Bayesian Inference for Switching Linear Dynamical Systems

D. Bacilieri, L. Barbiero, G. Bordin, A. Pitteri

11th March 2024

1 Model description

A switching linear dynamical system – also known as *switching state space model* – is defined, borrowing the notation from Linderman et al. [6], by the set of discrete-time stochastic equations

$$x_t = A_{z_t} x_{t-1} + b_{z_t} + v_t, \tag{1}$$
$$y_t = C_{z_t} x_t + d_{z_t} + w_t, \tag{2}$$

where $v_t \in \mathbb{R}^M$ and $w_t \in \mathbb{R}^N$ are Gaussian-distributed random vectors with mean zero and variance Q_{z_t}, S_{z_t} respectively. The vectors $y_t \in \mathbb{R}^N$, $t = 1, \dots, T$ may represent a time series of observations, while the $x_t \in \mathbb{R}^M$ are a set of continuous latent states linked together by linear dynamics defined by the matrices $A_k \in \mathbb{R}^{M \times M}$ and bias vectors $b_k \in \mathbb{R}^M$. The transformation between x and y is also linear, through the matrices $C_k \in \mathbb{R}^{N \times M}$ and bias vectors $d_k \in \mathbb{R}^N$.

The linear parameters A_k, b_k, C_k, d_k form a discrete set of K elements, and a discrete latent variable $z_t \in \{1, \dots, K\}$ sets the specific instances in use at time step t. The variable z evolves over time as a Markov process, meaning that z_t is conditionally independent of all previous states except for its immediate predecessor z_{t-1}:

$$p(z_t \mid z_{t-1}, z_{t-2}, \dots, z_1) = p(z_t \mid z_{t-1}). \tag{3}$$

We will denote the probability to transition from $z_{t-1} = j$ to $z_t = k$ with π_{jk}. The transition $z_{t-1} \to z_t$ effectively modifies the linear dynamics from x_{t-1} to x_t and the linear transformation from x_t to y_t, *switching* from one regime to another.

Given a set of data points y_t, the goal is then to infer the posterior distribution of the parameter set

$$\vartheta = \{\pi_k, A_k, b_k, Q_k, C_k, d_k, S_k\}, \tag{4}$$

where $x_{1:T}$ denotes the whole sequence $x_1, x_2, \dots, x_T$, and π_k the kth row of the transition matrix.

2 Model implementation

To set up a Monte Carlo sampling scheme, we chose to work with the Stan [8] programming language and its implementation in R through the package rstan [7]. Sampling in Stan is done by default using a variant of the Hamiltonian Monte Carlo scheme called 'No-U-Turn sampler' or NUTS [2].

The model cannot be implemented directly as it is, in the sense of specifying a categorical likelihood for the transition $z_t \to z_{t+1}$ and Gaussian likelihoods for $x_t \mid x_{t-1}$ and $y_t \mid x_t$, because Stan does not allow the definition of integer parameters: so, one should marginalize over the hidden discrete states. Besides Stan's limitations in this regard, the resulting strategy – known in the literature as *forward algorithm* – is more efficient than the straightforward implementation in sampling low-probability states, and is commonly used in similar inference problems involving hidden Markov models or other state space models [3].

2.1 The forward algorithm

The basic idea behind the forward algorithm is to exploit a recursive relationship to build the full likelihood: indeed, consider the quantity

$$\gamma_t(k) \coloneqq p(z_t = k, x_{1:t}, y_{1:t}). \tag{5}$$

By summing over the z states at $t-1$ first and then using the chain rule repeatedly, we can write

$$\begin{aligned} \gamma_t(k) &= \sum_{j=1}^{K} p(z_t = k, z_{t-1} = j, x_{1:t}, y_{1:t}) \\ &= p(y_t \mid z_t = k, x_t)\, p(x_t \mid z_t = k, x_{t-1}) \\ &\quad \cdot \sum_{j=1}^{K} \pi_{jk}\, p(z_{t-1} = j, x_{1:t-1}, y_{1:t-1}), \end{aligned} \tag{6}$$

where we have recognized the conditional probability $p(z_t = k \mid z_{t-1} = j, x_{1:t-1}, y_{1:t-1}) = p(z_t = k \mid z_{t-1} = j)$ as the element (j, k) of the transition matrix π. The first two terms outside the sum are the likelihoods of y_t and x_t, and because of the model definition they only depend on z_t, x_t and x_{t-1}. Also, they are simply Gaussian densities:

$$\mathcal{L}_k(y_t) \coloneqq p(y_t \mid z_t = k, x_t) = \mathcal{N}(C_k x_t + d_k, S_k), \tag{7}$$
$$\mathcal{L}_k(x_t) \coloneqq p(x_t \mid z_t = k, x_{t-1}) = \mathcal{N}(A_k x_{t-1} + b_k, Q_k). \tag{8}$$

Then, the remaining terms in the sum in Eq. (6) are nothing else than $\gamma_{t-1}(j)$, giving us the recursive relation we needed:

$$\gamma_t(k) = \mathcal{L}_k(y_t)\, \mathcal{L}_k(x_t) \sum_{j=1}^{K} \pi_{jk} \gamma_{t-1}(j). \tag{9}$$

Indeed, to retrieve the full joint likelihood of the sequences $x_{1:T}$ and $y_{1:T}$ we only need to marginalize γ at the last time step T over the discrete states $k = 1, 2, \dots, K$:

$$p(x_{1:T}, y_{1:T}) = \sum_{k=1}^{K} p(z_T = k, x_{1:T}, y_{1:T}) = \sum_{k=1}^{K} \gamma_T(k). \tag{10}$$

To recursively build γ_t up to time T we need $\mathcal{O}(TK^2)$ operations, because of the double marginalization over z_t and z_{t-1}. To initialize the recursion,

$$\begin{aligned}\gamma_1(k) &= p(z_1 = k, x_1, y_1) \\ &= \mathcal{L}_1(y_1)\, p(x_1 \mid z_1 = k)\, p(z_1 = k).\end{aligned} \tag{11}$$

The last two terms are the prior distributions on x_1 and z_1. We chose a multivariate Gaussian for the first and a uniform distribution over the K states for the second.

At this point, we also need the prior distributions for the dynamical parameters. Following the suggestion from Linderman et al. [6], we chose matrix-normal-inverse-Wishart priors:

$$(A_k, b_k), Q_k \sim \mathrm{MNIW}(M_x, \Omega_x, \Psi_x, \nu_x) \tag{12}$$

$$(C_k, d_k), S_k \sim \mathrm{MNIW}(M_y, \Omega_y, \Psi_y, \nu_y). \tag{13}$$

Here $M_x \in \mathbb{R}^{M\times(M+1)}$ and $M_y \in \mathbb{R}^{N\times(M+1)}$ are the mean matrices of the matrix normals, $\Omega_x, \Omega_y \in \mathbb{R}^{(M+1)\times(M+1)}$ their between-column covariance matrices, while $\Psi_x \in \mathbb{R}^{M\times M}$ and $\Psi_y \in \mathbb{R}^{N\times N}$ are the scale matrices of the inverse Wisharts and ν_x, ν_y their degrees of freedom. The returned random matrices with $M + 1$ columns are then split between the matrices A_k and C_k and their corresponding bias vectors b_k and d_k.

At the time of writing, Stan has not yet implemented a matrix normal distribution function [5]. Therefore, we have resorted to implementing it by defining the logarithm of the probability density. The explicit form of the matrix normal distribution with parameters $M \in \mathbb{R}^{p\times q}$, $\Sigma \in \mathbb{R}^{p\times p}$, $\Omega \in \mathbb{R}^{q\times q}$ is [4]

$$\begin{aligned}p(X \mid M, \Sigma, \Omega) = &\frac{1}{(2\pi)^{pq/2}|\Omega|^{p/2}|\Sigma|^{q/2}} \\ &\cdot \exp\left\{-\frac{1}{2}\operatorname{tr}\left[\Omega^{-1}(X - M)^\intercal \Sigma^{-1}(X - M)\right]\right\}.\end{aligned} \tag{14}$$

2.2 Reconstructing the hidden states

Since we marginalize out the z sequence during the sampling procedure, we need a way to recover them probabilistically. One way to do it is to search *a posteriori* for the most likely hidden sequence of z states conditioned to the observed $y_{1:T}$ and inferred $x_{1:T}$. This is done through the so-called 'Viterbi algorithm' [8], which is based on a recursive relation much like the forward algorithm.

Indeed, consider the quantity

$$\eta_t(k) \coloneqq \max_{z_{1:t-1}} p(z_{1:t-1}, z_t = k, x_{1:t}, y_{1:t}). \tag{15}$$

If we proceed similarly to (6), and using also the fact that

$$\max_{a,b}[f(a)g(a,b)] = \max_a[f(a) \max_b g(a,b)], \tag{16}$$

we can expand the definition as

$$\begin{aligned}\eta_t(k) = \max_{j\in\{1,\dots,K\}} &[\, p(y_t \mid z_t = k, x_t)\, p(x_t \mid z_t = k, x_{t-1}) \\ &\cdot \pi_{jk} \max_{z_{1:t-2}} p(z_{1:t-2}, z_{t-1} = j, x_{1:t-1}, y_{1:t-1})].\end{aligned} \tag{17}$$

We then recognize the last factor as η at the previous time step $t - 1$, giving us the recursive relation

$$\eta_t(k) = \mathcal{L}_k(y_t)\mathcal{L}_k(x_t) \max_{j\in\{1,\dots,K\}} [\pi_{jk}\, \eta_{t-1}(j)]. \tag{18}$$

Regarding the initial value, there is no z_{t-1} to maximize over, so we get

$$\eta_1(k) = p(z_1 = k, x_1, y_1), \tag{19}$$

which is the same initialization of $\gamma_1(k)$ as shown in the previous section, Eq. (11). Once we have the value of η at time T we can maximize over z_T and recover the maximum-probability sequence $\hat{z}_{1:T}$:

$$\hat{z}_{1:T} = \max_{z_{1:T}} p(z_{1:T}, x_{1:T}, y_{1:T}) = \max_{k\in\{1,\dots,K\}} \eta_T(k). \tag{20}$$

The procedure is thus substantially the same as the forward algorithm, but with maximization replacing summation.

3 Adding recursion

A possible improvement over the standard SLDS model is the one proposed by Barber [1], referred to by Linderman et al. [6] as *recurrent* SLDS. The recursivity here consists in adding a link between the current discrete hidden state z_t and the continuous hidden state at the previous time step x_{t-1}. In this way the regime switching can have a non-Markovian dependence on the continuous latent state, greatly enhancing the descriptive power in situations where there is strong multi-time step correlation in the linear dynamics.

To model the connection $z_t \mid x_{t-1}$, Linderman et al. [6] propose a linear transformation combined with a stick-breaking process to construct a vector π_{z_t} of transition probabilities. Specifically, let $\nu_t \in \mathbb{R}^{K-1}$, defined by

$$\nu_t = R_{z_{t-1}} x_{t-1} + r_{z_{t-1}}, \tag{21}$$

where $R_k \in \mathbb{R}^{K-1\times M}$ and $r_k \in \mathbb{R}^{K-1}$. The relative magnitudes of the matrix R and the bias vector r control the weight given to the recursive influence (R) compared to the pure Markov dynamics between the z states (r) [6]. The transition probabilities of z_t conditioned to x_{t-1} are then given by

$$z_t \mid x_{t-1} \sim \pi_{\mathrm{SB}}(\nu_t), \tag{22}$$

where $\pi_{\mathrm{SB}} \colon \mathbb{R}^{K-1} \to [0,1]^K$ is the stick-breaking function, mapping the vector ν_t to a normalized probability vector. Its kth component is defined as

$$\pi_{\mathrm{SB}}^{(k)}(\nu) = \begin{cases}\sigma(\nu_k) \prod_{j=1}^{k}(1 - \sigma(\nu_j)) & \text{if } k \leqslant K - 1, \\ \prod_{j=1}^{K-1}(1 - \sigma(\nu_j)) & \text{if } k = K,\end{cases} \tag{23}$$

where $\sigma(x) = 1/(1 + e^{-x})$ denotes the sigmoid function.

The forward algorithm described in Section 2.1 can readily accommodate this change in the transition probabilities. Indeed, recovering the expansion of $\gamma_t(k)$ from Eq. (6), we just need to adapt the term $p(z_t = k \mid z_{t-1} = j, x_{1:t-1}, y_{1:t-1})$. The same is valid for the Viterbi algorithm, of course. $p(z_t = k \mid z_{t-1} = j)$ becomes $p(z_t = k \mid z_{t-1} = j, x_{t-1})$, which is the kth component of the stick-breaking function as defined above. Working with logarithms, we can write this term quite simply as

$$\log p(z_t = k \mid z_{t-1} = j, x_{t-1}) = \nu_k - \sum_{j=1}^{k} \log(1 + e^{\nu_j}), \tag{24}$$

with $\nu = R_j x_{t-1} + r_j$ setting the dependence on x_{t-1}.

We also need additional priors for R_k and r_k. In keeping with the priors for the other sets of linear parameters, we chose a matrix normal distribution:

$$(R_k, r_k) \sim \mathrm{MN}(M_r, \Sigma_r, \Omega_r). \tag{25}$$

Here $M_r \in \mathbb{R}^{(K-1)\times(M+1)}$ is the mean matrix, while $\Sigma_r \in \mathbb{R}^{(K-1)\times(K-1)}$ and $\Omega_r \in \mathbb{R}^{(M+1)\times(M+1)}$ are the between-row and between-column covariance matrix respectively.

References

[1] David Barber. 'Expectation Correction for Smoothed Inference in Switching Linear Dynamical Systems'. In: *Journal of Machine Learning Research* 7.89 (2006), pp. 2515–2540.

[2] Bob Carpenter et al. 'Stan: A Probabilistic Programming Language'. In: *Journal of Statistical Software* 76.1 (2017), pp. 1–32. DOI: 10.18637/jss.v076.i01.

[3] Luis Damiano, Brian Peterson and Michael Weylandt. 'A tutorial on hidden Markov models using Stan'. In: StanCon 2018 (Asilomar Conference Center, California, 10th Jan. 2018). Asilomar. Zenodo, 10th Jan. 2018. DOI: 10.5281/zenodo.1284341.

[4] D. J. De Waal. 'Matrix-Valued Distributions'. In: *Encyclopedia of Statistical Sciences*. Ed. by Samuel Kotz et al. Found. by Norman L. Johnson. 2nd ed. Vol. 7. John Wiley & Sons, Ltd, Dec. 2005, pp. 4613–4620. ISBN: 9780471667193. DOI: https://doi.org/10.1002/0471667196.ess1565.pub2.

[5] Daniel Lee et al. *stan-dev/math: 2.17.1*. Version v2.17.1. Dec. 2017. DOI: 10.5281/zenodo.1101101.

[6] Scott W. Linderman et al. 'Bayesian Learning and Inference in Recurrent Switching Linear Dynamical Systems'. In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. Ed. by Aarti Singh and Jerry Zhu. Vol. 54. Proceedings of Machine Learning Research. PMLR, Apr. 2017, pp. 914–922.

[7] Stan Development Team. *RStan: the R interface to Stan*. R package version 2.32.6. 2024.

[8] Stan Development Team. *Stan Modeling Language Users Guide and Reference Manual*. Version 2.34. 2024.