z_t^n = k)$, $k = 1, \ldots, K$.

Acknowledgments

The author would like to thank David Barber for helpful discussions.

References

[1] M. Azzouzi and I. Nabney. Modelling financial time series with switching state space models. In *Proceedings of the IEEE/IAFE Conference on Computational Intelligence for Financial Engineering*, pages 240–249, 1999.

[2] D. Barber. Expectation correction for smoothing in switching linear Gaussian state space models. *Journal of Machine Learning Research*, 7:2515–2540, 2006.

[3] D. Barber and S. Chiappa. Unified inference for variational Bayesian linear Gaussian state-space models. In *Advances in Neural Information Processing Systems 19*, pages 81–88, 2007.

[4] S. Chiappa. Unsupervised Bayesian time-series segmentation based on linear Gaussian state-space models. Technical Report no. 171, MPI for Biological Cybernetics, Tübingen, Germany, 2008.

[5] S. Chiappa and D. Barber. Dirichlet mixtures of Bayesian linear Gaussian state-space models: a variational approach. Technical Report no. 161, MPI for Biological Cybernetics, Tübingen, Germany, 2007.

[6] J. Durbin and S. Koopman. *Time Series Analysis by State Space Methods*. Oxford University Press, 2001.

[7] L. Inoue, M. Neira, C. Nelson, M. Gleave, and R. Etzioni. Cluster-based network model for time-course gene expression data. *Biostatistics*, 8:507–525, 2007.

[8] D. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

[9] B. Mesot. *Inference in switching linear dynamical systems applied to noise robust speech recognition of isolated digits*. Ph.D. Thesis no. 4059, Ecole Polytechnique Fédérale de Lausanne, 2008.

[10] V. Pavlović, J. Rehg, and J. MacCormick. Learning switching linear models of human motion. In *Advances in Neural Information Processing Systems 13*, 2001.

[11] J. Quinn, C. Williams, and N. McIntosh. Factorial switching linear dynamical systems applied to physiological condition monitoring. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008.

[12] D. Wipf, J. Palmer, and B. Rao. Perspectives on sparse Bayesian learning. In *Advances in Neural Information Processing Systems 16*, 2004.

[13] Y. Xiong and D.-Y. Yeung. Mixtures of ARMA models for model-based time series clustering. In *Proceedings of the IEEE International Conference on Data Mining*, pages 717–720, 2002.