Variational Bayesian Inference and Learning for Continuous Switching Linear Dynamical Systems

Jack Goffinet
Department of Computer Science
Duke University
Durham, NC 27708
jack.goffinet@duke.edu

David E. Carlson
Department of Computer Science
Duke University
Durham, NC 27708
david.carlson@duke.edu

Abstract

Linear-Gaussian dynamical systems (LDSs) are computationally tractable because all latents and observations are jointly Gaussian. However, these systems are too restrictive to satisfactorily model many dynamical systems of interest. One generalization, the switching linear dynamical system (SLDS), trades analytic tractability for a more expressive model, allowing a discrete set of different linear regimes to model the data. Here we introduce a switching linear dynamical system with a continuum of linear regimes that are traversed continuously in time. We call this model a *continuous switching linear dynamical system* (CSLDS) and derive efficient variational Bayesian methods for inference and model learning.

1 Introduction

Linear dynamical systems (LDSs) are fundamental to probabilistic timeseries modeling due to their analytic tractability and the past several years have seen an increased interest in more expressive extensions of LDSs. The most common among these is the *switching* linear dynamical system (SLDS), which posits a set of discrete modes, each corresponding to a distinct LDS, with Markov transitions between the modes [3, 7]. Several extensions of SLDS exist, including recurrent variants that use both the discrete mode and the continuous state to determine mode dynamics [5, 6, 4].

The choice made by SLDS of maintaining a discrete set of linear regimes is computationally convenient given the Markov model structure of the modes. However, we note that an SLDS with a continuum of linear regimes can also be handled in a computationally convenient way. The resulting model, which we call a *continuous switching linear dynamical systems* (CSLDS), is better suited to modeling systems for which we expect the underlying dynamics vary smoothly in time.

The most similar model to the CSLDS in the literature is the warped autoregressive hidden Markov model (WAR-HMM) [2]. The autoregressive (AR)-HMM can be viewed as a discrete SLDS with a restricted observation model, which the WAR-HMM extends by associating with each discrete mode a univariate latent variable that modulates the mode's corresponding linear dynamics. The CSLDS, by contrast, modulates its linear dynamics with a continuously varying multivariate latent variable and extends the more flexible observation model of the SLDS.

2 Continuous Switching Linear Dynamical Systems

Setup Let time $t = 1, \ldots, T$ have associated observations $\boldsymbol{y}_t \in \mathbb{R}^N$ for all t. The model has two layers of latent variables: $\boldsymbol{\eta}_t \in \mathbb{R}^K$ for all $t < T$, and $\boldsymbol{x}_t \in \mathbb{R}^M$ for all t, with $M \geq K$. The graphical model is shown in Figure 1, which has an associated joint distribution of

$$p(\boldsymbol{\eta}_{1:T-1}, \boldsymbol{x}_{1:T}, \boldsymbol{y}_{1:T}) = p(\boldsymbol{\eta}_{1:T-1})p(\boldsymbol{x}_1)\prod_{t=2}^{T} p(\boldsymbol{x}_t|\boldsymbol{\eta}_{t-1}, \boldsymbol{x}_{t-1})\prod_{t=1}^{T} p(\boldsymbol{y}_t|\boldsymbol{x}_t)\ . \tag{1}$$

2022 NeurIPS Workshop on Gaussian Processes, Spatiotemporal Modeling, and Decision-making Systems.

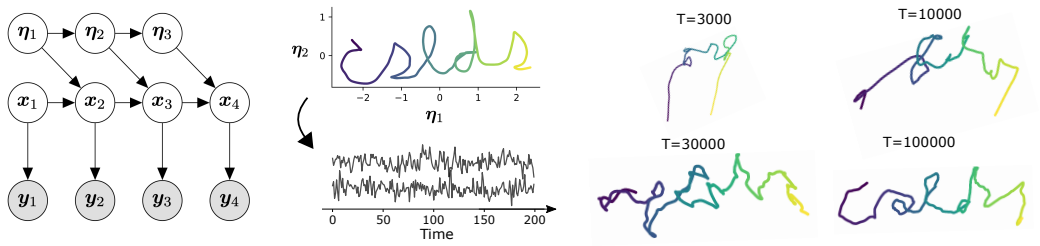

Figure 1: Left: Graphical model shared by the switching linear dynamical system (SLDS) and the proposed continuous switching linear dynamical system (CSLDS). Center: A synthetic dataset was made using continuously varying $\boldsymbol{\eta}$'s which spell out "cslds," where each setting of $\boldsymbol{\eta}$ corresponds to a distinct linear dynamical regime. Right: The proposed model is able to approximately recover the $\boldsymbol{\eta}$'s from the synthetic timeseries, with more accurate results obtained for datasets with more timepoints, T.

We choose an independent Gaussian process prior on each component of $\boldsymbol{\eta}$: $\boldsymbol{\eta}_{1:T-1}^{(k)} \sim \mathcal{GP}(\mathbf{0}, \kappa_k)$, where κ_k is a given kernel. We define the initial condition prior $\boldsymbol{x}_1 \sim \mathcal{N}(0, Q_{\text{IC}}^{-1})$, and the interaction between the $\boldsymbol{\eta}$'s and the $\boldsymbol{x}$'s as $\boldsymbol{x}_t|\boldsymbol{\eta}_{t-1}, \boldsymbol{x}_{t-1} \sim \mathcal{N}(A(\boldsymbol{\eta}_{t-1})\boldsymbol{x}_{t-1}, [Q(\boldsymbol{\eta}_{t-1})]^{-1})$. Lastly, we define a linear-Gaussian observation model: $\boldsymbol{y}_t|\boldsymbol{x}_t \sim \mathcal{N}(C\boldsymbol{x}_t + \boldsymbol{b}, R)$. In this way, the $\boldsymbol{\eta}$'s control the parameters of a linear-Gaussian state space model, producing an LDS with continuously varying parameters. We introduce two concrete choices for $A(\boldsymbol{\eta}_t)$ and $Q(\boldsymbol{\eta}_t)$, a simplex definition that is more consistent with the existing SLDS literature, and a factor definition that is more readily comparable to factor analysis. The *simplex* definition is

$$\boldsymbol{s} = \text{Softmax}(B\boldsymbol{\eta} + \boldsymbol{c}) \;, \quad A(\boldsymbol{\eta}) = \textstyle\sum_{k=1}^{K} s^{(k)} A^{(k)} \;, \quad [Q(\boldsymbol{\eta})]^{-1} = \textstyle\sum_{k=1}^{K} s^{(k)} (Q^{(k)})^{-1} \;, \tag{2}$$

where $A^{(k)}, Q^{(k)} \in \mathbb{R}^{M \times M}$ for $k = 1, \ldots, K$. The standard discrete SLDS is the limiting case using only the linear regimes in the corners of the simplex, i.e. $\boldsymbol{s}$ is constrained to be one-hot. The *factor* definition is given by

$$A(\boldsymbol{\eta}) = A^{(0)} + \textstyle\sum_{k=1}^{K} \eta^{(k)} A^{(k)} \;, \quad \boldsymbol{s} = \text{Softplus}(B\boldsymbol{\eta} + \boldsymbol{c}) \;, \quad [Q(\boldsymbol{\eta})]^{-1} = \textstyle\sum_{k=1}^{K} s^{(k)} (Q^{(k)})^{-1} \;. \tag{3}$$

In this work we only consider the *factor* variant and leave the *simplex* for consideration in future work. All told, the model parameters are $\theta = \{C, \boldsymbol{b}, R, B, \boldsymbol{c}, \{A^{(k)}\}_{k=0}^{K}, \{Q^{(k)}\}_{k=0}^{K}, \{\vartheta_k\}_{k=0}^{K}, Q_{\text{IC}}\}$ where the ϑ's denote the η GP kernel parameters, discussed below.

Variational Inference and Learning The multiplicative interaction between the $\boldsymbol{\eta}$'s in defining the A matrices and the $\boldsymbol{x}$'s produces a joint distribution of $\boldsymbol{\eta}$'s and $\boldsymbol{x}$'s that is not jointly Gaussian in general, hindering analytic inference. However, we note that the $\boldsymbol{x}$'s and $\boldsymbol{y}$'s form a linear-Gaussian state space model *conditioned on* a setting of $\boldsymbol{\eta}_{1:T-1}$. This suggests a model fitting procedure in which the $\boldsymbol{\eta}$'s are sampled from an approximate posterior, the log likelihood of the observations conditioned on the sampled $\boldsymbol{\eta}$ is calculated, and the model parameters are updated by stochastic gradient ascent.

Let $q(\boldsymbol{\eta}_{1:T-1}; \phi)$ denote an arbitrary joint distribution over the $\boldsymbol{\eta}$'s with parameters ϕ that admits sampling and density evaluation. We can lower bound the marginal observation log likelihood using the standard evidence lower bound, or ELBO:

$$\mathcal{L}(\boldsymbol{y}_{1:T}; \theta, \phi) \triangleq \mathbb{E}_{\boldsymbol{\eta}_{1:T-1} \sim q(\boldsymbol{\eta}_{1:T-1})} \log \left[\frac{p(\boldsymbol{y}_{1:T}, \boldsymbol{\eta}_{1:T-1})}{q(\boldsymbol{\eta}_{1:T-1})} \right] \leq \log p(\boldsymbol{y}_{1:T}) \;. \tag{4}$$

$q(\boldsymbol{\eta}_{1:T-1}; \phi)$ is known as an approximate posterior because the inequality becomes equality when $q(\boldsymbol{\eta}_{1:T-1}; \phi) = p(\boldsymbol{\eta}_{1:T-1}|\boldsymbol{y}_{1:T}; \theta)$, the true posterior. The following sections detail choices that allow for the efficient estimation of an ELBO, which will end up being a lower bound of the standard ELBO in Eq. 4.

The η Ornstein-Uhlenbeck GP We assume K independent GP priors, one for each component of $\boldsymbol{\eta}$:

$$\boldsymbol{\eta}^{(k)} \sim \mathcal{GP}(\mathbf{0}, \kappa_k) \ , \quad k = 1, \ldots, K \tag{5}$$

with kernel $\kappa_k(t, t') = \exp(-\vartheta_k |t - t'|)$ where $\vartheta_k \in \mathbb{R}^+$. This defines a real-valued Ornstein-Uhlenbeck process, a Markovian Gaussian process. The Markov property allows us to write the precision matrix of a set of equally spaced observations as a tridiagonal matrix:

$$p\left(\begin{bmatrix} \eta_1^{(k)} \\ \vdots \\ \eta_{T-1}^{(k)} \end{bmatrix}\right) = \mathcal{N}(\mathbf{0}, \Lambda(\vartheta_k)^{-1}), \quad \text{where } \Lambda(\vartheta)_{ij} = \begin{cases} \frac{1}{1-a^2} & \text{if } i = j = 0 \text{ or } i = j = T-1 \\ \frac{a^2+1}{1-a^2} & \text{if } i = j \text{ and } i \notin \{1, T-1\} \\ \frac{a}{a^2-1} & \text{if } |i-j| = 1 \\ 0 & \text{otherwise} \end{cases}, \tag{6}$$

and $a = \exp(-\vartheta)$. All together, the $\boldsymbol{\eta}$'s are assumed to be drawn from a zero-mean Gaussian with precision $\Lambda_{\text{prior}} = \sum_k \boldsymbol{e}_k \boldsymbol{e}_k^\top \otimes \Lambda(\vartheta_k)$, a block-diagonal matrix with K tridiagonal $(T-1)$-by-$(T-1)$ blocks.

Conditioning the OU GPs We condition the prior OU GPs to form an approximate posterior using precision and precision-mean *pseudo-observations*, which constitute ϕ, our approximate posterior parameters: $\phi = \{\Lambda_{\text{pseudo}}, \boldsymbol{r}_{\text{pseudo}}\}$ where $\boldsymbol{r}_{\text{pseudo}} \in \mathbb{R}^{K(T-1)}$ and $\Lambda_{\text{pseudo}} \in \mathbb{R}^{K(T-1) \times K(T-1)}$ is diagonal. These pseudo-observations are inducing points placed at each timestep and condition the prior GP to form the approximate posterior GP as if we had direct, uncertain observations of the $\boldsymbol{\eta}$'s at each timepoint. The approximate posterior is then taken as $q(\boldsymbol{\eta}_{1:T-1}) \sim \mathcal{N}(\mu_{\text{post}},\ \Lambda_{\text{post}}^{-1})$ where $\mu_{\text{post}} = \text{Solve}(\Lambda_{\text{post}}, \boldsymbol{r}_{\text{pseudo}})$ and $\Lambda_{\text{post}} = \Lambda_{\text{prior}} + \Lambda_{\text{pseudo}}$. The solve operation, sampling, and log density evaluation can all performed in time $\mathcal{O}(T)$ because Λ_{post} is tridiagonal. The KL divergence from the $\boldsymbol{\eta}$ approximate posterior to its prior is available in closed form as

$$D(q(\boldsymbol{\eta}_{1:T-1}) \,||\, p(\boldsymbol{\eta}_{1:T-1})) = \tfrac{1}{2}\left[\ \text{tr}(\Lambda_{\text{prior}}\Lambda_{\text{post}}^{-1}) - K(T-1) + \mu_{\text{post}}^\top \Lambda_{\text{prior}} \mu_{\text{post}} + \log \tfrac{|\Lambda_{\text{post}}|}{|\Lambda_{\text{prior}}|}\right] . \tag{7}$$

The log determinant term here can be computed in $\mathcal{O}(T)$ time by performing block-tridiagonal Cholesky decompositions on both Λ's. The quadratic form can also be computed in $\mathcal{O}(T)$ time because Λ_{prior} is block-tridiagonal. Lastly, we choose to estimate the trace term using Hutchinson's trace estimator,

$$\text{tr}(\Lambda_{\text{prior}}\Lambda_{\text{post}}^{-1}) = \mathbb{E}_{\boldsymbol{\epsilon} \sim \mathcal{N}(0,I)}\left[\boldsymbol{\epsilon}^\top \Lambda_{\text{prior}} \Lambda_{\text{post}}^{-1} \boldsymbol{\epsilon}\right] \ , \tag{8}$$

where $\Lambda_{\text{post}}^{-1}\boldsymbol{\epsilon}$ can be computed with an $\mathcal{O}(T)$ block-tridiagonal solve operation.

Deriving the likelihood For fixed $\boldsymbol{\eta}$'s and $t = 1, \ldots, T-1$ let $A_t = A(\boldsymbol{\eta}_t)$ and $Q_t = Q(\boldsymbol{\eta}_t)$. Then the precision of the $\boldsymbol{x}$ latents, $\text{cov}(\boldsymbol{x}_{1:T})^{-1} = \Sigma_{xx}^{-1} = \Lambda_{xx}$ is given by

$$\begin{bmatrix} Q_{\text{IC}} + A_1^\top Q_1 A_1 & -A_1^\top Q_1 & & & \\ -Q_1 A_1 & Q_1 + A_2^\top Q_2 A_2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & Q_{T-2} + A_{T-1}^\top Q_{T-1} A_{T-1} & -A_{T-1}^\top Q_{T-1} \\ & & & -Q_{T-1} A_{T-1} & Q_{T-1} \end{bmatrix} . \tag{9}$$

The corresponding covariance matrix of the observations is

$$\text{cov}(\boldsymbol{y}_{1:T}) = \Sigma_{yy} = \Lambda_{yy}^{-1} = (C \otimes I_T)\Sigma_{xx}(C^\top \otimes I_T) + (R \otimes I_T) \ . \tag{10}$$

To evaluate the likelihood of the data under a Gaussian distribution with this covariance matrix, we need efficient methods to estimate both its log determinant and evaluate the quadratic form $\boldsymbol{y}^\top \Lambda_{yy} \boldsymbol{y}$. In the Appendix, we derive a log determinant bound $\log|\Sigma_{yy}| \leq T \log|R| + T\ \text{tr}(C^\top R^{-1} C)$ and describe an efficient computational method for the quadratic form. This results in the following likelihood lower bound:

$$\begin{aligned} \log p(\boldsymbol{y}|\boldsymbol{\eta}_{1:T-1}) &= -\frac{1}{2}\left[\text{logdet}(\Sigma_{yy}) + \boldsymbol{y}^\top \Lambda_{yy} \boldsymbol{y} + NT \log(2\pi)\right] \\ &\geq -\frac{1}{2}\left[T\ \text{logdet}(R) + T\ \text{tr}(C^\top R^{-1} C) + \boldsymbol{y}^\top \Lambda_{yy} \boldsymbol{y} + NT \log(2\pi)\right] \end{aligned} \tag{11}$$

Eqs. 7 and 11 together form our ELBO objective, which can be estimated in $\mathcal{O}(T)$ time.

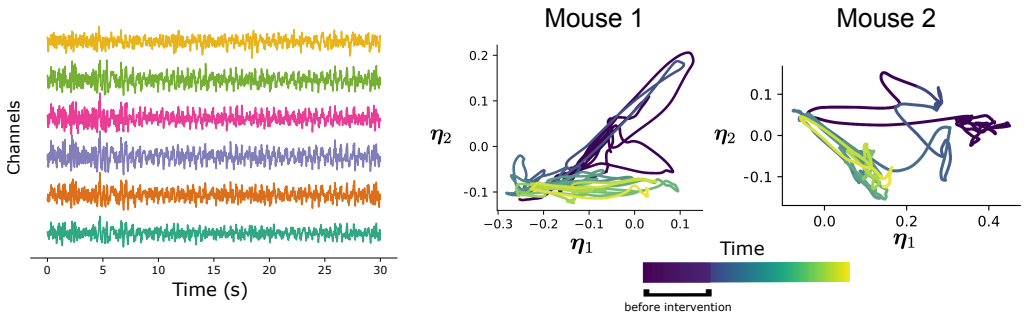

Figure 2: Left: Two CSLDS models were trained on six-channel, 2000 second local field potentials recorded from two mice before an after undergoing a pharmacological intervention. Right: The inferred $\boldsymbol{\eta}$ traces clearly distinguish between the dynamics before the intervention (dark purple) and after (lighter colors).

Identifiability of $\boldsymbol{\eta}$'s We assert under certain conditions, namely that the A matrices have distinct eigenvalues and are sufficiently close to some A_0, that the true $\boldsymbol{\eta}$'s underlying a CSLDS are identifiable up to linear transformation in the slow-$\boldsymbol{\eta}$ limit. Here we present a brief logical sketch and in Figure 1 we present empirical evidence of this claim, leaving a formal proof for future work. In the slow-$\boldsymbol{\eta}$ limit, we may window the $\boldsymbol{y}$ timeseries and fit a separate LDS model for each window corresponding to a specific $\boldsymbol{\eta} \in \mathbb{R}^K$. In general, the dynamics matrix A in an LDS is only identifiable up to matrix similarity [1], which reduces to identifiability up to eigenvalues given the distinct eigenvalue assumption. Under the small perturbation assumption, we can match eigenvalues across windows unambiguously and identify a subspace of eigenvalues. Lastly, in this low perturbation regime, we can use first-order eigenvalue perturbations to confirm that the true $\boldsymbol{\eta}$'s are given by a linear transformation of coordinates in this eigenvalue subspace.

3 Experiments

Synthetic Data As a test of the proposed model, we asked whether the CSLDS could recover ground-truth $\boldsymbol{\eta}$'s in a synthetic dataset. First, we drew "cslds" in cursive (Figure 1, center) to represent the time-evolution of a 2-dimensional $\boldsymbol{\eta}$. Multiple datasets were created that divided the "cslds" path into $T = 3 \cdot 10^3, 10^4, 3 \cdot 10^4$, and 10^5 timesteps. The ground-truth model was randomly initialized with $K, M, N = 2$ and held fixed between experiments. Given the discussion on the identifiability of the $\boldsymbol{\eta}$'s above, we expect that we should be able to infer the correct $\boldsymbol{\eta}$'s up to linear transformation with a large amount of data. We fit a CSLDS model to each timeseries and confirmed that it was able to approximately resconstruct the ground-truth $\boldsymbol{\eta}$'s (Figure 1, right, shown rotated to best match the ground-truth $\boldsymbol{\eta}$'s).

Modeling Multichannel Neural Data We next used CSLDS to model multichannel local field potential data from two freely behaving mice before and after a pharmacological intervention, asking what effect, if any, the intervention has on the inferred $\boldsymbol{\eta}$'s. The data for each mouse comprises 6 channels and 2000 seconds sampled at 50Hz for 10^5 total timesteps and the CSLDS model parameters are $K = 2$ and $M = 4$. The inferred $\boldsymbol{\eta}$'s for both recordings, shown in Figure 2, clearly distinguish the times before the intervention (dark purple) and after the intervention (lighter colors).

4 Conclusion

We propose the continuous switching linear dynamical systems (CSLDS), which traverses a continuum of linear dynamical systems continuously in time. The model admits computationally feasible methods for inference and learning, allowing us to recover ground-truth latents and model neural recordings in mice. One particularly promising direction of future work is to use the CSLDS to study transitions between nominally discrete states, for example, by modeling the neural dynamics underlying sleep state transitions.

References

[1] KS Arun and SY Kung. "Balanced approximation of stochastic systems". In: *SIAM journal on matrix analysis and applications* 11.1 (1990), pp. 42–68.

[2] Julia C Costacurta et al. "Distinguishing discrete and continuous behavioral variability using warped autoregressive HMMs". In: *bioRxiv* (2022).

[3] Emily Fox et al. "Nonparametric Bayesian learning of switching linear dynamical systems". In: *Advances in neural information processing systems* 21 (2008).

[4] Joshua Glaser et al. "Recurrent switching dynamical systems models for multiple interacting neural populations". In: *Advances in neural information processing systems* 33 (2020), pp. 14867–14878.

[5] Scott Linderman et al. "Bayesian learning and inference in recurrent switching linear dynamical systems". In: *Artificial Intelligence and Statistics*. PMLR. 2017, pp. 914–922.

[6] Josue Nassar et al. "Tree-Structured Recurrent Switching Linear Dynamical Systems for Multi-Scale Modeling". In: *International Conference on Learning Representations*. 2019. URL: https://openreview.net/forum?id=HkzRQhR9YX.

[7] Sang Min Oh et al. "Learning and inferring motion patterns using parametric segmental switching linear dynamic systems". In: *International Journal of Computer Vision* 77.1 (2008), pp. 103–124.

[8] Virginia Rutten et al. "Non-reversible Gaussian processes for identifying latent dynamical structure in neural data". In: *Advances in Neural Information Processing Systems* (2020).

ELBO details

This section addresses details necessary to compute the CSLDS ELBO objective efficiently.

An efficient log determinant bound Here we derive an upper bound for the log determinant of the data covariance, Σ_{yy}. We begin with a derivation that applies generally to the "GP factor analysis" literature, and can be found in [8]. In what follows, we prioritize the common case $M < N \ll T$. Applying the Woodbury matrix inversion lemma to Eq. 10, we get

$$\begin{aligned}\Lambda_{yy} &= (R^{-1} \otimes I_T) - (R^{-1}C \otimes I_T)(L^{-\top}L^{-1} + C^\top R^{-1} C \otimes I_T)^{-1}(C^\top R^{-1} \otimes I_T) \\ &= (R^{-\frac{1}{2}} \otimes I_T)(I_{NT} - A^\top B^{-1} A)(R^{-\frac{1}{2}} \otimes I_T)\ ,\end{aligned} \tag{12}$$

where $LL^\top = \Sigma_x$, $A = L^\top(C^\top R^{-\frac{1}{2}} \otimes I_T)$, and $B = I + AA^\top$. We can then simplify the log determinant:

$$\begin{aligned}\text{logdet}(\Lambda_{yy}) &= 2\ \text{logdet}(R^{-1/2} \otimes I_T) + \ \text{logdet}(I - A^\top B^{-1} A) \\ &= -T\ \text{logdet}(R) + \ \text{logdet}(B - AA^\top) - \ \text{logdet}(B) \\ &= -T\ \text{logdet}(R) + \ \text{logdet}(I) - \ \text{logdet}(B) \\ &= -T\ \text{logdet}(R) - \ \text{logdet}(I + L^\top(C^\top R^{-1} C \otimes I_T)L) \\ \Rightarrow \quad \text{logdet}(\Sigma_{yy}) &= T\ \text{logdet}(R) + \ \text{logdet}(I + L^\top(C^\top R^{-1} C \otimes I_T)L)\end{aligned} \tag{13}$$

Note that the matrix in the second logdet term is MT-by-MT, so a naive logdet calculation would take time $\mathcal{O}(M^3T^3)$, which is an improvement on the $\mathcal{O}(N^3T^3)$ required for the full NT-by-NT matrix, but still prohibitive for large T. Instead of performing an exact log determinant calculation, we deviate from the GP factor analysis approach and opt to upper bound the log determinant using the identity $\text{logdet}(A) \leq \ \text{tr}(A - I)$:

$$\begin{aligned}\text{logdet}(\Sigma_{yy}) &\leq T\ \text{logdet}(R) + \ \text{tr}(L^\top(C^\top R^{-1} C \otimes I_T)L) \\ &= T\ \text{logdet}(R) + \sum_m [C^\top R^{-1} C]_{mm}\ \text{tr}(L_m L_m^\top) \\ &= T\ \text{logdet}(R) + \sum_m [C^\top R^{-1} C]_{mm}\ \text{tr}(\Sigma_{xx,m}) \\ &= T(\ \text{logdet}(R) + \sum_m [C^\top R^{-1} C]_{mm}) \\ &= T[\ \text{logdet}(R) + \ \text{tr}(C^\top R^{-1} C)]\end{aligned} \tag{14}$$

Note that $\text{tr}(\Sigma_{xx,m}) = T$ relies on our choice of kernel, which guarantees $\kappa(t,t) = 1$. If we restrict R to be diagonal, this log determinant bound can be calculated in time $\mathcal{O}(MN)$, which is notably independent of T, the number of timepoints.

In what situations is this bound tight? It can be easily checked that the bound $\text{logdet}(A) \leq \ \text{tr}(A - I)$ is tight when $A = I$. The bound can be tightened in a flexible way by using a lower triangular matrix W. We have:

$$\text{logdet}(A) = \ \text{logdet}(WW^\top A) - 2\ \text{logdet}(W) \leq \ \text{tr}(WW^\top A - I) - 2\ \text{logdet}(W)\ . \tag{15}$$

The log determinant of W can be calculated efficiently because W is triangular. The bound is tight when $WW^\top A = I$, and for a positive definite A, which is our use case, we are guaranteed the existence of a W such that $WW^\top A = I$. This suggests an optimization procedure in which W is a free parameter to be optimized that keeps the log determinant bound tight during the course of model training. We leave an implementation and evaluation of this adaptive log determinant bound strategy to future work.

Quadratic form calculation A naive calculation of the quadratic form $\boldsymbol{y}^\top \Lambda_{yy} \boldsymbol{y}$ takes time $\mathcal{O}(N^2T^2)$, which is prohibitive for large T. The key step is to perform matrix/vector products with

$$\Lambda_{yy} = (R^{-\frac{1}{2}} \otimes I_T)(I_{NT} - A^\top B^{-1} A)(R^{-\frac{1}{2}} \otimes I_T)\ . \tag{16}$$

Multiplication with the first and third terms are easy. Remember $A = L^\top (C^\top R^{-\frac{1}{2}} \otimes I_T)$. Matrix/vector multiplication with $(C^\top R^{-\frac{1}{2}} \otimes I_T)$ is easy, but we do not have direct access to $L^\top$. However, we can efficiently find $L^{-\top} L^{-1} = \Lambda_{xx}$ by a Cholesky decomposition because Λ_{xx} is block-tridiagonal. Then we have $L^\top \boldsymbol{b} = \text{Solve}(L^{-\top}, \boldsymbol{b})$, with an efficient block-bidiagonal solve. Therefore, $A\boldsymbol{b}$ can be efficiently calculated as $\text{Solve}(L^{-\top}, (C^\top R^{-\frac{1}{2}} \otimes I_T)\boldsymbol{b})$.

Now for $B^{-1}\boldsymbol{b}$. We can write $B = L^\top [C^\top R^{-1} C \otimes I_T + \Lambda_{xx}] L$. Then

$$B^{-1} = L^{-1} [C^\top R^{-1} C \otimes I_T + \Lambda_{xx}]^{-1} L^{-\top} \ , \tag{17}$$

and as we saw before, matrix/vector multiplication with L^{-1} and $L^{-\top}$ is efficient. If P is the permutation matrix that switched the order of the Kronecker product expansion, then $P(C^\top R^{-1} C \otimes I_T)$ is block-diagonal with T separate $C^\top R^{-1} C$ blocks and $P\Lambda_{xx}$ is symmetric block-tridiagonal with diagonal blocks. It follows that $P[C^\top R^{-1} C \otimes I_T + \Lambda_{xx}]$ is block-tridiagonal with $3T-2$ separate M-by-M blocks. Solve operations can be performed efficiently with block tridiagonal matrices, $\mathcal{O}(M^3 T)$ in this case. Note the permutation P can be applied as a $\mathcal{O}(1)$ reshaping operation as opposed to a $\mathcal{O}(M^2 T^2)$ matrix/vector multiplication.

To summarize, the quadratic form calculation is given by:

1. $\boldsymbol{b} \leftarrow (R^{-\frac{1}{2}} \otimes I_T)\boldsymbol{y}$
2. $\widetilde{\boldsymbol{b}} \leftarrow \text{Solve}(L^{-\top}, (C^\top R^{-\frac{1}{2}} \otimes I_T)\boldsymbol{b})$
3. $\widetilde{\boldsymbol{b}} \leftarrow P^\top \text{Solve}(P[C^\top R^{-1} C \otimes I_T + \Lambda_{xx}], \widetilde{\boldsymbol{b}})$
4. $\widetilde{\boldsymbol{b}} \leftarrow (R^{-\frac{1}{2}} C \otimes I_T)\text{Solve}(L^{-1}, \widetilde{\boldsymbol{b}})$
5. $\boldsymbol{b} \leftarrow \boldsymbol{b} - \widetilde{\boldsymbol{b}}$
6. $\boldsymbol{b} \leftarrow (R^{-\frac{1}{2}} \otimes I_T)\boldsymbol{b}$
7. Return $\boldsymbol{y}^\top \boldsymbol{b}$

ELBO estimation The complete CSLDS objective function is

$$-\frac{1}{2}\left[T \,\text{logdet}(R) + T \,\text{tr}(C^\top R^{-1} C) + \boldsymbol{y}^\top \Lambda_{yy} \boldsymbol{y} + NT \log(2\pi)\right] - D(q(\boldsymbol{\eta}_{1:T-1}) \,||\, p(\boldsymbol{\eta}_{1:T-1})) \\ \leq \mathcal{L}(\boldsymbol{y}_{1:T}; \theta, \phi) \leq \log p(\boldsymbol{y}_{1:T}) \ , \tag{18}$$

where the KL divergence $D(q(\boldsymbol{\eta}_{1:T-1}) \,||\, p(\boldsymbol{\eta}_{1:T-1}))$ is given in Eq. 7. This is a lower bound to the standard ELBO $\mathcal{L}(\boldsymbol{y}_{1:T})$ from Eq. 4 (due to the Σ_{yy} log determinant bound in Eq. 14), which is in turn a lower bound to the observation marginal log likelihood $\log p(\boldsymbol{y}_{1:T})$.