A Variational inference method for Switching Linear Dynamic Systems

Sang Min Oh, Ananth Ranganathan, James M. Rehg, Frank Dellaert
The BORG Lab @ The College of Computing
Georgia Institute of Technology, Atlanta, GA
Technical Report number GIT-GVU-05-16
{sangmin,ananth,rehg,dellaert}@cc.gatech.edu

Drafted on 2005 June 4th.

Abstract

This paper aims to present a structured variational inference algorithm for switching linear dynamical systems (SLDSs) which was initially introduced by Pavlovic and Rehg [14]. Starting with the need for the variational approach, we proceed to the derivation of the generic (model-independent) variational update formulas which are obtained under the mean field assumption. This leads us to the derivation of an approximate variational inference algorithm for an SLDS. The details of deriving the SLDS-specific variational update equations are presented.

1 Introduction

Switching Linear Dynamical System (SLDS) models have been studied in a variety of problem domains. Representative examples include computer vision [2, 11, 14, 15, 16], computer graphics [21], speech recognition [4, 13, 18], econometrics [8], machine learning [7, 10], biology [12] and statistics [20]. While there are several versions of SLDS in the literature, this paper addresses the model structure depicted in Figure 2. An SLDS model represents the nonlinear dynamic behavior of a complex system by the switching among a set of linear dynamic models over time. In contrast to HMM's, the Markov process in an SLDS selects from a set of continuously-evolving linear Gaussian dynamics, rather than a fixed Gaussian mixture density. As a consequence, an SLDS has potentially greater descriptive power. Offsetting this advantage is the fact that exact inference in an SLDS is intractable, which complicates estimation and parameter learning [9].

The structured variational inference method for SLDS was first introduced by Pavlovic and Rehg [14][1]. Though, the presentation in [14, 16] is rather brief and hence, understanding the presented method needs substantial knowledge of both SLDS and variational methods. In addition, the derivation was based on a constrained SLDS with a fixed measurement model. This paper aims to provide the knowledge-base to facilitate the understanding of the variational inference method for SLDS [14, 16], and present thorough derivations of more generic variational inference formulas for an unconstrained SLDS model. We also demonstrate that the constrained model can be easily obtained as a special case of the generic variational inference formulas, and show that the results match those reported by Pavlovic and Rehg [14].

[1] The variational methods presented by Gharahmani and Hinton [7] is very closely related to the variational method presented in this paper. However, the involved model is slightly different.

2 Need for Variational methods

In the probabilistic inference framework, we aim to evaluate a posterior probability $P(H|Z)$ on the hidden variables H given a set of visible (evidence) variables Z as follows :

$$P(H|Z) \quad = \quad \frac{P(Z,H)}{P(Z)} \tag{1}$$

The joint probability $P(Z,H)$ on a set of all variables on r.h.s. of (13) can be obtained exactly for an arbitrary set of values Z,H once the priors and the conditional dependencies of the model are properly set up. Thus, inference on the hidden nodes H is straightforward once we can compute the evidence $P(Z)$. It may look simple. However, the the evidence given a model can be evaluated only through the exhaustive enumeration/integration of all the possible values for the hidden variables H w.r.t the joint probability (model) :

$$P(Z) \quad = \quad \int_H P(Z,H) \tag{2}$$

In general, the integration in (2) becomes computationally intractable for several reasons. First, the hidden variables H are continuous variables but there may not be an analytical solution available for (2). Second, the interdependencies within and across the hidden variables H may be exponentially complex, i.e., (it can even form a complete graph from the graph theoretical point of view), which will cause exponential increase in the number of enumerations for the concatenated hidden variable values. Clearly, albeit the hidden variables are discrete, once the number of variables is above some thresholds, the necessary enumeration would be intractable. Third, it is probable that the number of variables will monotonically increase with time in case we need to deal with the inference tasks on a temporal model. In such cases, the number of variables will increase linearly with time, which will become intractable soon with few exceptions such as Kalman filtering or RTS smoothing for linear dynamical system [1, 19] or forward-backward algorithm for HMM [17]. Consequently, one has to resort to an alternative approach to compute the posterior in (13) in case the evaluation of evidence in (2) is intractable.

As mentioned above, for most realistic models the joint posterior $P(Z,H)$ is intractable and consequently, the integration in (2) cannot be performed analytically. A common approach to overcome this intractability is to use sampling techniques such as Markov Chain Monte Carlo (MCMC) to evaluate the integral. Such techniques, while providing any level of accuracy desired, tend to be slow in general.

Variational methods evaluate the integral (2) by computing an approximating distribution to the actual joint posterior. The approximating distribution is chosen so that the integration becomes easy to perform and can be handled analytically. In the process, a tractable approximating distribution to the posterior over hidden variables $P(H|Z)$ is also obtained. In this case it is said that a posterior on hidden variables has been learned. Hence, variational learning can be seen as almost a side effect of the use of variational techniques to compute the evidence of a model.

3 Variational method

To evaluate the evidence $P(Z)$ and equivalently to assess the posterior $P(H|Z)$ as exactly as possible, variational methods find a functional form of a lower bound for the evidence $P(Z)$ and maximize that functional until it converges. The lower bound functional associates variational approximation distribution $Q(H)$ which is an approximate posterior for the target (true) posterior $P(H|Z)$. However, as mentioned earlier, we should be able to handle the updates (lower bound maximizations) analytically by carefully selecting the variational distribution $Q(H)$. Generally, variational methods take divide-and-conquer approach, i.e. we have an approximate posterior $Q(H)$ of a factored form where the factors are either a set of clusters or even fully factorized individual variables with which analytical updates are feasible. Then, we update the factors in $Q(H)$ iteratively until they all converge. In this way, the lower bound is guaranteed to improve monotonically, and we can obtain a good approximation for the evidence once it converges.

3.1 Lower bounding the evidence

Variational methods use Jensen's inequality to construct a lower bound to the log evidence of the model. This is done as follows :

$$\begin{aligned}\log P(Z) &= \log \int_H P(Z,H) \\ &= \log \int_H Q(H) \frac{P(Z,H)}{Q(H)} \\ &\geq \int_H Q(H) \log \frac{P(Z,H)}{Q(H)} \qquad (3)\end{aligned}$$

where $Q(H)$ is an arbitrary distribution on the hidden variables used to construct the lower bound, and the last relation is obtained using Jensen's inequality. This bound can also be alternately obtained as

$$\begin{aligned}\log P(Z) &= \log \frac{P(Z,H)}{P(H|Z)} \\ &= \int_H Q(H) \log \frac{Q(H)}{Q(H)} \frac{P(Z,H)}{P(H|Z)} \\ &= \int_H Q(H) \log \frac{P(Z,H)}{Q(H)} + \int_H Q(H) \log \frac{Q(H)}{P(H|Z)} \\ &= \mathcal{L}(Q) + \mathrm{KL}(Q||P) \qquad (4)\end{aligned}$$

where we have made use of the fact that $\int_H Q(H) = 1$. $\mathrm{KL}(Q||P)$ is the KL-divergence between $Q(H)$ and the posterior over hidden variables $P(H|Z)$. Since the KL-divergence is always positive (or, of course, zero), $\mathcal{L}(Q)$ is a lower bound on the evidence, which confirms our result from (3).

Note that the sum in the right hand side of (4) is constant since it is the log evidence of the model. Hence, we can proceed to approximate the evidence by minimizing the KL-divergence $\mathrm{KL}(Q||P)$ or by maximizing the lower bound $\mathcal{L}(Q)$, both of which are equivalent and yield the same result. In either case the optimization is performed wrt the $Q(H)$ distribution. The approximate posterior on the hidden variables that is obtained using variational learning is then given by

$$P^{\star}(H|Z) = \underset{Q(H)}{\mathrm{argmax}}\ \mathcal{L}(Q) = \underset{Q(H)}{\mathrm{argmin}}\ \mathrm{KL}(Q||P) \qquad (5)$$

In this document, we will take the approach of maximizing the lower bound $\mathcal{L}(Q)$ since this is more popular.

3.2 Deriving the variational updates

From the preceding discussion, it can be seen that learning the posterior on the hidden variables (and evaluating the evidence) through a variational approximation can be done by maximizing the lower bound on the log evidence. However, maximizing $\mathcal{L}(Q)$ without any constraints on $Q(H)$ is fruitless. To see this, we differentiate $\mathcal{L}(Q)$ wrt $Q(H)$ after adding a Lagrangian term that ensures that $Q(H)$ is a proper distribution, and set the derivative to zero :

$$\frac{d}{dQ(H)} \left(\mathcal{L}(Q) + \lambda \left(1 - \int_H Q(H) \right) \right) = \log P(Z,H) - \log Q(H) - 1 - \lambda = 0 \qquad (6)$$

Hence,

$$Q(H) \propto P(Z,H)$$

Obviously, the true joint distribution maximizes the bound but this is exactly the distribution that we were trying to avoid having to work with in the first place. Also note that we have not introduced any approximations in the above derivation.

To make $Q(H)$ tractable, we make the assumption, called the *mean field assumption*, that all the variables in the set H are independent, so that $Q(H)$ can be factored into the distributions on the individual variables in H, i.e. $Q(H) = \prod_i Q_i(H_i)\ \forall i$ s.t. $H_i \in H$. In making this assumption, we constrain $Q(H)$ to those distributions that adhere to the independence assumptions.

Given the mean field assumption, we can write the lower bound on the log evidence:

$$\begin{aligned}\mathcal{L}(Q) &= \int_H \left(\prod_i Q_i(H_i)\right) \log \frac{P(Z,H)}{\prod_i Q_i(H_i)} \\ &= \int_H \left(\prod_i Q_i(H_i)\right) \log P(Z,H) - \sum_i \int_{H_i} Q_i(H_i) \log Q_i(H_i) \qquad (7)\end{aligned}$$

which in turn can be written by separating out the terms in $Q_i(H_i)$ as

$$\mathcal{L}(Q) = \int_H Q_i(H_i) \langle \log P(Z,H) \rangle_{\sim Q_i(H_i)} - \int_{H_i} Q_i(H_i) \log Q_i(H_i) + \text{other terms} \qquad (8)$$

where $\langle \log P(Z,H) \rangle_{\sim Q_i(H_i)}$ is the expectation of the log joint posterior wrt to the product of all the Qs except Q_i. We can maximize the bound by performing the maximization wrt to each Q_i individually. In detail, we perform the maximization for 8 wrt Q_i in exactly the same fashion as (6) :

$$\langle \log P(Z,H) \rangle_{\sim Q_i(H_i)} - \log Q_i(H_i) - 1 - \lambda = 0 \qquad (9)$$

Finally, we get the variational update equation for every factor/cluster $Q_i(H_i)$ under the mean field assumption :

$$Q_i(H_i) = \frac{1}{c} \exp \langle \log P(Z,H) \rangle_{\sim Q_i(H_i)} \qquad (10)$$

Above, c denotes a normalizing constant. Note that these equations are coupled since each Q_i depends on all the others. The set of equations obtained using (10) are also called *fixed point equations*, as they describe the properties that should be maintained between the components in the decoupled mean field model. Finally, Q distributions are iteratively updated using the corresponding fixed point equations in (10).

The optimization performed above is unconstrained regarding the form of the Q distributions. The form of the Q distributions is given by the form of the true joint posterior $P(Z,H)$ from (10). However, it is also possible to assume some parametric distribution for the Q_i and constrain the optimization to this family. If the parametric form of the distributions is written as $Q_i(H_i|\lambda_i)$, the optimization performed above can be modified to directly obtain the values of the λ_i, referred to as the *variational parameters*, instead of the Q_i. Such a direct optimization is only tractable if the lower bound in (7) can be computed relatively efficiently. This is turn is possible if the log joint posterior $\log P(Z,H)$ is a polynomial and the parametric distributions $Q_i(H_i|\lambda_i)$ have only a few moments. If these conditions are satisfied, variational learning can be reduced to a simultaneous, (possibly) non-linear minimization problem on the variational parameters. We show a specific example of such variational approximate inference for switching linear dynamic systems (SLDSs). We first describe SLDSs and the notations to be used in Section 4. Then, we proceed and present the structured variational inference method for SLDSs in Section 5.

4 Switching Linear Dynamic Systems

A switching linear dynamic systems (SLDS) model describes the dynamics of a complex physical process by the switching between a set of linear dynamic systems (LDS).

4.1 Linear Dynamic Systems

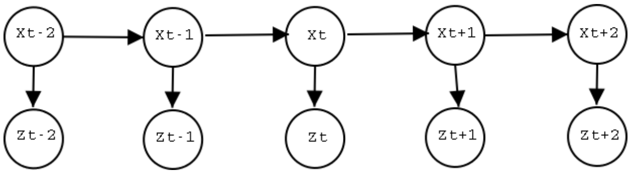


Figure 1: A linear dynamic system (LDS)

An LDS is a time-series state-space model[1, 19] that comprises a linear Gaussian dynamics model and a linear Gaussian observation model. The graphical representation of an LDS is shown in Fig.1. The Markov chain at the top represents the state evolution of the continuous hidden states x_t. The prior density p_1 on the initial state x_1 is assumed to be normal with mean x_0 and covariance Σ_0, i.e., $x_1 \sim \mathcal{N}(x_0, \Sigma_0)$.

The state x_t is obtained by the product of state transition (system) matrix F and the previous state x_{t-1} corrupted by the additive white noise w_t, zero-mean and normally distributed with covariance matrix U:

$$x_t = F x_{t-1} + w_t \text{ where } w_t \sim \mathcal{N}(0, U) \tag{11}$$

In addition, the measurement z_t is generated from the current state x_t through the observation matrix H, which is then corrupted by the white observation noise v_t:

$$z_t = H x_t + v_t \text{ where } v_t \sim \mathcal{N}(0, R) \tag{12}$$

Thus, an LDS model M is defined by the tuple $M \triangleq \{(x_0, \Sigma_0), (F, U), (H, R)\}$. Exact inference in an LDS can be done efficiently using the RTS smoother [1, 19].

4.2 Switching Linear Dynamic Systems

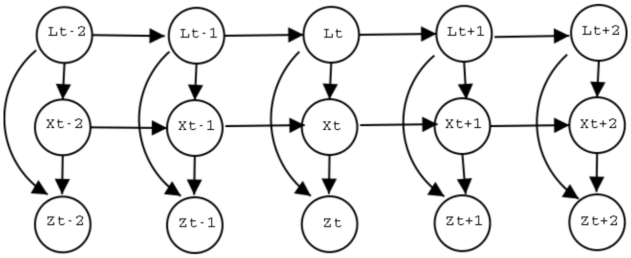


Figure 2: Switching linear dynamic systems (SLDS)

An SLDS is a natural extension of an LDS, where we assume the existence of n distinct LDS models $M \triangleq \{M_i | 1 \leq i \leq n\}$, where each model M_i is defined by the LDS parameters. The graphical model corresponding to an SLDS is shown in Fig.2. The middle chain, representing the hidden state sequence $X \triangleq \{x_t | 1 \leq t \leq T\}$, together with the observations $Z \triangleq \{z_t | 1 \leq t \leq T\}$ at the bottom, is identical to an LDS in Fig.1. However, we now have an additional discrete Markov chain $L \triangleq \{l_t | 1 \leq t \leq T\}$ that determines which of the n models M_i is being used at every time-step. We call $l_t \in M$ the *label* at time t and L a *label sequence*.

In addition to a set of LDS models M, we specify two additional parameters: a multinomial distribution $\pi(l_1)$ over the initial label l_1 and an $n \times n$ transition matrix B that defines the switching behavior between the n distinct LDS models, i.e. $B_{ij} \overset{\Delta}{=} P(l_j|l_i)$. In summary, an SLDS model is completely defined by the tuple $\Theta \overset{\Delta}{=} \left\{\pi, B, M \overset{\Delta}{=} \{M_i | i = 1..n\}\right\}$.

4.3 Learning in SLDS via EM

The EM algorithm [3] can be used to obtain the maximum-likelihood parameters $\hat{\Theta}$ of an SLDS. The hidden variables in EM are the label sequence L and the state sequence X. Given the observation data Z, EM iterates between the two steps as in Algorithm 1.

Algorithm 1 EM for Learning in SLDS

- **E-step :** Inference to obtain the posterior distribution :

$$f^i(L,X) \overset{\Delta}{=} P(L,X|Z,\Theta^i) \tag{13}$$

  over the hidden variables L and X, using a current guess for the SLDS parameters Θ^i.

- **M-step :** Maximize the expected log-likelihoods :

$$\Theta^{i+1} \leftarrow \underset{\Theta}{\operatorname{argmax}} \ \ \langle \log P(L,X,Z|\Theta) \rangle_{f^i(L,X)} \tag{14}$$

Above, $\langle \cdot \rangle_W$ denotes the expectation of a function $(\cdot)$ under a distribution W. The exact E-step in (13) is proved to be intractable[9] and motivates the development of approximate inference techniques.

4.4 Alternative approximate methods for inference in SLDS

Other than the variational inference method, previous work on SLDSs introduced various alternative approximate inference schemes. The early examples include GPB2 [1], and Kalman filtering [2]. More recent examples include an approximate Viterbi method [16, 15], expectation propagation [22], sequential Monte Carlo methods [6], iterative Monte Carlo methods [5], Data-Driven MCMC [12] and Gibbs sampling [18].

5 Structured Variational Approximation for SLDS

This section describes a structured variational approximations for switching linear dynamic system (SLDS) [13, 14]. The graphical representation of a standard switching linear dynamic system (SLDS) is shown in Fig.3. There is a switching between the discrete states L, the label sequence, at the top chain. Additionally, the states generate an observation with the switching measurement models, which is represented by the arcs from a discrete state l_t to a corresponding observation node z_t.

The exact inference in SLDS is proved to be intractable [9]. By exact inference, we mean the exact evaluation of $P(L,X|Z)$. Thus, we instead rely on an approximate variational inference technique to evaluate the approximate posterior $\tilde{P}(L,X|Z)$.

To approximate an intractable exact posterior $P(L,X,|Z)$, we use an approximate posterior $Q(L,X)$ which can be factorized into two separate distributions $Q(L)$ and $Q(X)$ based on the presented mean field assumption :

$$P(L,X|Z) \quad \approx \quad Q(L,X) = Q(L)Q(X) \tag{15}$$

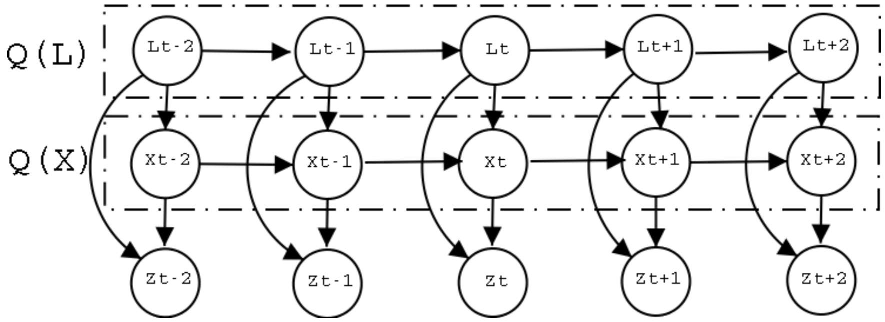


Figure 3: Switching Linear Dynamic System (SLDS) with a fixed measurement model.

Given the factorized forms in (15), we can obtain the update steps for both $Q(L)$ and $Q(X)$ using the generic update formula (10) :

$$Q(X) \quad \leftarrow \quad \frac{1}{c_X} \exp \langle \log P(L,X,Z) \rangle_{Q(L)} \tag{16}$$

$$Q(L) \quad \leftarrow \quad \frac{1}{c_L} \exp \langle \log P(L,X,Z) \rangle_{Q(X)} \tag{17}$$

The update formulas (16) and (17) look deceptively simple. In fact, a significant amount of work is involved in actually obtaining the expected joint log-likelihood $\langle \mathcal{L} \rangle \stackrel{\Delta}{=} \langle \log P(L,X,Z) \rangle$ wrt $Q(L)$ or $Q(X)$. In the following discussion, we expand the log-likelihood $\mathcal{L}$ in (18) and present the detailed derivations of each update formula in the separate subsections.

The joint log-likelihood $\mathcal{L} \stackrel{\Delta}{=} \log P(L,X,Z)$ of an SLDS model can be written as :

$$\mathcal{L} \quad = \quad \log P(L) + \log P(X|L) + \log P(Z|L,X)$$

Then, $\mathcal{L}$ can be expressed in a slightly more expanded form using the notation introduced in Section 4 :

$$\begin{aligned}
\mathcal{L} \quad = \quad & \log P(l_1|\pi) + \sum_{t=2}^{T} \log P(l_t|l_{t-1}, B) \\
& -\frac{1}{2} \left\{ \left[\left(x_1 - x_0^{(l_1)} \right)' \left(\Sigma_0^{(l_1)} \right)^{-1} \left(x_1 - x_0^{(l_1)} \right) \right] + \log |\Sigma_0^{(l_1)}| + n \log(2\pi) \right\} \\
& -\frac{1}{2} \sum_{t=2}^{T} \left\{ \left[(x_t - F_{l_t} x_{t-1})' U_{l_t}^{-1} (x_t - F_{l_t} x_{t-1}) \right] + \log |U_{l_t}| + n \log(2\pi) \right\} \\
& -\frac{1}{2} \sum_{t=1}^{T} \left\{ \left[(z_t - H_{l_t} x_t)' R_{l_t}^{-1} (z_t - H_{l_t} x_t) \right] + \log |R_{l_t}| + n \log(2\pi) \right\}
\end{aligned} \tag{18}$$

The matrices with superscripts or subscripts in (18) denote that the matrices are associated with the corresponding LDS components.

5.1 Update for $Q(X)$.

To solve (16) to update $Q(X)$, we need to evaluate the expectation of (18) wrt $Q(L)$:

$$\begin{aligned}
\langle \mathcal{L} \rangle_{Q(L)} \quad \equiv \quad & -\frac{1}{2}\left\{ x_1' \left\langle \left(\Sigma_0^{(l_1)}\right)^{-1} \right\rangle_{Q(l_1)} x_1 - 2x_1' \left\langle \left(\Sigma_0^{(l_1)}\right)^{-1} x_0^{(l_1)} \right\rangle_{Q(l_1)} + \left\langle x_0^{(l_1)'} \left(\Sigma_0^{(l_1)}\right)^{-1} x_0^{(l_1)} \right\rangle_{Q(l_1)} \right\} \\
& -\frac{1}{2}\sum_{t=2}^{T}\left\{ x_t' \left\langle U_{l_t}^{-1} \right\rangle_{Q(l_t)} x_t - 2x_t' \left\langle U_{l_t}^{-1} F_{l_t} \right\rangle_{Q(l_t)} x_{t-1} + x_{t-1}' \left\langle F_{l_t}' U_{l_t}^{-1} F_{l_t} \right\rangle_{Q(l_t)} x_{t-1} \right\} \\
& -\frac{1}{2}\sum_{t=2}^{T}\left\{ z_t' \left\langle R_{l_t}^{-1} \right\rangle_{Q(l_t)} x_t - 2z_t' \left\langle R_{l_t}^{-1} H_{l_t} \right\rangle_{Q(l_t)} x_t + x_t' \left\langle H_{l_t}' R_{l_t}^{-1} H_{l_t} \right\rangle_{Q(l_t)} x_t \right\}
\end{aligned} \tag{19}$$

Above, the PDFs $Q(l_1)$ and $Q(l_t)$ are the marginalized densities of current $Q(L)$. The notation $\equiv$ denotes that the two terms on the left and right sides are equivalent up to a constant. It can be observed that the form of the expected log-likelihood in (19) has the form of the joint log-likelihood function of a time-varying LDS. Thus, we *assume* in advance that the expected log-likelihood $\langle \mathcal{L} \rangle_{Q(L)}$ can be re-expressed as (21) by introducing a set of new variational parameters λ_X, and investigate the relevant parameters. The set of variational parameters λ_X are defined in (20) :

$$\lambda_X \quad \stackrel{\Delta}{=} \quad \left\{ \{\hat{R}_t\}_{t=1}^{T}, \{\hat{H}_t\}_{t=1}^{T}, \{\hat{U}_t\}_{t=2}^{T}, \{\hat{F}_t\}_{t=2}^{T}, \hat{x}_0, \hat{\Sigma}_0 \right\} \tag{20}$$

$$\begin{aligned}
\langle \mathcal{L} \rangle_{Q(L)} \quad \equiv \quad & -\frac{1}{2}\left[(x_1 - \hat{x}_0)' \left(\hat{\Sigma}_0\right)^{-1} (x_1 - \hat{x}_0) \right] \\
& -\frac{1}{2}\sum_{t=2}^{T}\left[\left(x_t - \hat{F}_t x_{t-1}\right)' \hat{U}_t^{-1} \left(x_t - \hat{F}_t x_{t-1}\right) \right] \\
& -\frac{1}{2}\sum_{t=1}^{T}\left[\left(z_t - \hat{H}_t x_t\right)' \hat{R}_t^{-1} \left(x_t - \hat{H}_t x_t\right) \right]
\end{aligned} \tag{21}$$

Every component in the variational parameters λ_X can be obtained exactly and efficiently by matching the coefficients of every term in (19) and (21). The procedure presented in Algorithm 2 provides an simple and efficient way to do so.

With the components obtained under this scheme, one can easily prove that (19) and (21) are equivalent[2]. The results in Algorithm 2 are slightly more involved than the update formulas reported by Pavlovic and Rehg, Eq.(6) in [14]. This is due to the fact that Pavlovic and Rehg [14] adopt a *constrained* SLDS model with a fixed measurement model. Thus, the derivation presented in this section is more general in the sense that it is derived without such constraints. The derivation of variational inference method for a constrained SLDS model is described in Section 5.3.

Now, we can observe that the expected log-likelihood $\langle \mathcal{L} \rangle_{Q(L)}$ is exactly equivalent to the joint log-likelihood of a time-varying LDS with the set of obtained variational parameters λ_X. The corresponding model is illustrated in Fig.4.

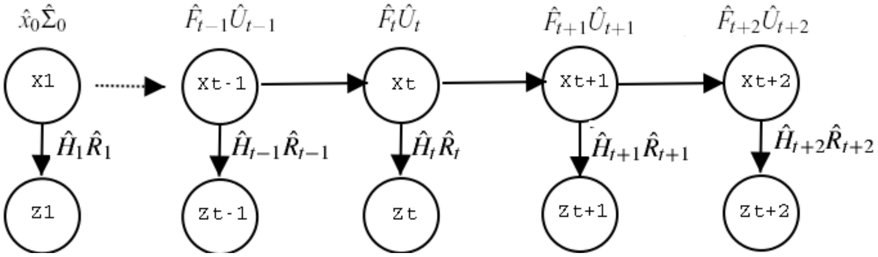
Thus, we perform RTS-smoothing on a time-varying LDS with the obtained variational parameters which are actually the series of varying LDS parameters $\lambda_X = \left\{ \{\hat{R}_t\}_{t=1}^{T}, \{\hat{H}_t\}_{t=1}^{T}, \{\hat{U}_t\}_{t=2}^{T}, \{\hat{F}_t\}_{t=2}^{T}, \hat{x}_0, \hat{\Sigma}_0 \right\}$. In other words, we simply evaluate $P(X|\lambda_X, Z)$. Finally, we update $Q(X)$:

$$Q(X) \quad \leftarrow \quad P(X|\lambda_X, Z) \tag{22}$$

[2] In fact, there are $4T+1$ terms in (19), and there are only $4T$ variational parameters in (21). Hence, the presented algorithm is not exact. However, it is often the case that the priors are uninformative Gaussians with very large covariances. Thus, the slight loss in exactness for the priors does not affect the performance of the algorithms in general. In case an SLDS has a strong prior, this can be easily resolved by introducing an additional variational parameter. However, that complicates the derivations and is omitted for the brevity of the presentation.

Algorithm 2 Variational parameters to update $Q(X)$.

```
Obtain $\lambda_X \triangleq \left\{ \{\hat{R}_t\}_{t=1}^{T}, \{\hat{H}_t\}_{t=1}^{T}, \{\hat{U}_t\}_{t=2}^{T}, \{\hat{F}_t\}_{t=2}^{T}, \hat{x}_0, \hat{\Sigma}_0 \right\}$ as follows :
For $t = T$ to 1 do
$\hat{R}_t^{-1} \leftarrow \left\langle R_{l_t}^{-1} \right\rangle_{Q(l_t)}$
$\hat{H}_t \leftarrow \hat{R}_t \left\langle R_{l_t}^{-1} H_{l_t} \right\rangle_{Q(l_t)}$
$\hat{U}_t^{-1} \leftarrow \begin{cases} \left\langle U_{l_T}^{-1} \right\rangle_{Q(l_T)} + \left\langle H_{l_T}' R_{l_T}^{-1} H_{l_T} \right\rangle_{Q(l_T)} - \hat{H}_T' \hat{R}_T^{-1} \hat{H}_T & t = T \\ \left\langle U_{l_t}^{-1} \right\rangle_{Q(l_t)} + \left\langle F_{l_{t+1}}' U_{l_{t+1}}^{-1} F_{l_{t+1}} \right\rangle_{Q(l_{t+1})} - \hat{F}_{t+1}' \hat{U}_{t+1}^{-1} \hat{F}_{t+1} + \left\langle H_{l_t}' R_{l_t}^{-1} H_{l_t} \right\rangle_{Q(l_t)} - \hat{H}_t' \hat{R}_t^{-1} \hat{H}_t & 2 \leq t < T \end{cases}$
$\hat{F}_t \leftarrow \hat{U}_t \left\langle U_{l_t}^{-1} F_{l_t} \right\rangle_{Q(l_t)}$    $2 \leq t \leq T$
$\hat{\Sigma}_0^{-1} \leftarrow \left\langle U_{l_2}^{-1} \right\rangle_{Q(l_2)} + \left\langle F_{l_2}' U_{l_2}^{-1} F_{l_2} \right\rangle_{Q(l_2)} - \hat{F}_2' \hat{U}_2^{-1} \hat{F}_2 + \left\langle H_{l_1}' R_{l_1}^{-1} H_{l_1} \right\rangle_{Q(l_1)} - \hat{H}_1' \hat{R}_1^{-1} \hat{H}_1$    $t = 1$
$\hat{x}_0 \leftarrow \hat{\Sigma}_0 \left\langle \left( \Sigma_0^{(l_1)} \right)^{-1} x_0^{(l_1)} \right\rangle_{Q(l_1)}$    $t = 1$
```



$\hat{x}_0\hat{\Sigma}_0$
$\hat{F}_{t-1}\hat{U}_{t-1}$
$\hat{F}_t\hat{U}_t$
$\hat{F}_{t+1}\hat{U}_{t+1}$
$\hat{F}_{t+2}\hat{U}_{t+2}$
X1
Xt-1
Xt
Xt+1
Xt+2
$\hat{H}_1\hat{R}_1$
$\hat{H}_{t-1}\hat{R}_{t-1}$
$\hat{H}_t\hat{R}_t$
$\hat{H}_{t+1}\hat{R}_{t+1}$
$\hat{H}_{t+2}\hat{R}_{t+2}$
Z1
Zt-1
Zt
Zt+1
Zt+2

Figure 4: Time-varying LDS with a set of variational parameters λ_X

5.2 Update for $Q(L)$.

We now solve (17) to update $Q(X)$. We again expand (18) into an involved form up to a constant :

$$
\begin{aligned}
\langle \mathcal{L} \rangle_{Q(X)} \equiv\ & \log(\pi_{l_1}) + \sum_{t=2}^{T} \log P(l_t | l_{t-1}) \\
& - \frac{1}{2} \left\langle \left[\left(x_1 - x_0^{(l_1)} \right)' \left(\Sigma_0^{(l_1)} \right)^{-1} \left(x_1 - x_0^{(l_1)} \right) \right] + \log |\Sigma_0^{(l_1)}| \right\rangle_{Q(x_1)} \\
& - \frac{1}{2} \sum_{t=2}^{T} \left\langle \left[(x_t - F_{l_t} x_{t-1})' U_{l_t}^{-1} (x_t - F_{l_t} x_{t-1}) \right] + \log |U_{l_t}| \right\rangle_{Q(x_t)} \\
& - \frac{1}{2} \sum_{t=1}^{T} \left\langle \left[(z_t - H_{l_t} x_t)' R_{l_t}^{-1} (z_t - H_{l_t} x_t) \right] + \log |R_{l_t}| \right\rangle_{Q(x_t)}
\end{aligned}
\tag{23}
$$

As before, $Q(x_1)$ and $Q(x_t)$ denote the marginalized densities of the current PDF $Q(X)$. From the form of (23), we can observe that it has a form of the joint log-likelihood function of a Hidden Markov Model (HMM). Readers not familiar with the concepts of the likelihood of a state $q_t(\cdot)$ are referred to Rabiner and Juang's tutorial [17]. An equivalent HMM can be found by setting the observation log-likelihoods $\log q_t(l_t)$ as follows :

$$\log q_t(i) \quad = \quad \begin{cases} -\frac{1}{2}\left\{\left\langle dx_1(i)'\left(\Sigma_0^{(i)}\right)^{-1} dx_1(i)\right\rangle_{Q(x_1)} + \left\langle dz_t(i)' R_i^{-1} dz_t(i)\right\rangle_{Q(x_1)} + \log|\Sigma_0^{(i)}||R_i|\right\} & t=1 \\ -\frac{1}{2}\left\{\left\langle dx_t(i)' U_i^{-1} dx_t(i)\right\rangle_{Q(x_t)} + \left\langle dz_t(i)' R_i^{-1} dz_t(i)\right\rangle_{Q(x_t)} + \log|U_i||R_i|\right\} & t>1 \end{cases} \tag{24}$$

Above, F_i, U_i, H_i, R_i denote the parameters of the ith LDS model. The notations dx_1, dx_t and dz_t used in (24) are defined as follows : $dx_1(i) \triangleq \left(x_1 - x_0^{(i)}\right)$, $dx_t(i) \triangleq (x_t - F_i x_{t-1})$ and $dz_t(i) \triangleq (z_t - H_i x_t)$.

The log-likelihood value $\log q_t(i)$ can be effectively evaluated using the sufficient statistics $\langle x_t x_t' \rangle, \langle x_t x_{t-1}' \rangle, \langle x_t \rangle$ from $Q(X)$. The set of values $q_t(i)$ comprise the variational parameters $\lambda_L \triangleq \{q_t(i) | 1 \le i \le n\}_{t=1}^{T}$ for an equivalent HMM. The graphical representation of the equivalent HMM is illustrated in Fig.5.

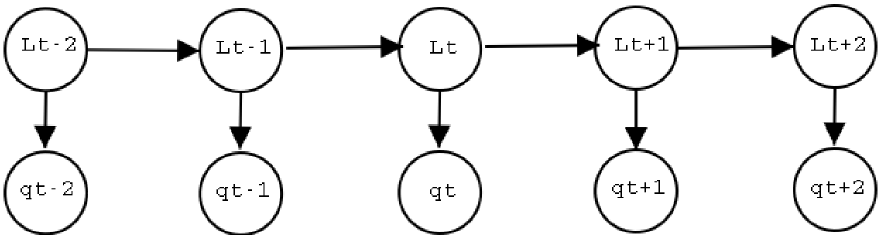


Figure 5: HMM with a set of variational parameters λ_L

The standard forward-backward algorithm for HMM is applied with a set of variational parameters λ_L, and the approximate posterior on the label sequence $P(L|\pi, B, \lambda_L)$ is obtained. Finally, we update $Q(L)$:

$$Q(L) \quad \leftarrow \quad P(L|\pi, B, \lambda_L) \tag{25}$$

5.3 SLDSs with a fixed measurement model

This section presents the variational method for an SLDS with an additional assumption which constrains the SLDS model to have a *fixed* measurement model. The graphical representation of an SLDS with a fixed measurement model is shown in Fig.6. It can be observed that the dependencies from every discrete node l_t to a corresponding observation z_t are removed. This model has been supported with the argument that the measurements may not depend on the current states of an object being tracked. Rather, they are dependent on the characteristics of a measurement device (which does not change with the target's states)[4].

The overall derivation of the variational updates are analogous to those presented in Section 5.1 and 5.2, resulting in less involved forms.

First, we again write the expected log-likelihood $\langle \mathcal{L} \rangle$ wrt $Q(L)$ upto a constant :

$$\begin{aligned} \langle \mathcal{L} \rangle_{Q(L)} \quad \equiv \quad & -\frac{1}{2}\left\{x_1' \left\langle \left(\Sigma_0^{(l_1)}\right)^{-1} \right\rangle_{Q(l_1)} x_1 - 2x_1' \left\langle \left(\Sigma_0^{(l_1)}\right)^{-1} x_0^{(l_1)} \right\rangle_{Q(l_1)} + \left\langle x_0^{(l_1)'} \left(\Sigma_0^{(l_1)}\right)^{-1} x_0^{(l_1)} \right\rangle_{Q(l_1)} \right\} \\ & -\frac{1}{2}\sum_{t=2}^{T}\left\{x_t' \left\langle U_{l_t}^{-1} \right\rangle_{Q(l_t)} x_t - 2x_t' \left\langle U_{l_t}^{-1} F_{l_t} \right\rangle_{Q(l_t)} x_{t-1} + x_{t-1}' \left\langle F_{l_t}' U_{l_t}^{-1} F_{l_t} \right\rangle_{Q(l_t)} x_{t-1} \right\} \end{aligned} \tag{26}$$

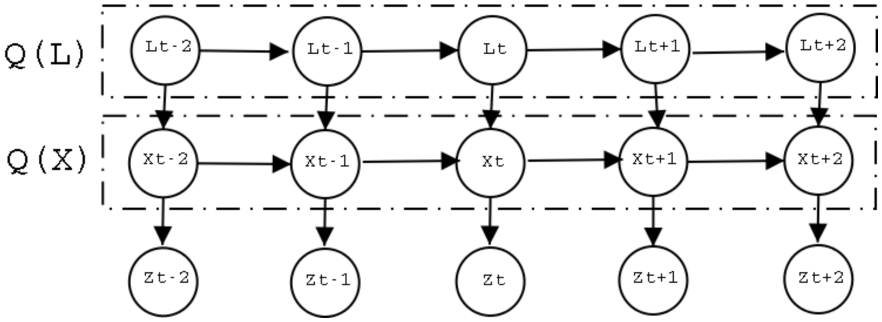


Figure 6: Switching Linear Dynamic System (SLDS) with a fixed measurement model.

Note that the terms with the measurement parameters, e.g. H, R, disappear in (26). Then, we investigate an equivalent time-varying LDS by introducing a set of variational parameters λ_X :

$$\lambda_X \quad \triangleq \quad \left\{ \left\{\hat{U}_t\right\}_{t=2}^{T}, \left\{\hat{F}_t\right\}_{t=2}^{T}, \hat{x}_0, \hat{\Sigma}_0 \right\} \tag{27}$$

The target LDS has a joint log-likelihood function below :

$$\begin{aligned} \langle \mathcal{L} \rangle_{Q(L)} \quad \equiv \quad & -\frac{1}{2}\left[\left(x_1 - \hat{x}_0\right)' \hat{\Sigma}_0^{-1} \left(x_1 - \hat{x}_0\right) \right] \\ & -\frac{1}{2}\sum_{t=2}^{|Z|} \left[\left(x_t - \hat{F}_t x_{t-1}\right)' \hat{U}_t^{-1} \left(x_t - \hat{F}_t x_{t-1}\right) \right] \end{aligned} \tag{28}$$

We observe that (26) and (28) should be equivalent, and find the solutions for the set of variational parameters λ_X. The procedure is described in Algorithm 3.

Algorithm 3 Evaluation of variational parameters λ_X for a constrained SLDS.

```
$\lambda_X \triangleq \left\{ \left\{\hat{U}_t\right\}_{t=2}^{T}, \left\{\hat{F}_t\right\}_{t=2}^{T}, \hat{x}_0, \hat{\Sigma}_0 \right\}$ as follows :
For $t = T$ to 1 do
$\hat{U}_t^{-1} \leftarrow \begin{cases} \left\langle U_{l_T}^{-1} \right\rangle_{Q(l_T)} & t = T \\ \left\langle U_{l_t}^{-1} \right\rangle_{Q(l_t)} + \left\langle F'_{l_{t+1}} U_{l_{t+1}}^{-1} F_{l_{t+1}} \right\rangle_{Q(l_{t+1})} - \hat{F}'_{t+1} \hat{U}_{t+1}^{-1} \hat{F}_{t+1} & 2 \le t < T \end{cases}$
$\hat{F}_t \leftarrow \hat{U}_t \left\langle U_{l_t}^{-1} F_{l_t} \right\rangle_{Q(l_t)} \qquad 2 \le t \le T$
$\hat{\Sigma}_0^{-1} \leftarrow \left\langle U_{l_1}^{-1} \right\rangle_{Q(l_1)} + \left\langle F'_{l_2} U_{l_2}^{-1} F_{l_2} \right\rangle_{Q(l_2)} - \hat{F}'_2 \hat{U}_2^{-1} \hat{F}_2 \qquad t = 1$
$\hat{x}_0 \leftarrow \hat{\Sigma}_0 \left\langle \left(\Sigma_0^{(l_1)}\right)^{-1} x_0^{(l_1)} \right\rangle_{Q(l_1)} \qquad t = 1$
```

The results in Algorithm 3 matches the update formulas reported by Pavlovic and Rehg, Eq.(6) in [14]. It can be observed that the variational updates for this constrained SLDS with a fixed measurement model are obtained simply by removing all the terms regarding the switching measurement models from the more generic derivations presented in Section 5.1 and 5.2.

As before, we perform RTS-smoothing on a time-varying LDS with the obtained variational parameters $\lambda_X = \left\{ \{\hat{U}_t\}_{t=2}^{T}, \{\hat{F}_t\}_{t=2}^{T}, \hat{x}_0, \hat{\Sigma}_0 \right\}$, i.e., we evaluate $P(X|\lambda_X, Z)$. The final update of $Q(X)$ is identical to (22) : $Q(X) \leftarrow P(X|\lambda_X, Z)$.

Again, we can obtain the update formulas for $Q(L)$ in an analogous manner. While further derivation details are omitted, the update formulas are shown in Eq.29. Once the variational parameters $\lambda_L \overset{\Delta}{=} \{q_t(i)\}_{t=1}$ are obtained, we perform forward-backward algorithm for HMM, and update $Q(L) \leftarrow P(X|\lambda_X, Z)$.

$$\log q_t(i) \quad = \quad \begin{cases} -\frac{1}{2}\left\langle \left(x_1 - x_0^{(i)}\right)' \left(\Sigma_0^{(i)}\right)^{-1} \left(x_1 - x_0^{(i)}\right) \right\rangle - \frac{1}{2}\log|\Sigma_0^{(i)}| & t = 1 \\ -\frac{1}{2}\left\langle (x_t - F_i x_{t-1})' U_i^{-1} (x_t - F_i x_{t-1}) \right\rangle - \frac{1}{2}\log|U_i| & t > 1 \end{cases} \tag{29}$$

6 Conclusion

The structured variational inference method for SLDSs is presented in this paper. Full derivations of the variational inference method for a generic SLDS is demonstrated, and it is shown that the results reported by Pavlovic and Rehg [14] can be obtained as a special case once a reasonable constraint on the model structure is added.

The final variational posterior $Q(LX) \approx Q(L)Q(X)$, which approximates the exact posterior $P(L, X|Z)$, is obtained as a side effect while we iteratively improve the expected log-likelihoods of an SLDS model $\langle \mathcal{L} \rangle$ wrt the factorized variational posteriors $Q(X)$ and $Q(L)$ in turn. The approximate variational inference method was used in the domain of human figure tracking [14, 16], and has been reported to be comparable to some of the alternative approximate inference methods that are described in Section 4.4. However, the competency of the variational inference method presented here against the competing methods in broad application domains needs to be further investigated. We expect active contributions toward resolving this question in the near future.

References

[1] Y. Bar-Shalom and X. Li, *Estimation and Tracking: principles, techniques and software*, Artech House, Boston, London, 1993.

[2] C. Bregler, *Learning and recognizing human dynamics in video sequences*, in IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 1997.

[3] A. Dempster, N. Laird, and D. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the Royal Statistical Society, Series B, 39 (1977), pp. 1–38.

[4] V. Digalakis, J. R. Rohlicek, and M. Ostendorf, *A dynamical system approach to continuous speech recognition*, IEEE Trans. Speech Audio Processing, 1 (1993), pp. 431–442.

[5] A. Doucet and C. Andrieu, *Iterative algorithms for state estimation of jump markov linear systems*, IEEE Trans. Signal Processing, 49 (2001).

[6] A. Doucet, N. J. Gordon, and V. Krishnamurthy, *Particle filters for state estimation of jump Markov linear systems*, IEEE Trans. Signal Processing, 49 (2001).

[7] Z. Ghahramani and G. E. Hinton, *Variational learning for switching state-space models*, Neural Computation, 12 (1998), pp. 963–996.

[8] C.-J. Kim, *Dynamic linear models with Markov-switching*, Journal of Econometrics, 60 (1994).

[9] U. Lerner and R. Parr, *Inference in hybrid networks: Theoretical limits and practical algorithms*, in Proc. 17th Annual Conference on Uncertainty in Artificial Intelligence (UAI-01), Seattle, WA, 2001, pp. 310–318.

[10] U. Lerner, R. Parr, D. Koller, and G. Biswas, *Bayesian fault detection and diagnosis in dynamic systems*, in Proc. AAAI, Austin, TX, 2000.

[11] B. North, A. Blake, M. Isard, and J. Rottscher, *Learning and classification of complex dynamics*, IEEE Trans. Pattern Anal. Machine Intell., 22 (2000), pp. 1016–1034.

[12] S. M. Oh, J. M. Rehg, T. Balch, and F. Dellaert, *Data-driven MCMC for learning and inference in switching linear dynamic systems*, in AAAI Nat. Conf. on Artificial Intelligence, 2005.

[13] M. Ostendorf, V. V. Digalakis, and O. A. Kimball, *From hmm's to segment models : A unified view of stochastic modeling for speech recognition*, IEEE Transactions on Speech and Audio Processing, 4 (1996), pp. 360–378.

[14] V. Pavlović and J. Rehg, *Impact of dynamic model learning on classification of human motion*, in IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2000.

[15] V. Pavlović, J. Rehg, T.-J. Cham, and K. Murphy, *A dynamic Bayesian network approach to figure tracking using learned dynamic models*, in Intl. Conf. on Computer Vision (ICCV), 1999.

[16] V. Pavlović, J. Rehg, and J. MacCormick, *Learning switching linear models of human motion*, in Advances in Neural Information Processing Systems (NIPS), 2000.

[17] L. Rabiner and B. Juang, *An introduction to hidden Markov models*, in IEEE ASSP Magazine, 1986.

[18] A.-V. Rosti and M. Gales, *Rao-blackwellised Gibbs sampling for switching linear dynamical systems*, in Intl. Conf. Acoust., Speech, and Signal Proc. (ICASSP), vol. 1, 2004, pp. 809–812.

[19] S. Roweis and Z. Ghahramani, *A unifying review of linear gaussian models*, Neural Computation, 11 (1999), pp. 305–345.

[20] R. Shumway and D. Stoffer, *Dynamic linear models with switching*, Journal of the American Statistical Association, 86 (1992), pp. 763–769.

[21] Y.Li, T.Wang, and H.-Y. Shum, *Motion texture : A two-level statistical model for character motion synthesis*, in SIGGRAPH, 2002.

[22] O. Zoeter and T. Heskes, *Hierarchical visualization of time-series data using switching linear dynamical systems*, IEEE Trans. Pattern Anal. Machine Intell., 25 (2003), pp. 1202–1215.