Max–Planck–Institut für biologische Kybernetik
Max Planck Institute for Biological Cybernetics

Technical Report No. 161

Dirichlet Mixtures of Bayesian Linear Gaussian State-Space Models: a Variational Approach

Silvia Chiappa[1], David Barber[2]

March 2007

[1] Department Schölkopf, email: silvia.chiappa@tuebingen.mpg.de; [2] University College London, Department of Computer Science, email: d.barber@cs.ucl.ac.uk

Dirichlet Mixtures of Bayesian Linear Gaussian State-Space Models: a Variational Approach

Silvia Chiappa, David Barber

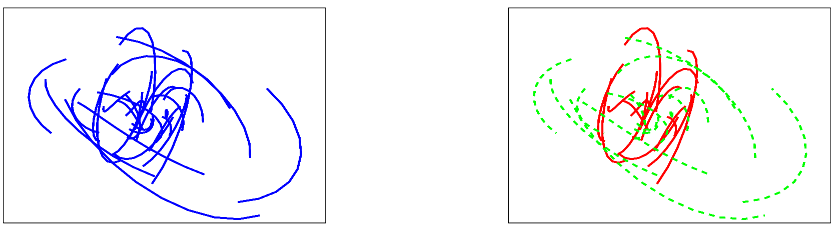
Abstract. We describe two related models to cluster multidimensional time-series under the assumption of an underlying linear Gaussian dynamical process. In the first model, times-series are assigned to the same cluster when they show global similarity in their dynamics, while in the second model times-series are assigned to the same cluster when they show simultaneous similarity. Both models are based on Dirichlet Mixtures of Bayesian Linear Gaussian State-Space Models in order to (semi) automatically determine an appropriate number of components in the mixture, by biasing the components to a parsimonious parameterization. The resulting models are formally intractable and to deal with this we describe a deterministic approximation based on a novel implementation of Variational Bayes.

1 Introduction

Clustering is a large topic in machine learning and related areas, and the aim of this report is to provide an additional methodology which may be better suited to applications in which a dynamical system is believed to be underlying the data. This situation is common in time-series based on natural phenomena, since the equations believed to be describing the physical world can often be modeled as Markovian dynamics on an underlying state-system.

Perhaps the most straightforward approach to perform clustering of a set of time-series is to consider each time-series as a 'static' vector, thereby transforming the time-series clustering problem into a more standard clustering-of-vectors problem. For clustering vectors, any of a number of methods may be applied, ranging from K-Means [1], to more recent methods based on probabilistic mixture models [2, 3, 4, 5]. However, for long or high dimensional time-series, this approach becomes computationally problematic, and therefore features of the signal are used instead, typically extracted from short windows of the time-series. Each time-series is then represented by either a single, or set of feature vectors, which may be then clustered by any standard static clustering technique [6]. However, it is not always clear what the appropriate feature extraction method should be. For example, times-series generated by the same dynamical system with different initial conditions can look highly dissimilar (see Fig. 1a). In such cases, it not obvious which features should be used to perform clustering. On the other hand, a method which can explicitly model the dynamics of the time-series would be able to perform clustering without the need of preliminary feature extraction.

Our method is therefore motivated by the desiderata to perform clustering by explicitly modeling the dynamics of the time-series. Furthermore, we are interested in the case in which the number of clusters is not known in



(a) Unclustered Trajectories

(b) Clustered Trajectories

Figure 1: (a) Thirty trajectories length $T = 10$ resulting from the dynamics of two different LGSSMs, both with hidden dimension $H = 4$. Plotted are the points $([v_t]_1, [v_t]_2)$. (b) The trajectories as labeled by our algorithm – all labels are consistent with the generating mechanism.

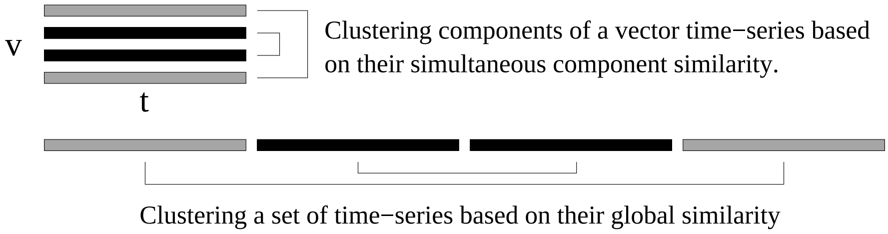


Figure 2: (Top) Time-series clustering based on simultaneous similarity versus (Bottom) time-series clustering based on global similarity.

advance, and therefore in a model which can automatically determine an appropriate number of clusters. To prevent overfitting, we would also like to encourage each cluster to be described by a parsimonious parameterization.

Recently, there has been considerable interest in Dirichlet Process Mixture Models [2, 3, 4, 5] which provide a (semi) automatic way to determine an appropriate number of clusters. The Dirichlet mixture results in a Polya distribution on the cluster assignments, enabling an appropriate number of clusters to be found by the model. In a continued theme of development of this class of techniques, our time-series model will be based on a Dirichlet Mixture of Bayesian Linear Gaussian State-Space models (DMBLGSSMs). The Bayesian approach places a prior on the parameters of each mixture component, encouraging them to have the smallest parameterization consistent with the data.

We will give two main variants of the DMBLGSSM, in order to perform clustering based on either *simultaneous* or *global* similarity of the time-series dynamics. The distinction between these two approaches will be spelled out mathematically in Section 4, whilst an informal sketch is presented in Fig. 2:

Simultaneous Similarity Our simultaneous similarity clustering approach assigns two time-series to the same cluster if they are derived from the *same realization* of a dynamical process.

Global Similarity The global similarity method will assign two time-series to the same cluster if they are generated by *different realizations* of the same dynamical process.

The resulting Bayesian time-series clustering model is formally computationally intractable, and therefore approximations need to be considered. To the best of our knowledge, whilst sampling methods have been applied in a similar more constrained context [7], a Variational Bayesian treatment of this class of models is new, as is our application to clustering based on simultaneous dynamical similarity. Variational Bayes is a deterministic approximation scheme which has the potential advantage of speed over sampling techniques.

We will first describe the BLGSSM in Section 2 and the general procedure of Dirichlet Process mixture models in Section 3. We will then marry the two in Section 4 to form the clustering methods based on simultaneous and global dynamical similarity. In Section 5 we will give an illustrative demonstration of the performance of the two approaches.

2 Bayesian Linear Gaussian State-Space Models

In a Linear Gaussian State-Space Model (LGSSM)[1] [8, 9, 10], a sequence of observations $v_{1:T} \equiv v_1, \ldots, v_T$ is generated from an underlying dynamical system on $h_{1:T}$ according to:

$$v_t = Bh_t + \eta_t^v, \;\; \eta_t^v \sim \mathcal{N}(\mathbf{0}_V, \Sigma_V), \qquad h_t = Ah_{t-1} + \eta_t^h, \;\; \eta_t^h \sim \mathcal{N}\left(\mathbf{0}_H, \Sigma_H\right), \;\; h_1 \sim \mathcal{N}\left(\mu, \Sigma\right),$$

where $\mathcal{N}(m, S)$ denotes a Gaussian with mean m and covariance S, and $\mathbf{0}_X$ denotes an X-dimensional zero vector. The observation v_t has dimension V and the hidden state h_t has dimension H. Probabilistically, the LGSSM is defined by:

$$p(v_{1:T}, h_{1:T}|\Theta) = p(v_1|h_1)p(h_1)\prod_{t=2}^{T} p(v_t|h_t)p(h_t|h_{t-1}),$$

with

$$p(v_t|h_t) = \mathcal{N}\left(Bh_t, \Sigma_V\right), \qquad p(h_t|h_{t-1}) = \mathcal{N}\left(Ah_{t-1}, \Sigma_H\right), \qquad p(h_1) = \mathcal{N}(\mu, \Sigma)$$

[1] Also called Kalman Filter/Smoother, Linear Dynamical System.

and where $\Theta = \{A, B, \Sigma_H, \Sigma_V, \mu, \Sigma\}$ denotes the model parameters. Thanks to the simple structure of the model, most quantities of interest, such as the posterior density $p(h_t|v_{1:T}, \Theta)$ and likelihood

$$p(v_{1:T}|\Theta) = \int_{h_{1:T}} p(v_{1:T}, h_{1:T}|\Theta) \tag{1}$$

can be computed efficiently in $O(T)$ operations [11].

In a Bayesian treatment of the model, a parameter prior $p(\Theta|\hat{\Theta})$ is defined, where $\hat{\Theta}$ are the associated hyperparameters, resulting in the marginal likelihood

$$p(v_{1:T}|\hat{\Theta}) = \int_{\Theta, h_{1:T}} p(v_{1:T}, h_{1:T}|\Theta) p(\Theta|\hat{\Theta}). \tag{2}$$

In a full Bayesian treatment we would define additional prior distributions over the hyperparameters $\hat{\Theta}$. Here we take instead the ML-II ('evidence') framework, in which the optimal set of hyperparameters is found by maximizing $p(v_{1:T}|\hat{\Theta})$ with respect to $\hat{\Theta}$ [12, 13, 14].

Whilst the integral required to compute the likelihood (Eq. (1)) is tractable, the result of this integral couples the parameters Θ in the integrand of the marginal likelihood (Eq. (2)). An exact implementation of the Bayesian LGSSM is then formally intractable. Recently Variational Bayes (VB) has been applied to this model as a route to a computationally efficient approximate implementation [12, 13, 14, 15, 16, 17].

The most challenging part of implementing the VB method is performing inference over $h_{1:T}$. Some authors [12, 17] have developed their own specialized routines, based on Belief Propagation (see [18]), since standard LGSSM inference routines appear, at first sight, not to be applicable. However, in [15] is shown how the VB treatment of the LGSSM can be implemented using any standard LGSSM inference routine, including those specifically addressed to improve numerical stability [9, 19, 20]. This approach can also be used for the Dirichlet Mixture case, as we will see below.

For the parameter priors, here we define Gaussians on the elements of A and on the columns of B:

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

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

where $\hat{A}$ and $\hat{B}$ are hyperparameters defining our preferred values for the transition and emission matrices. The dependency of the priors on Σ_H and Σ_V renders the VB implementation feasible. The choice $\hat{A} \equiv 0, \hat{B} \equiv 0$ creates an Automatic Relevance Determination (ARD) bias towards a simple dynamics and eliminates unnecessary parameters of the model [21]. This form of prior, in which individual elements of the transition A and columns on the emission B are biased to be small, is appropriate for the global similarity clustering of Section 4.1[2]. For the simultaneous similarity clustering of Section 4.2 we modify the emission prior by biasing elements of B to be small. The conjugate priors for general inverse covariances Σ_H^{-1} and Σ_V^{-1} are Wishart distributions. In the simpler case of diagonal covariances $\Sigma_H^{-1} = dg\left(\tau\right)$[3] and $\Sigma_V^{-1} = dg\left(\rho\right)$ these become Gamma distributions (see Appendix A.1). For μ, we define $p(\mu|\Sigma) = \mathcal{N}\left(\mu_\mu, \Sigma_\mu\Sigma\right)$, while for Σ we define a Wishart or Gamma prior. The hyperparameters to optimize are then α, β, μ_μ, Σ_μ and the parameters of Gamma or Wishart distributions.

3 Dirichlet Mixture Models

Given a set of observations $v^{1:N} \equiv v^1, \ldots, v^N$, the clustering task is to assign each observation v^n to one of a finite set of cluster labels $k = 1, \ldots, K$. To do so, we introduce a cluster indicator variable $z^n \in \{1, \ldots, K\}$ for each observation. To model the joint cluster allocations we define

$$p\left(z^{1:N}|\gamma\right) = \int_{\pi} \left\{\prod_{n} p\left(z^n|\pi\right)\right\} p(\pi|\gamma), \tag{3}$$

[2] A prior that prefer simultaneously the i^{th} row and column of A to be small would be preferable. However, it is not straightforward to make the VB feasible in this case.

[3] The notation $dg\left(\tau\right)$ indicates diagonal matrix with the elements of the vector τ on the main diagonal.

where $p\left(z^n = k|\pi\right) \equiv \pi_k$, and $p(\pi|\gamma)$ is has symmetric Dirichlet distribution $p(\pi|\gamma) \propto \prod_{k=1}^{K} \pi_k^{\gamma/K-1}$. The integral in Eq. (3) gives rise to the Polya distribution (see Appendix A.3):

$$p\left(z^{1:N}|\gamma\right) = \frac{\Gamma(\gamma)}{\Gamma(N+\gamma)} \prod_{k=1}^{K} \frac{\Gamma(N_k + \gamma/K)}{\Gamma(\gamma/K)}, \tag{4}$$

where $N_k \equiv \sum_{n=1}^{N} I[z^n = k]$ counts the number of times that state k occurs in the indicators[4]. The joint distribution of all observations is then given by

$$p\left(v^{1:N}|\Theta^{1:K}, \gamma\right) = \sum_{z^{1:N}} \left\{ \prod_n p\left(v^n|z^n, \Theta^{1:K}\right) \right\} p\left(z^{1:N}|\gamma\right), \tag{5}$$

where $p\left(v^n|z^n = k, \Theta^{1:K}\right) \equiv p\left(v^n|\Theta^k\right)$ denotes that the parameters of cluster k are used to determine the probability of observation v^n. In the limit of infinite K, the prior expected number of clusters for a set of N sequences is [22]

$$\sum_{n=1}^{N} \frac{\gamma}{\gamma + n - 1} \approx \gamma \log\left(\frac{N}{\gamma} + 1\right),$$

which remains finite for finite γ. In our experiments γ will typically be treated as a hyperparameter and optimized with respect to the marginal likelihood.

In our work, we will consider K to be finite. This is in contrast to Dirichlet *Process* Mixture Models [3, 4, 5], in which the $K \to \infty$ limit is formally taken. This can be achieved, for example, by writing down a sampling algorithm for the finite dimensional case, and then taking the limit $K \to \infty$. If the sampler is initialized with a small number of clusters, the sampling algorithm generates new clusters until sufficiently many are present to explain the data well. In practice, since only a finite number of mixture components is effectively used, we prefer the finite K case. An advantage of this is that we retain an explicit expression for the marginal likelihood which is then amenable to fast deterministic approximation schemes.

Computing the resulting model likelihood in Eq. (5) is intractable, and a useful deterministic approximation is to use a lower bound based on the 'collapsed'[5] variational KL divergence KL $\left(\prod_{n=1}^{N} q\left(z^n\right) || p\left(z^{1:N}|v^{1:N}, \gamma\right)\right)$ [23][6].

4 Dirichlet Mixture of Bayesian LGSSMs

Our approach to clustering is to form a Dirichlet mixture of Bayesian LGSSMs (DMBLGSSMs). This has the advantage of determining the number of clusters, where each cluster may also be biased towards a dynamical system of a preferred form. We will start by describing a model for time-series clustering based on global similarity. We will then introduce a clustering method based on simultaneous similarity as a modification of this first model.

4.1 Clustering based on Global Similarity

The graphical representation of Dirichlet Mixture of Bayesian LGSSMs is given in Fig. 3. The likelihood term $p\left(v^n|z^n, \Theta^{1:K}\right)$ for each temporal sequence $v^n \equiv v^n_{1,\ldots,T}$ in Eq. (5) is defined by the likelihood of the LGSSM (Eq. (1)). The approach we take to extend this to a Dirichlet mixture of the *Bayesian* LGSSMs is to introduce a distribution q, for which we assume that the following factorizations hold:

$$\begin{aligned} q\left(h^{1:N}_{1:T}|z^{1:N}, \Theta^{1:K}\right) &\equiv q\left(h^{1:N}_{1:T}|z^{1:N}\right) \\ q\left(z^{1:N}, \Theta^{1:K}\right) &\equiv q\left(z^{1:N}\right) q\left(\Theta^{1:K}\right) \\ q\left(z^{1:N}\right) &\equiv \prod_{n=1}^{N} q\left(z^n\right). \end{aligned}$$

[4] $I[a = b] = 1$ if $a = b$ and 0 otherwise.

[5] The 'uncollapsed' joint approximation of $p\left(z^{1:N}, \pi|v^{1:N}\right)$ is seductive since it is more straightforward to compute the KL divergence in the joint $(z^{1:N}, \pi)$ space under the factorized assumption $q(z^{1:N})q(\pi)$. However, due to the strong implicit coupling between $z^{1:N}$ and π, this factorized approximation can be insufficiently accurate [23]. However, we experimentally observed that, in our model, the two different approaches perform similarly.

[6] Here and in the rest of the report (if this does not cause confusion), we omit the conditioning on the observations for q.

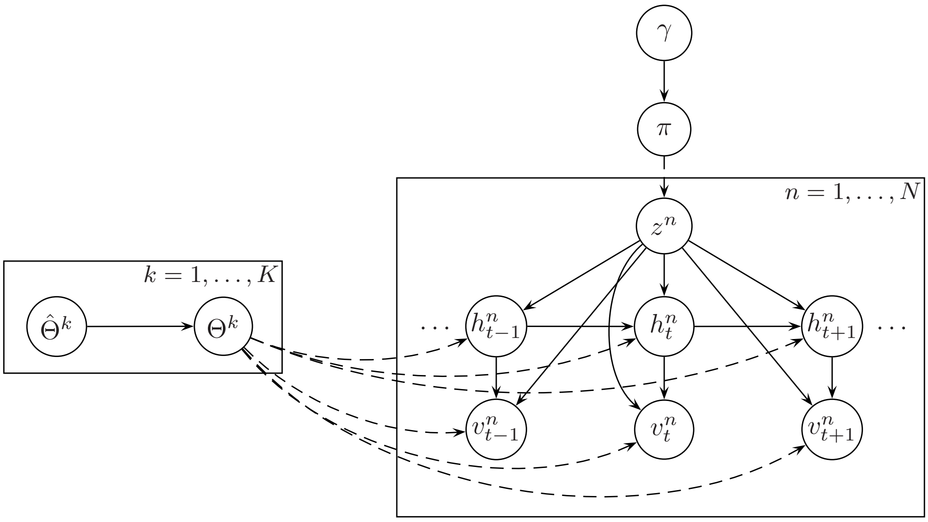


Figure 3: Graphical representation of the Dirichlet Mixture of Bayesian LGSSMs for performing clustering based on global similarity.

We then consider the variational approximation[7]:

$$p\left(z^{1:N}, h^{1:N}_{1:T}, \Theta^{1:K} | v^{1:N}_{1:T}, \hat{\Theta}^{1:K}, \gamma\right) \approx \left\{\prod_{n=1}^{N} q\left(z^n\right) q\left(h^n_{1:T}|z^n\right)\right\} \prod_{k=1}^{K} q\left(\Theta^k\right).$$

Taking the KL divergence between the right and left hand sides of the above then gives the following lower bound on the log-likelihood, $\log p\left(v^{1:N}_{1:T}|\hat{\Theta}^{1:K}, \gamma\right) \geq \mathcal{F}\left(\hat{\Theta}^{1:K}, \gamma, q\right)$, of the Dirichlet mixture of Bayesian LGSSMs:

$$\begin{aligned}\mathcal{F} \equiv & \sum_{k=1}^{K} H_q(\Theta^k) + \sum_{n=1}^{N}\sum_{k=1}^{K} q\left(z^n = k\right) H_q\left(h^n_{1:T}|z^n = k\right) + \sum_{n=1}^{N} H_q\left(z^n\right) + \sum_{k=1}^{K}\left\langle \log p\left(\Theta^k|\hat{\Theta}^k\right)\right\rangle_{q(\Theta^k)} \\ & + \left\langle \log p\left(z^{1:N}|\gamma\right)\right\rangle_{\prod_n q(z^n)} + \sum_{n=1}^{N}\sum_{k=1}^{K} q(z^n = k)\left\langle \log p\left(v^n_{1:T}, h^n_{1:T}|\Theta^k\right)\right\rangle_{q(\Theta^k) q\left(h^n_{1:T}|z^n = k\right)}, \end{aligned} \tag{6}$$

where $H_q(x)$ denotes the entropy of the distribution $q(x)$, and $\langle\cdot\rangle_q$ denotes expectation with respect to q.

Variational Bayes then proceeds by iteratively maximizing the lower bound with respect to the q distributions for fixed hyperparameters $\hat{\Theta}, \gamma$ and vice-versa until no further improvement is found. The resulting updates for q are given by:

$$\begin{aligned} q\left(\Theta^k\right) &\propto p\left(\Theta^k|\hat{\Theta}^k\right) e^{\sum_{n=1}^{N} q(z^n = k)\left\langle \log p\left(v^n_{1:T}, h^n_{1:T}|\Theta^k\right)\right\rangle_{q\left(h^n_{1:T}|z^n = k\right)}} \\ q\left(z^n = k\right) &\propto e^{H_q\left(h^n_{1:T}|z^n = k\right) + \left\langle \log p\left(z^n = k|z^{\neg n}, \gamma\right)\right\rangle_{\prod_{m \neq n} q(z^m)} + \left\langle \log p\left(v^n_{1:T}, h^n_{1:T}|\Theta^k\right)\right\rangle_{q\left(h^n_{1:T}|z^n = k\right) q(\Theta^k)}} \\ q\left(h^n_{1:T}|z^n = k\right) &\propto e^{\left\langle \log p\left(v^n_{1:T}, h^n_{1:T}|\Theta^k\right)\right\rangle_{q(\Theta^k)}} \end{aligned}$$

where $z^{\neg n}$ indicates all indicator variables except for z^n. We will discuss each specific update below. Full details are given in Appendix B.2.

Missing Observations

One of the advantages using a LGSSM model for each cluster is the ease with which missing observations can be dealt with. Indeed if the multidimensional vector v_t has some missing components corresponding to unobserved

[7] From the assumptions on p and q, it follows that the q that maximizes the lower bound $\mathcal{F}$ (Eq. (6)) satisfies $q\left(h^{1:N}_{1:T}|z^{1:N}\right) = \prod_{n=1}^{N} q\left(h^n_{1:T}|z^n\right)$ and $q\left(\Theta^{1:K}\right) = \prod_{k=1}^{K} q\left(\Theta^k\right)$ (see Appendix B.1).

or corrupted measurements, we can integrate these out in the likelihood. This can be formally introduced in our model by the replacement $B \leftarrow W_t^n B$ in the bound $\mathcal{F}$, where W_t^n is the identity matrix with diagonal elements corresponding to missing observation components replaced by zeros, and by replacing missing components in v_t^n by zeros. We will give the updates for this general framework.

Throughout, we assume that Σ_V is diagonal in the case in which there are missing observations. This assumption makes the formula for the mean update independent of Σ_V, which is a computational convenience.

Updates for $q\left(B^k, \left[\Sigma_V^k\right]^{-1}\right)$

To simplify the notation, in the following (and elsewhere where clear from the context) we will omit the dependency of the model parameters Θ^k and hyperparameters $\hat{\Theta}^k$ on the mixture k. The choice Normal-Wishart(Gamma) prior $p\left(B, \Sigma_V^{-1}\right)$ give rise to a Normal-Wishart(Gamma) approximated posterior $q\left(B, \Sigma_V^{-1}\right)$:

$$q\left(B, \Sigma_V^{-1}\right) \propto e^{\sum_{n=1}^N q(z^n=k) \sum_{t=1}^T \left\langle \log p\left(v_t^n | h_t^n, B, \Sigma_V^{-1}\right)\right\rangle_{q(h_t^n|z^n=k)}} p\left(B|\beta, \Sigma_V^{-1}\right) p\left(\Sigma_V^{-1}|\hat{\Theta}\right).$$

In the following, we decompose the joint $q\left(B, \Sigma_V^{-1}\right) \equiv q\left(B|\Sigma_V^{-1}\right) q\left(\Sigma_V^{-1}\right)$.

$q\left(B|\Sigma_V^{-1}\right)$**:** In Appendix B.2.1 we show that $q\left(vc\left(B\right)|\Sigma_V^{-1}\right) = \mathcal{N}\left(\mu_B, \Sigma_B\right)$ where[8]:

$$\Sigma_B = \left(I_H \otimes \Sigma_V\right) \left(\underbrace{\sum_{n=1}^N q(z^n = k) \sum_{t=1}^T \left\langle h_t^n (h_t^n)^\mathsf{T} \right\rangle_{q(h_t^n|z^n=k)} \otimes W_t^n + dg\left(\beta\right) \otimes I_V}_{H_{BM}} \right)^{-1} \tag{7}$$

$$\mu_B = H_{BM}^{-1} vc \left(\underbrace{\sum_{n=1}^N q(z^n = k) \sum_{t=1}^T W_t^n v_t^n \left\langle h_t^n \right\rangle^\mathsf{T}_{q(h_t^n|z^n=k)} + \hat{B} dg\left(\beta\right)}_{N_B} \right).$$

We remind the reader that in the case of missing observations, Σ_V in Eq. (7) is constrained to be diagonal. For the case of no missing observations, $W_t^n = I_V$, Eq. (7) simplifies, as shown in Appendix B.2.1.

$q\left(\Sigma_V^{-1}\right)$**:** For the case in which there are no missing observations $W_t^n = I_V$, we may consider the general Wishart prior $p(\Sigma_V^{-1}|\nu_V, S_V) = \mathcal{W}(\nu_V, S_V)$, for which the updates are (see Appendix B.2.2):

$$q(\Sigma_V^{-1}) = \mathcal{W}\left(\nu_V + T\sum_{n=1}^N q(z^n = k), \left(S_V^{-1} + \sum_{n=1}^N q(z^n = k) \sum_{t=1}^T v_t^n (v_t^n)^\mathsf{T} - N_B H_B^{-1} N_B^\mathsf{T} + \hat{B} dg\left(\beta\right) \hat{B}^\mathsf{T} \right)^{-1}\right),$$

with

$$H_B \equiv \sum_{n=1}^N q(z^n = k) \sum_{t=1}^T \left\langle h_t^n (h_t^n)^\mathsf{T} \right\rangle_{q(h_t^n|z^n=k)} + dg\left(\beta\right).$$

Under the simpler diagonal constraint $\Sigma_V^{-1} = dg\left(\rho\right)$, where each diagonal element ρ_i follows a Gamma prior $\mathcal{G}(b_1^i, b_2^i)$, the optimal updates are (see Appendix B.2.2):

$$q(\rho_i) = \mathcal{G}\left(b_1^i + \frac{T}{2}\sum_{n=1}^N q(z^n = k), b_2^i + \frac{1}{2}\left(\sum_{n=1}^N q(z^n = k) \sum_{t=1}^T \left[v_t^n\right]_i^2 - \left[G_B\right]_{ii} + \sum_j \beta_j \hat{B}_{ij}^2\right)\right), \tag{8}$$

where

$$G_B \equiv M_B N_B^\mathsf{T}, \quad M_B \equiv N_B H_B^{-1} \tag{9}$$

[8] $vc\left(B\right)$ denotes the vector formed by stacking the columns of the matrix B. This column vector formulation simplifies mathematical notations. $\otimes$ indicates the Kronecker product and I_X is the identity matrix of dimension $X \times X$.

For the case of missing observations, we are restricted to the above diagonal covariance constraint, in which case we replace M_B in Eq. (9) by the $V \times H$ matrix

$$M_{BM} = mc\left(H_{BM}^{-1} vc\left(N_B\right)\right)$$

where $mc\left(x\right)$ denotes reshaping the vector to form a matrix by reverse column-stacking.

Updates for $q\left(A^k, \left[\Sigma_H^k\right]^{-1}\right)$

The optimal $q\left(A, \Sigma_H^{-1}\right)$ is given by:

$$q\left(A, \Sigma_H^{-1}\right) \propto e^{\sum_{n=1}^{N} q(z^n=k) \sum_{t=2}^{T} \left\langle \log p\left(h_t^n | h_{t-1}^n, A, \Sigma_H^{-1}\right)\right\rangle_{q\left(h_{t-1:t}^n | z^n=k\right)}} p\left(A|\alpha, \Sigma_H^{-1}\right) p\left(\Sigma_H^{-1}|\hat{\Theta}\right).$$

As above, without loss of generality, we decompose this as $q(A|\Sigma_H^{-1})q(\Sigma_H^{-1})$.

$q\left(vr\left(A\right)|\Sigma_H^{-1}\right)$**:** In Appendix B.2.3 we show that $q\left(vr\left(A\right)|\Sigma_H^{-1}\right) = \mathcal{N}\left(\mu_A, \Sigma_A\right)$ where[9]:

$$\Sigma_A = bdg\left(H_{1A}^{-1}\left[\Sigma_H\right]_{11}, \ldots, H_{HA}^{-1}\left[\Sigma_H\right]_{HH}\right)$$
$$\mu_A = vert\left(\left(\left[N_A\right]_{1'} H_{1A}^{-1}\right)^{\mathsf{T}}, \ldots, \left(\left[N_A\right]_{H'} H_{HA}^{-1}\right)^{\mathsf{T}}\right)$$

where $\left[N_A\right]_{i'}$ indicates the i-th row of matrix N_A, defined as

$$\left[N_A\right]_{ij} = \sum_{n=1}^{N} q(z^n = k) \sum_{t=2}^{T} \left\langle \left[h_{t-1}^n\right]_j \left[h_t^n\right]_i \right\rangle_{q(h_{t-1:t}^n | z^n=k)} + \alpha_{ij}\hat{A}_{ij}.$$

In addition, we define

$$\left[H_{iA}\right]_{jl} \equiv \sum_{n=1}^{N} q(z^n = k) \sum_{t=2}^{T} \left\langle \left[h_{t-1}^n\right]_j \left[h_{t-1}^n\right]_l \right\rangle_{q(h_{t-1}^n | z^n=k)} + \alpha_{ij}\delta_{jl}$$

$q\left(\Sigma_H^{-1}\right)$**:** For $\Sigma_H^{-1} = dg\left(\tau\right)$, where each diagonal element τ_i follows a Gamma prior $\mathcal{G}(a_1^i, a_2^i)$, the updates are (see Appendix B.2.4):

$$q(\tau_i) = \mathcal{G}\left(a_1^i + \frac{T-1}{2}\sum_{n=1}^{N} q(z^n = k), a_2^i + \frac{1}{2}\left(\sum_{n=1}^{N} q(z^n = k) \sum_{t=2}^{T} \left\langle \left[h_t^n\right]_i^2 \right\rangle_{q(h_t^n | z^n=k)} - \left[G_A\right]_i + \sum_j \alpha_{ij}\hat{A}_{ij}^2\right)\right),$$

where $\left[G_A\right]_i \equiv \left[N_A\right]_{i'} H_{iA}^{-1} \left[N_A\right]_{i'}^{\mathsf{T}}$.

Updates for $q\left(\mu^k, \left[\Sigma^k\right]^{-1}\right)$

The optimal $q\left(\mu, \Sigma^{-1}\right)$ is a Gaussian-Wishart(Gamma) distribution given by:

$$q\left(\mu, \Sigma^{-1}\right) \propto e^{-\frac{1}{2}\sum_{n=1}^{N} q(z^n=k) \left\langle (h_1^n - \mu)^{\mathsf{T}} \Sigma^{-1} (h_1^n - \mu)\right\rangle_{q(h_1 | z^n=k)}} p(\mu|\Sigma^{-1}) p(\Sigma^{-1}|\hat{\Theta})$$

$q\left(\mu|\Sigma^{-1}\right)$**:** In Appendix B.2.5 we show that $q\left(\mu|\Sigma^{-1}\right) = \mathcal{N}\left(\mu_h, \Sigma_h\right)$ where

$$\mu_h = \left(\sum_{n=1}^{N} q(z^n = k) I_H + \Sigma_\mu^{-1}\right)^{-1} \left(\underbrace{\sum_{n=1}^{N} q(z^n = k)\left\langle h_1^n \right\rangle + \Sigma_\mu^{-1}\mu_\mu}_{m}\right)$$

$$\Sigma_h = \left(\sum_{n=1}^{N} q(z^n = k) I_H + \Sigma_\mu^{-1}\right)^{-1} \Sigma.$$

[9] $vr\left(A\right)$ denotes the vector formed by stacking the rows of the matrix A. Unlike for B, the choice of a row vector formulation is more appropriate for simplifying mathematical notations. $bdg\left(x_1, \ldots, x_n\right)$ indicates the block diagonal matrix with blocks $x_1, \ldots, x_n$, while $vert\left(x_1, \ldots, x_n\right)$ stands for vertically concatenating the arguments $x_1, \ldots, x_n$.

$q\left(\Sigma^{-1}\right)$**:** For the case in which $p\left(\Sigma^{-1}\right) = \mathcal{W}\left(\nu, S\right)$ the updates are (see Appendix B.2.6):

$$\mathcal{W}\left(\nu + \sum_{n=1}^{N} q(z^n = k), S^{-1} - m\mu_h^{\mathsf{T}} + \sum_{n=1}^{N} q(z^n = k)\left\langle h_1^n (h_1^n)^{\mathsf{T}}\right\rangle + \Sigma_\mu^{-1}\mu_\mu\mu_\mu^{\mathsf{T}}\right)$$

If each element λ_i of Σ^{-1} follows a gamma distribution $p\left(\Sigma^{-1}\right) = \mathcal{G}\left(\sigma_1^i, \sigma_2^i\right)$, we have:

$$q(\lambda_i) = \mathcal{G}\left(\sigma_1^i + \frac{1}{2}\sum_{n=1}^{N} q(z^n = k), \sigma_2^i + \frac{1}{2}\left(-\left[m\mu_h^{\mathsf{T}}\right]_{ii} + \sum_{n=1}^{N} q(z^n = k)\left\langle \left[h_1^n\right]_i^2\right\rangle + \left[\Sigma_\mu^{-1}\mu_\mu\mu_\mu^{\mathsf{T}}\right]_{ii}\right)\right)$$

Updates for $q\left(z^n\right)$

The optimal $q\left(z^n = k\right)$ is given by:

$$q(z^n = k) \propto e^{H_q(h_{1:T}^n|z^n=k) + \langle \log p(z^n=k|z^{\neg n},\gamma)\rangle_{q(z^{\neg n})} + \left\langle \log p\left(v_{1:T}^n, h_{1:T}^n|\Theta^k\right)\right\rangle_{q\left(h_{1:T}^n|z^n=k\right)q(\Theta^k)}}$$

The first term in the exponent can be computed exactly as described in the Appendix B.7. The average $\langle \log p\left(z^n = k|z^{\neg n}, \gamma\right)\rangle_{(z^{\neg n})}$ needs attention since, naively, $p\left(z^n = k|z^{\neg n}, \gamma\right)$ possesses little structure to enable the average to be tractable. Whilst the naive exponential complexity can be reduced, we employ the second order Taylor expansion approximation of [23], as shown in detail in Appendix B.3.

Inference on $q\left(h_{1:T}^n|z^n = k\right)$

The optimal $q\left(h_{1:T}^n|z^n = k\right)$ is given by:

$$q(h_{1:T}^n|z^n = k) \propto e^{\left\langle \log p\left(v_{1:T}^n, h_{1:T}^n|\Theta^k\right)\right\rangle_{q(\Theta^k)}}. \tag{10}$$

This term is closely related to a standard VB approximation to a Bayesian LGSSM. Clearly the structure of $q(h_{1:T}^n|z^n = k)$ is a pairwise Markov chain, and inference algorithms such as Belief Propagation can be used [12, 17]. However, we take the approach discussed in [15], which reformulates the problem such that standard LGSSM inference routines can be applied. This both simplifies the development and can be advantageous in regimes of numerical instability. The central idea is to use the following decomposition:

$$\left\langle (v_t^n - W_t^n B h_t^n)^{\mathsf{T}}\Sigma_V^{-1}(v_t^n - W_t^n B h_t^n)\right\rangle_{q(B,\Sigma_V^{-1})} = \underbrace{(v_t^n - W_t^n\langle B\rangle h_t^n)^{\mathsf{T}}\left\langle\Sigma_V^{-1}\right\rangle(v_t^n - W_t^n\langle B\rangle h_t^n)}_{mean} + \underbrace{(h_t^n)^{\mathsf{T}}(S_B)_t^n h_t^n}_{fluctuation},$$

and similarly

$$\left\langle (h_t^n - A h_{t-1}^n)^{\mathsf{T}}\Sigma_H^{-1}(h_t^n - A h_{t-1}^n)\right\rangle_{q(A,\Sigma_H^{-1})} = \underbrace{(h_t^n - \langle A\rangle h_{t-1}^n)^{\mathsf{T}}\left\langle\Sigma_H^{-1}\right\rangle(h_t^n - \langle A\rangle h_{t-1}^n)}_{mean} + \underbrace{(h_{t-1}^n)^{\mathsf{T}} S_A h_{t-1}^n}_{fluctuation},$$

where

$$(S_B)_t^n \equiv \left\langle B^{\mathsf{T}} W_t^n \Sigma_V^{-1} W_t^n B\right\rangle - \langle B\rangle^{\mathsf{T}} W_t^n\left\langle\Sigma_V^{-1}\right\rangle W_t^n\langle B\rangle, \qquad S_A \equiv \left\langle A^{\mathsf{T}}\Sigma_H^{-1}A\right\rangle - \langle A\rangle^{\mathsf{T}}\left\langle\Sigma_H^{-1}\right\rangle\langle A\rangle.$$

The analytical expressions for these covariances are given in Appendix B.5. The mean terms represent the contribution of a standard LGSSM with parameters A, B, Σ_H^{-1} and Σ_V^{-1} replaced by their average values.

The key observation is to consider the extra 'fluctuation' terms as having been generated from fictitious zero-valued observations. This way, we can see Eq. (10) as the posterior of a standard LGSSM for which any of the standard algorithms in the literature [11] may be applied to perform inference.

More specifically we want to represent Eq. (10) directly as the posterior distribution $\tilde{q}(h_{1:T}^n|\tilde{v}_{1:T}^n)$ of an LGSSM by augmenting v_t^n and B as[10]:

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

[10]There are several ways of achieving a similar augmentation. We chose this since, in the non-Bayesian limit $U_A = (U_B)_t^n = \mathbf{0}_{HH}$, no numerical instabilities would be introduced ($\mathbf{0}_{HH}$ is a $H \times H$ matrix of zero elements).

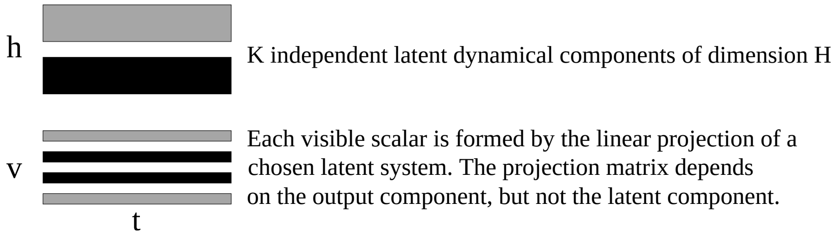


Figure 4: Time-series clustering based on simultaneous similarity.

where U_A is the Cholesky decomposition of S_A, so that $U_A^{\mathsf{T}} U_A = S_A$ (similarly, $(U_B)_t^n$ is the Cholesky decomposition of $(S_B)_t^n$). The equivalent LGSSM $\tilde{q}(h_{1:T}^n|\tilde{v}_{1:T}^n)$ is then completed by specifying[11]:

$$\tilde{A} \equiv \langle A \rangle\,, \quad \tilde{\Sigma}_H \equiv \left\langle \Sigma_H^{-1} \right\rangle^{-1}\,, \quad \tilde{\Sigma}_V \equiv bdg\left(\left\langle W_t^n \Sigma_V^{-1} \right\rangle^{-1}, I_H, I_H \right)\,, \quad \tilde{\mu} \equiv \langle \mu \rangle\,, \quad \tilde{\Sigma} \equiv \left\langle \Sigma^{-1} \right\rangle^{-1}\,.$$

The validity of this parameter assignment can be checked by showing that, up to negligible constants, the exponent of this augmented LGSSM has the same form as the exponent in Eq. (10). We can now apply any standard inference routine to compute $q(h_t^n|v_{1:T}^n) = \tilde{q}(h_t^n|\tilde{v}_{1:T}^n)$ [8, 9, 20][12]. In Appendix B.4 we describe the standard predictor-corrector form of the Kalman Filter, together with the Rauch-Tung-Striebel Smoother [9] and we show that a slight modification of the predictor-corrector algorithm produces a more efficient procedure obviating the need to consider fictitious outputs explicitly.

4.2 Clustering based on Simultaneous Similarity

In Section 4.1 we described a time-series clustering approach based on global similarity of the dynamics. We now introduce a different method, in which time-series are clustered on the basis of their simultaneous dynamical similarity. We will show that this can be achieved by a modification of the previously described framework.

As a motivating scenario, consider a situation in which at each time t, the stock prices of a set of V companies is known. Making a model of the stock prices of the companies over a long time is a considerable challenge. However, it is also suspected that certain stocks follow a similar underlying temporal profile, at least over a short time-scale. For example, the stocks of high-tech companies might be strongly dependent in the sense that their movements are correlated. Similarly, the stocks of oil companies might be more strongly related to each other than to companies outside the oil group. Our interest is therefore to cluster the companies into groups of 'correlated' activities. Whilst making a model for the long-time resulting vector time-series of the companies is very complex, grouping stocks together based on their simultaneous dynamical behavior over a short time-scale may be much simpler since we can compare at each time-point how the movements of the stocks are correlated with each other.

For clarity, consider only the task of clustering a set of V *unidimensional* time-series $v_{1:T}^{1:V}$; the extension to the multidimensional case is straightforward). Unlike the global clustering method, we are now interested in simultaneous dynamical similarity, for which we assume that all time-series have the same length T[13]. It is useful to compactly rewrite this set as one V-dimensional time-series $v_{1:T}$ with components $[v_{1:T}]_i \equiv [v_1]_i\,, \ldots, [v_T]_i\,, i = 1, \ldots, V$. We are thus interested in clustering the components i, that is to assign the i^{th} sequence $[v_{1:T}]_i$ into one of K clusters on the basis of its simultaneous dynamical similarity with other sequences in the same cluster. In order to do so, we consider K independent dynamical systems. Each unidimensional time-series is then a one-dimensional projection from one of the dynamical systems. The assignment of an output to a latent dynamical system is fixed throughout the time-series. See Fig. 4 for an informal sketch of this setup. More precisely, our model assumes that the emission parameters of the LGSSM $\Theta_\epsilon \equiv \{B, \Sigma_V\}$ do not depend on the cluster k, while the transition parameters $\Theta_\tau^k \equiv \{A^k, \Sigma_H^k, \mu^k, \Sigma^k\}$ depend on the cluster k. We may cluster components by using an indicator z^i for each component and define $p\left([v_t]_i\,|h_t^{1:K}, z^i = k, \Theta_\epsilon\right) \equiv p\left([v_t]_i\,|h_t^k, \Theta_\epsilon\right)$. That is, when z^i is

[11]At time T, $\tilde{B}_T^n \equiv vert\left(W_T^n \langle B \rangle, \mathbf{0}_{HH}, (U_B)_T^n\right)$. At time 1, $\tilde{B}_1^n \equiv vert\left(W_1^n \langle B \rangle, U_A, (U_B)_1^n, \mathbf{0}_H^{\mathsf{T}}\right)$, $\tilde{v}_1^n \equiv vert\left(v_1^n, \mathbf{0}_H, \mathbf{0}_H, \left(\left\langle \mu^{\mathsf{T}} \Sigma^{-1} \mu \right\rangle - \langle \mu \rangle^{\mathsf{T}} \left\langle \Sigma^{-1} \right\rangle \langle \mu \rangle\right)^{\frac{1}{2}}\right)$, $\tilde{\Sigma}_V \equiv bdg\left(\left\langle W_t^n \Sigma_V^{-1} \right\rangle^{-1}, I_H, I_H, 1\right)$.

[12]Note that, since the augmented LGSSM $\tilde{q}(h_{1:T}^n|\tilde{v}_{1:T}^n)$ is designed to match the *fully* clamped distribution $q(h_{1:T}^n|v_{1:T}^n)$, the filtered posterior $\tilde{q}(h_t^n|\tilde{v}_{1:t}^n)$ does not correspond to $q(h_t^n|v_{1:t}^n)$.

[13]For synchronized time-series off differing lengths, in principle one could treat this as a missing-data problem, along the lines previously described.

in state k, the i^{th} component is drawn from the dynamics of the k^{th} LGSSM. More precisely, the emission term is given by

$$p\left([v_t]_i \,|h_t^{1:K}, z^i = k, \Theta_\epsilon\right) \equiv \mathcal{N}\left(B_{i'} h_t^k, [\Sigma_V]_{ii}\right),$$

while the transition term is given by:

$$p\left(h_t^k | h_{t-1}^k, \Theta_\tau^k\right) \equiv \mathcal{N}\left(A^k h_{t-1}^k, \Sigma_H^k\right).$$

The emissions and transitions over time give:

$$p\left([v_{1:T}]_i \,|h_{1:T}^{1:K}, z^i = k, \Theta_\epsilon\right) = \prod_{t=1}^{T} p\left([v_t]_i \,|h_t^{1:K}, z^i = k, \Theta_\epsilon\right)$$
$$p\left(h_{1:T}^k | \Theta_\tau^k\right) = p\left(h_1^k | \Theta_\tau^k\right) \prod_{t=2}^{T} p\left(h_t^k | h_{t-1}^k, \Theta_\tau^k\right).$$

Each of the K linear dynamical systems proceeds independently of the rest, giving:

$$p\left(h_{1:T}^{1:K} | \Theta_\tau^{1:K}\right) = \prod_{k=1}^{K} p\left(h_{1:T}^k | \Theta_\tau^k\right).$$

The covariance Σ_V is constrained to be diagonal, so that

$$p\left(v_{1:T} | h_{1:T}^{1:K}, z^{1:V}, \Theta_\epsilon\right) = \prod_{i=1}^{V} p\left([v_{1:T}]_i \,|h_{1:T}^{1:K}, z^i, \Theta_\epsilon\right).$$

The graphical representation of this model is presented in Fig. 5. In the Bayesian version, we define the joint density

$$p\left(v_{1:T}, h_{1:T}^{1:K}, z^{1:V}, \Theta_\epsilon, \Theta_\tau^{1:K} | \hat{\Theta}_\epsilon, \hat{\Theta}_\tau^{1:K}, \gamma\right)$$
$$= \prod_t \left\{ \prod_i p\left([v_t]_i \,|h_t^{1:K}, z^i, \Theta_\epsilon\right) \prod_k p\left(h_t^k | h_{t-1}^k, \Theta_\tau^k\right) \right\} p\left(z^{1:V} | \gamma\right) p\left(\Theta_\epsilon | \hat{\Theta}_e\right) p\left(\Theta_\tau^{1:K} | \hat{\Theta}_t^{1:K}\right).$$

As in Section 4, we can automatically learn the number of clusters K by placing a Polya distribution on $p(z^{1:V}|\gamma)$ (Eq. (4)), where now N is replaced with V. For the LGSSM parameters, we define Gaussian-Gamma priors, similar to the one described in Section 2.

To form a tractable marginal log-likelihood bound, we use the variational approximation[14]:

$$p\left(h_{1:T}^{1:K}, z^{1:V}, \Theta_\epsilon, \Theta_\tau^{1:K} | v_{1:T}, \hat{\Theta}_\epsilon, \hat{\Theta}_\tau^{1:K}, \gamma\right) \approx q\left(h_{1:T}^{1:K}\right) q(z^{1:V}) q(\Theta_\epsilon, \Theta_\tau^{1:K}),$$

and further make the factorization assumption $q(z^{1:V}) \equiv \prod_i q(z^i)$. The independence assumptions on p and q imply $q(h_{1:T}^{1:K}) = \prod_k q(h_{1:T}^k)$ and $q(\Theta_\tau^{1:K}) = \prod_k q(\Theta_\tau^k)$. Then the lower bound on $\log p(v_{1:T}|\hat{\Theta}_\epsilon, \hat{\Theta}_\tau^{1:K}, \gamma)$ is given by:

$$\mathcal{F} \equiv \sum_k H_q(h_{1:T}^k) + \sum_i H_q(z^i) + H_q(\Theta_\epsilon) + \sum_k H_q(\Theta_\tau^k)$$
$$+ \sum_{i,t} \left\langle \log p\left([v_t]_i \,|h_t^{1:K}, z^i, \Theta_\epsilon\right)\right\rangle_{q(z^i)q(h_t^{1:K})q(\Theta_\epsilon)} + \sum_{t,k} \left\langle \log p\left(h_t^k | h_{t-1}^k, \Theta_\tau^k\right)\right\rangle_{q(h_{t-1:t}^k)q(\Theta_\tau^k)}$$
$$+ \left\langle \log p\left(z^{1:V}|\gamma\right)\right\rangle_{\prod_i q(z^i)} + \left\langle \log p\left(\Theta_\epsilon | \hat{\Theta}_\epsilon\right)\right\rangle_{q(\Theta_\epsilon)} + \sum_k \left\langle \log p\left(\Theta_\tau^k | \hat{\Theta}_\tau^k\right)\right\rangle_{q(\Theta_\tau^k)}.$$

[14]The factorization $q\left(h_{1:T}^{1:K} | z^{1:V}\right) \equiv q\left(h_{1:T}^{1:K}\right)$ is assumed in order to avoid the computational issue of having to consider all possible combinations of $z^{1:V}$.

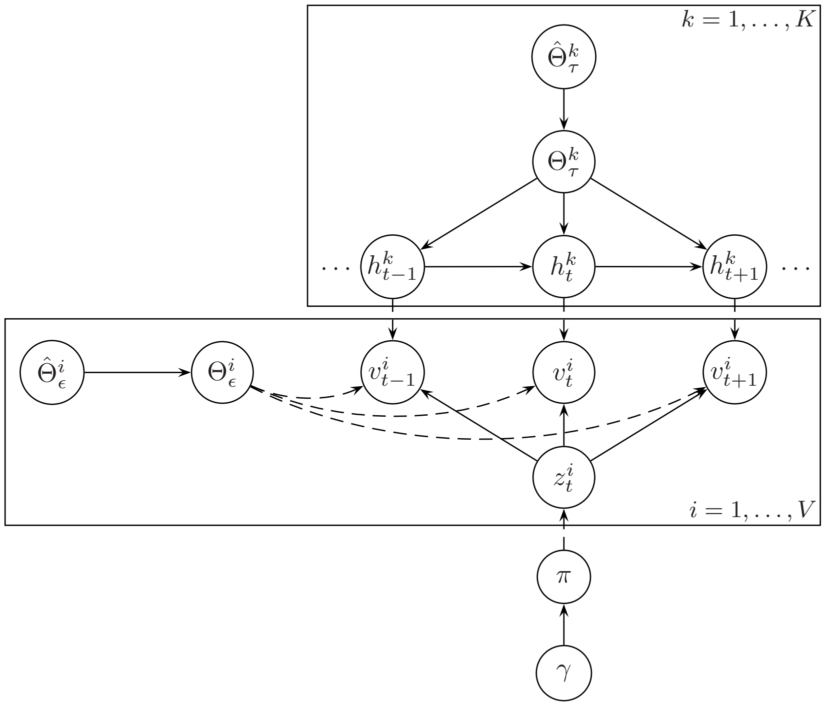


Figure 5: Graphical representation of the Dirichlet Mixture of Bayesian LGSSMs for performing clustering based on simultaneous similarity.

The resulting optimal q distributions are:

$$q\left(A^k, (\Sigma^k_H)^{-1}\right) \propto p\left(A^k, \left(\Sigma^k_H\right)^{-1} |\hat{\Theta}^k_\tau\right) e^{\sum_t \left\langle \log p\left(h^k_t | h^k_{t-1}, \Theta^k_\tau\right)\right\rangle_{q\left(h^k_{t-1:t}\right)}}$$

$$q\left(B, \Sigma_V^{-1}\right) \propto p\left(B, \Sigma_V^{-1} | \hat{\Theta}_e\right) e^{\sum_{i,t} \left\langle \log p\left([v_t]_i | h^{1:K}_t, z^i, \Theta_\epsilon\right)\right\rangle_{q(z^i) q\left(h^{1:K}_t\right)}}$$

$$q\left(h^k_{1:T}\right) \propto e^{\sum_{i,t} q(z^i = k) \left\langle \log p\left([v_t]_i | h^k_t, \Theta_\epsilon\right)\right\rangle_{q(\Theta_e)} + \sum_t \left\langle \log p\left(h^k_t | h^k_{t-1}, \Theta^k_\tau\right)\right\rangle_{q(\Theta^k_\tau)}}$$

$$q\left(z^i = k\right) \propto e^{\left\langle \log p\left(z^i | z^{\neg i}\right)\right\rangle_{\prod_{j \neq i} z^j} + \sum_t \left\langle \log p\left(v^i_t | h^k_t, \Theta_\epsilon\right)\right\rangle_{q\left(h^k_t\right) q(\Theta_\epsilon)}}$$

The specific updates can be derived similarly as for the method described in Section 4.1. We point out that for performing inference on $q(h^k_{1:T})$ a similar approach as the one described in Section 4.1 can be applied with the replacement $\langle \rho_i \rangle \leftarrow q(z^i = k) \langle \rho_i \rangle$, $S_B \leftarrow \sum_i q(z^i = k) H^{-1}_{iB}$ for the case in which there are not missing observations.

4.3 Relation to Previous Work

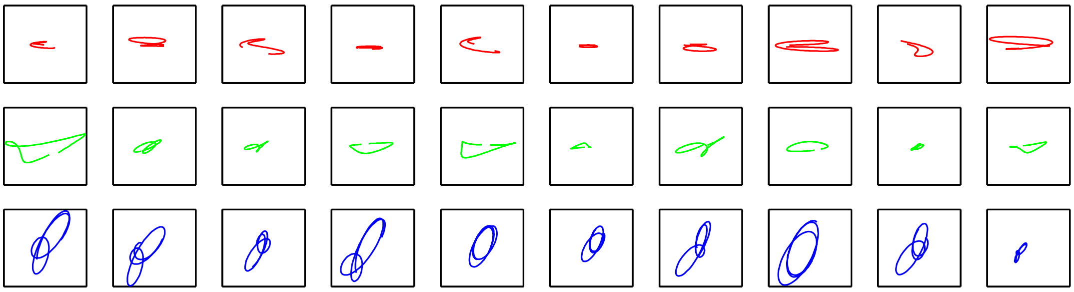
Older works based on temporal linear models include [24], which uses a mixture of ARMA models with the number of mixtures determined by the BIC criterion. A Bayesian approach based on Gibbs sampling has been used in a model similar but more constrained than the DMBLGSSM [7]. However, in this model the number of clusters was not determined automatically but using the BIC criterion, since no preference about a parsimonious parametrization was build into the model.

5 Demonstration

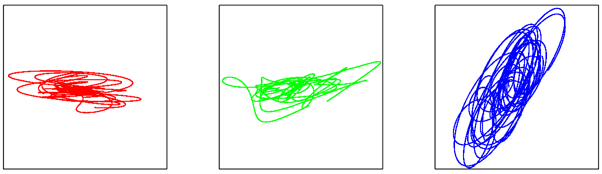
We performed several experiments to test the two models presented above on their clustering ability in 'difficult' situations. We will give some illustrative examples below.

5.1 Clustering based on Global Similarity

We tested the model on thirty synthetic sequences generated by three LGSSMs with $V = 2$, $H = 5$ and $T = 140$. The parameters were chosen so that all time-series had visually dissimilar dynamical trajectories, see Fig. 6a. Our



(a) Unclustered Trajectories



(b) Clustered Trajectories

Figure 6: (a) Thirty trajectories length $T = 140$ resulting from the dynamics of three different LGSSMs, all with hidden dimension $H = 5$. Plotted are the points $([v_t]_1, [v_t]_2)$. Different colors correspond to different underlying LGSSMs. (b) Our method correctly identifies three clusters, with all cluster labels consistent with the known generating mechanism. The trajectories belonging to the same cluster are plotted in the same subfigure.

model with initial $K = 5$ clusters and a $H = 10$ latent dimension perfectly clustered the data into three groups (see Fig. 6b). Thanks to the priors enforcing a low number of clusters, and simplicity of each cluster model, we consistently found the same clustering using different initial K and H, provided they are sufficiently large. This illustrates that the DMBLGSSM is capable of clustering time-series for which a common method based on feature extraction would be more difficult to apply, since it is not clear which type of features can be used to group the time-series.

5.1.1 Missing Observations

We generated fifty synthetic sequences from two LGSSMs with $V = 2$, $H = 5$ and $T = 30$. The dynamics A^k of the two models where set to differ by a small amount (the eigenvalues are close and have the same stability properties). The mixing matrices, the noise covariances and prior means where set to be independent on the LGSSM, that is $B^k = B$, $\Sigma^k = \Sigma$, $\Sigma^k_H = \Sigma_H$ and $\Sigma^k_V = \Sigma_V$. Therefore the two models differ slightly only in the deterministic part of the dynamics A^k. We removed randomly 10% of the data from each channel (for a total 20% of the available time values). In Fig. 7 we plot four samples, two from LGSSM1 (blue) and two from LGSSM2 (red). Notice that, as expected from the setup, it is not possible to visually identify a common structure for two samples from the same cluster or a structure specific to each cluster. We run the DMBLGSSM with initial number of mixture $M = 8$ and hidden dimensionality $H = 7$. The model could correctly learn the appropriate number of mixture components and assigned all samples to the correct cluster. However, when removing 15% of the data from each channel, the model incorrectly assigned all samples to one cluster only.

5.2 Clustering based on Simultaneous Similarity

As a simple illustration of the clustering method based on simultaneous similarity, in Fig. 8 we plot a set of $V = 6$ output sequences of length T=250 which were generated by projecting from two independent LGSSM of dimension $H = 6$ (different colors correspond to different underlying LGSSMs). We trained our model on this data, assuming $K = 4$ latent linear dynamical systems, each of dimension $H = 8$. Pleasingly, the method correctly discarded two of the unneeded clusters, and identified the first three outputs (from top to bottom) as belonging to cluster 1, and the bottom three as belonging to cluster 2, consistent with the way the data was generated.

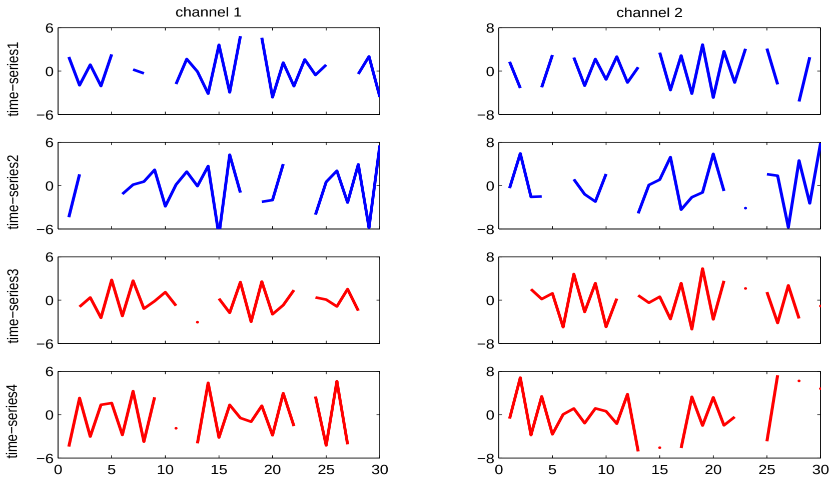


Figure 7: Four synthetic time-series generated from two LGSSMs with V=2, H=5 and T=30. Different colors correspond to different underlying LGSSMs. Plotted in the first and second columns are the first and second component of v_t respectively. No specific structure which identifies each cluster is visible.

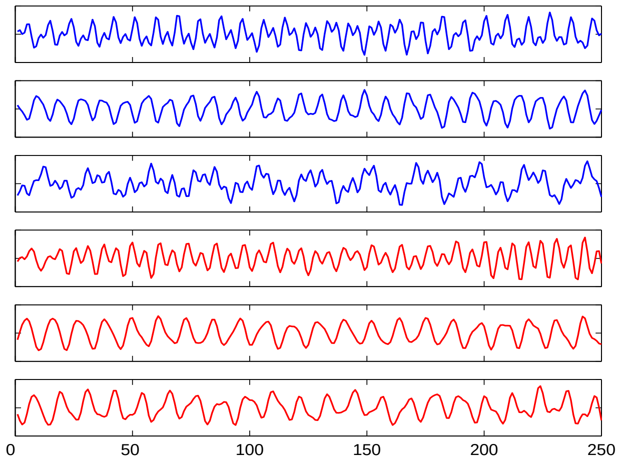


Figure 8: Clustering based on Simultaneous Similarity. Our model correctly identifies two clusters, assigning the top three output sequences to one cluster and the bottom three to another.

APPENDIX

A

A.1 Wishart and Gamma Distributions

Wishart Distribution

Let Σ be a $s \times s$ positive definite symmetric matrix of random variables and let S be a positive definite matrix of size $s \times s$. Then, Σ has a Wishart distribution $\mathcal{W}(\nu, S)$ if it has a probability density function given by:

$$p(\Sigma|\nu, S) = \frac{1}{Z}\left|\Sigma\right|^{(\nu-s-1)/2} e^{-\frac{1}{2}tr\left[S^{-1}\Sigma\right]},$$

where $Z = 2^{\nu s/2}\left|S\right|^{\nu/2}\pi^{s(s-1)/4}\prod_{i=1}^{s}\Gamma\left(\frac{\nu+1-i}{2}\right)$.

Gamma Distribution

A random variable σ has a Gamma distribution $\mathcal{G}(\nu_1, \nu_2)$ if it has a probability density function given by:

$$p(\sigma|\nu_1, \nu_2) = \frac{\nu_2^{\nu_1}}{\Gamma(\nu_1)} \sigma^{\nu_1 - 1} e^{-\nu_2 \sigma}.$$

A.2 Kullback-Leibler Divergence of Gaussian, Gamma and Wishart Distributions

The Kullback-Leibler divergence $\mathrm{KL}(q||p) = \langle \log q/p \rangle_q$ between two s-dimensional Gaussian distributions $q(x|\mu_q, \Sigma_q) = \mathcal{N}(\mu_q, \Sigma_q)$ and $p(x|\mu_p, \Sigma_p) = \mathcal{N}(\mu_p, \Sigma_p)$ is given by:

$$\mathrm{KL}(q||p) = \frac{1}{2}\left(\log\left(\frac{\det \Sigma_p}{\det \Sigma_q}\right) + tr\left[\Sigma_p^{-1}\Sigma_q\right] + (\mu_p - \mu_q)^\mathsf{T}\Sigma_p^{-1}(\mu_p - \mu_q) - s\right).$$

The KL divergence between two Gamma distributions $q(\sigma|q_1, q_2) = \mathcal{G}(q_1, q_2)$ and $p(\sigma|p_1, p_2) = \mathcal{G}(p_1, p_2)$ is given by:

$$\mathrm{KL}(q||p) = q_1 \log q_2 - p_1 \log p_2 - \log\frac{\Gamma(q_1)}{\Gamma(p_1)} + (q_1 - p_1)(\psi(q_1) - \log q_2) - q_1\left(1 - \frac{p_2}{q_2}\right),$$

where $\psi(\cdot)$ derivative of the gamma function logarithm. The KL divergence between two s-dimensional Wishart distributions $q(\Sigma|\nu_q, S_q) = \mathcal{W}(\nu_q, S_q)$ and $p(\Sigma|\nu_p, S_p) = \mathcal{W}(\nu_p, S_p)$:

$$\mathrm{KL}(q||p) = \log\frac{Z_{\nu_p S_p}}{Z_{\nu_q S_q}} + \frac{\nu_q - \nu_p}{2}\langle \ln|\Sigma| \rangle_q + \frac{1}{2}\nu_q tr\left[S_p^{-1}S_q - I_s\right],$$

where $\langle \ln|\Sigma| \rangle_q = \sum_{i=1}^{s} \psi(\frac{\nu_q + 1 - i}{2}) + s\log 2 + \log|S_q|$.

A.3 Dirichlet Distribution

The Dirichlet distribution of order $K \geq 2$ with parameters $\alpha_1, \ldots, \alpha_K$ has a probability density function given by:

$$\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K} x_k^{\alpha_k - 1}, \qquad \text{with} \sum_{k=1}^{K} x_k = 1.$$

The factor $\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}$ is the normalizing constant, that is $\frac{\prod_{k=1}^{K}\Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K}\alpha_k)} = \prod_{k=1}^{K}\int_0^1 x_k^{\alpha_k - 1} dx_k$. In order to see that, we first show that $\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$[15].

$$\begin{aligned}
\Gamma(\alpha)\Gamma(\beta) &= \int_0^\infty t^{\alpha-1}e^{-t}dt \int_0^\infty s^{\beta-1}e^{-s}ds \\
&= \int_0^\infty \int_0^\sigma \tau^{\alpha-1}(\sigma - \tau)^{\beta-1}e^{-\sigma}d\sigma d\tau \\
&= \int_0^\infty \int_0^1 y^{\alpha-1}x^{\alpha-1}y^{\beta-1}(1-x)^{\beta-1}e^{-y}ydydx \\
&= \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx\Gamma(\alpha+\beta),
\end{aligned}$$

[15]This is the normalizing constant of the Beta distribution.

with the change of variables $\sigma = t + s$, $\tau = t$, which implies $\sigma \in [0 \;\; \infty]$, $\tau \in [0 \;\; \sigma]$ ($s = \sigma - \tau > 0$) and $|J| = 1$; and the change of variables $x = \frac{\tau}{\sigma}$, $y = \sigma$, which gives $|J| = y$. Using this result, we find:

$$\begin{aligned}
\prod_{k=1}^{K} \Gamma(\alpha_k) &= \int_0^\infty t_1^{\alpha_1 - 1} e^{-t_1} dt_1 \cdots \int_0^\infty t_K^{\alpha_K - 1} e^{-t_K} dt_K \\
&= \int_0^\infty t_1^{\alpha_1 - 1} e^{-t_1} \cdots \int_0^\infty t_{K-2}^{\alpha_{K-2} - 1} e^{-t_{K-2}} \int_0^1 u_{K-1}^{\alpha_{K-1} - 1} (1 - u_{K-1})^{\alpha_K - 1} \int_0^\infty \sigma^{\alpha_{K-1} + \alpha_K - 1} e^{-\sigma} \\
&= \dots \\
&= \left[\prod_{k=1}^{K-1} \int_0^1 u_k^{\alpha_k - 1} (1 - u_k)^{\sum_{j=k+1}^{K} \alpha_j - 1} \right] \Gamma\left(\sum_{k=1}^{K} \alpha_k \right).
\end{aligned}$$

Therefore we have to show that $\prod_{k=1}^{K-1} \int_0^1 u_k^{\alpha_k - 1} (1 - u_k)^{\sum_{j=k+1}^{K} \alpha_j - 1} = \prod_{k=1}^{K} \int_0^1 x_k^{\alpha_k - 1} dx_k$. This can be done by induction. If we assume:

$$\prod_{k=1}^{K-2} \int_0^1 u_k^{\alpha_k - 1} (1 - u_k)^{\sum_{j=k+1}^{K-1} \alpha_j - 1} = \int_0^1 \cdots \int_0^1 \prod_{k=1}^{K-2} x_k^{\alpha_k - 1} \left(1 - \sum_{k=1}^{K-2} x_k \right)^{\alpha_{K-1} - 1} dx_{1:K-2},$$

then

$$\underbrace{\prod_{k=1}^{K-2} \int_0^1 u_k^{\alpha_k - 1} (1 - u_k)^{\sum_{j=k+1}^{K} \alpha_j - 1}}_{A} = \int_0^1 \cdots \int_0^1 \prod_{k=1}^{K-2} x_k^{\alpha_k - 1} \left(1 - \sum_{k=1}^{K-2} x_k \right)^{\alpha_{K-1} + \alpha_K - 1} dx_{1:K-2},$$

and

$$\begin{aligned}
& A \int_0^1 u_{K-1}^{\alpha_{K-1} - 1} (1 - u_{K-1})^{\alpha_K - 1} \\
&= \int_0^1 \cdots \int_0^1 \prod_{k=1}^{K-2} x_k^{\alpha_k - 1} \left(1 - \sum_{k=1}^{K-2} x_k \right)^{\alpha_{K-1} + \alpha_K - 1} dx_{1:K-2} \int_0^1 u_{K-1}^{\alpha_{K-1} - 1} (1 - u_{K-1})^{\alpha_K - 1} \\
&= \int_0^1 \cdots \int_0^1 \prod_{k=1}^{K-2} y_k^{\alpha_k - 1} \left(1 - \sum_{k=1}^{K-2} y_k \right)^{\alpha_{K-1}} \left(1 - \frac{y_K}{1 - \sum_{k=1}^{K-2} y_k} \right)^{\alpha_{K-1} - 1} y_K^{\alpha_K - 1} \frac{1}{1 - \sum_{k=1}^{K-2} y_k} dy_{1:K-2,K} \\
&= \int_0^1 \cdots \int_0^1 \prod_{k=1}^{K-2,K} y_k^{\alpha_k - 1} \left(1 - \sum_{k=1}^{K-2,K} y_k \right)^{\alpha_{K-1} - 1} dy_{1:K-2,K},
\end{aligned}$$

with the change of variables $y_K = (1 - \sum_{k=1}^{K-2} x_k)(1 - u_{K-1}), y_1 = x_1, \dots, y_{K-2} = x_{K-2}$, which implies $|J| = \frac{1}{1 - \sum_{k=1}^{K-2} y_k}$.

Polya Distribution

Consider a multinomial distribution $p(z^{1:N} | \pi) = \prod_{k=1}^{K} \pi_k^{N_k}$ and a symmetric Dirichlet distribution $p(\pi) = \frac{\Gamma(\gamma)}{\Gamma(\gamma/K)^K} \prod_{k=1}^{K} \pi_k^{\gamma/K - 1}$. From what we have seen above:

$$p\left(z^{1:N}\right) = \frac{\Gamma(\gamma)}{\Gamma(\gamma/K)^K} \int_\pi \prod_{k=1}^{K} \pi_k^{\gamma/K + N_k - 1} = \frac{\Gamma(\gamma)}{\Gamma(\gamma/K)^K} \frac{\prod_{k=1}^{K} \Gamma(N_k + \gamma/K)}{\Gamma(N + \gamma)},$$

and

$$
\begin{aligned}
p\left(z^1=1|z^{2:N}\right) &= \frac{p(z^1=1,z^{2:N})}{p(z^{2:N})}\\
&= \frac{\prod_{k=1}^{K}\Gamma(\gamma/K+N_k)}{\Gamma(\gamma+N)}\frac{\Gamma(\gamma+N-1)}{\Gamma(\gamma/K+N_1-1)\prod_{k=2}^{K}\Gamma(\gamma/K+N_k)}\\
&= \frac{\Gamma(\gamma/K+N_1)}{\Gamma(\gamma+N)}\frac{\Gamma(\gamma+N-1)}{\Gamma(\gamma/K+N_1-1)} = \frac{\gamma/K+N_1-1}{\gamma+N-1}.
\end{aligned}
$$

A.4 Kronecker Product and Vectorization

The following properties of Kronecker product and matrix vectorization hold:

$$
\begin{aligned}
&tr\left[A^{\mathsf{T}}B\right] = vc\,(A)^{\mathsf{T}}\,vc\,(B) = vr\,(A)^{\mathsf{T}}\,vr\,(B)\\
&(A\otimes B)^{\mathsf{T}} = A^{\mathsf{T}}\otimes B^{\mathsf{T}}\\
&(A\otimes B)^{-1} = A^{-1}\otimes B^{-1}\\
&vc\,(ABC) = (C^{\mathsf{T}}\otimes A)vc\,(B)\\
&vc\,(ABC)^{\mathsf{T}} = vc\,(B)^{\mathsf{T}}\,(C\otimes A^{\mathsf{T}})\\
&vr\,(ABC)^{\mathsf{T}} = vr\,(B)^{\mathsf{T}}\,(A^{\mathsf{T}}\otimes C)\\
&(A\otimes B)(C\otimes D) = AC\otimes BD
\end{aligned}
$$

B

In this Section we describe in details the model introduced in Section 4.1. Details for the model introduced in Section 4.2 are similar and thus omitted.

B.1 Independence Assumptions on the q Distribution

The independence assumptions made on the q distribution are the following: $q(h_{1:T}^{1:N}|z^{1:N},\Theta^{1:K}) \equiv q(h_{1:T}^{1:N}|z^{1:N})$, $q(z^{1:N},\Theta^{1:K}) \equiv q(z^{1:N})q(\Theta^{1:K})$ and $q(z^{1:N}) \equiv \prod_{n=1}^{N} q(z^n)$. The first two assumptions on q and the assumptions on p imply that the optimal q satisfies: $q(h_{1:T}^{1:N}|z^{1:N}) = \prod_{n=1}^{N} q\left(h_{1:T}^{n}|z^n\right)$ and $q(\Theta^{1:K}) = \prod_{k=1}^{K} q(\Theta^k)$. Indeed the lower bound is given by:

$$
\begin{aligned}
\mathcal{F} \equiv& H_q(\Theta^{1:K}) + \sum_{k^{1:N}} q\left(z^{1:N}=k^{1:N}\right) H_q\left(h_{1:T}^{1:N}|z^{1:N}=k^{1:N}\right) + H_q\left(z^{1:N}\right) + \left\langle \sum_{k=1}^{K}\log p\left(\Theta^k|\hat{\Theta}^k\right)\right\rangle_{q(\Theta^{1:K})}\\
&+ \left\langle \log p\left(z^{1:N}\right)\right\rangle_{q(z^{1:N})} + \sum_{k^{1:N}} q(z^{1:N}=k^{1:N}) \left\langle \sum_{n=1}^{N}\log p\left(v_{1:T}^n, h_{1:T}^n|\Theta^{k^n}\right)\right\rangle_{q(\Theta^{1:K})q\left(h_{1:T}^{1:N}|z^{1:N}=k^{1:N}\right)}.
\end{aligned}
$$

By maximizing $\mathcal{F}$ with respect to $q\left(h_{1:T}^{1:N}\right)$ and $q(\Theta^{1:K})$, we obtain:

$$
\begin{aligned}
q\left(h_{1:T}^{1:N}|z^{1:N}=k^{1:N}\right) &\propto e^{\left\langle \sum_{n=1}^{N}\log p\left(v_{1:T}^n,h_{1:T}^n|\Theta^{k^n}\right)\right\rangle_{q(\Theta^{1:K})}}\\
&= e^{\sum_{n=1}^{N}\left\langle \log p\left(v_{1:T}^n,h_{1:T}^n|\Theta^{k^n}\right)\right\rangle_{q(\Theta^{k^n})}}\\
&= \prod_{n=1}^{N} q\left(h_{1:T}^n|z^n=k^n\right),
\end{aligned}
$$

and

$$
\begin{aligned}
q(\Theta^{1:K}) &\propto \prod_{k=1}^{K} p\left(\Theta^k|\hat{\Theta}^k\right) e^{\left\langle \sum_{n=1}^{N} \log p\left(v_{1:T}^n, h_{1:T}^n|\Theta^{k^n}\right)\right\rangle_{q\left(h_{1:T}^{1:N}, z^{1:N}=k^{1:N}\right)}} \\
&= \prod_{k=1}^{K} p\left(\Theta^k|\hat{\Theta}^k\right) e^{\sum_{n=1}^{N} q(z^n=k)\left\langle \log p\left(v_{1:T}^n, h_{1:T}^n|\Theta^k\right)\right\rangle_{q\left(h_{1:T}^n|z^n=k\right)}} \\
&= \prod_{k=1}^{K} q(\Theta^k).
\end{aligned}
$$

Notice that, under these assumptions, the optimal $q(z^{1:N})$ would be:

$$
q(z^{1:N}=k^{1:N}) \propto p\left(z^{1:N}=k^{1:N}\right) e^{H_q\left(h_{1:T}^{1:N}|z^{1:N}=k^{1:N}\right)+\left\langle \sum_{n=1}^{N} \log p\left(v_{1:T}^n, h_{1:T}^n|\Theta^{k^n}\right)\right\rangle_{q(\Theta^{1:K})q\left(h_{1:T}^{1:N}|z^{1:N}=k^{1:N}\right)}},
$$

which does not factorize because $p\left(z^{1:N}\right)$ does not factorize.

B.2 Parameter Updates

To simplify the notation, it is useful to use a column vectorization for B, $vc\,(B)$, and a row vectorization for A, $vr\,(A)$ for the respective distributions. For the same reason, we will omit the dependency of the model parameter Θ^k and hyperparameter $\hat{\Theta}^k$ on the mixture k.

Updates for $q\left(B^k, \left[\Sigma_V^k\right]^{-1}\right)$

The optimal $q\left(B, \Sigma_V^{-1}\right)$ is a Gaussian-Wishart(Gamma) distribution given by:

$$
\frac{1}{|2\pi\Sigma_V|^{\frac{T}{2}\sum_{n=1}^{N} q(z^n=k)}} e^{-\frac{1}{2}\sum_{n=1}^{N} q(z^n=k)\sum_{t=1}^{T}\left\langle (v_t^n - W_t^n B h_t^n)^\mathsf{T}\Sigma_V^{-1}(v_t^n - W_t^n B h_t^n)\right\rangle_{q(h_t|z^n=k)}} p(B|\beta, \Sigma_V^{-1})p(\Sigma_V^{-1}|\hat{\Theta}) \tag{11}
$$

The exponent (excluding $p(\Sigma_V^{-1}|\hat{\Theta})$) is given by $-\frac{1}{2}\mathcal{E}$, where:

$$
\begin{aligned}
\mathcal{E} = \sum_{n=1}^{N} q(z^n=k)\left(\sum_{t=1}^{T}(v_t^n)^\mathsf{T}\Sigma_V^{-1}v_t^n - 2\sum_{t=1}^{T}\left\langle h_t^n\right\rangle^\mathsf{T} B^\mathsf{T} W_t^n \Sigma_V^{-1} v_t^n + \sum_{t=1}^{T}\left\langle (h_t^n)^\mathsf{T} B^\mathsf{T} W_t^n \Sigma_V^{-1} W_t^n B h_t^n\right\rangle\right) \\
+ \sum_j \left(\beta_j B_j^\mathsf{T}\Sigma_V^{-1}B_j - 2\hat{B}_j^\mathsf{T}\Sigma_V^{-1}B_j + \hat{B}_j^\mathsf{T}\Sigma_V^{-1}\hat{B}_j\right).
\end{aligned}
$$

B.2.1 Determining $q\left(B^k|\left[\Sigma_V^k\right]^{-1}\right)$

Optimally, $q\left(B|\Sigma_V^{-1}\right)$ is a Gaussian. If we assume that Σ_V is diagonal for the case in which there are missing observations, then $W_t^n\Sigma_V^{-1}W_t^n = W_t^n\Sigma_V^{-1}$ and, using the properties described in Appendix A.4, we can write the quadratic term in B of $\mathcal{E}$ as

$$
\begin{aligned}
&\sum_{n=1}^{N} q(z^n=k)\sum_{t=1}^{T}\left\langle (h_t^n)^\mathsf{T} B^\mathsf{T} W_t^n \Sigma_V^{-1} B h_t^n\right\rangle + \sum_j \beta_j B_j^\mathsf{T}\Sigma_V^{-1}B_j \\
&= tr\left[\sum_{n=1}^{N} q(z^n=k)\sum_{t=1}^{T}\left\langle h_t^n (h_t^n)^\mathsf{T}\right\rangle B^\mathsf{T} W_t^n \Sigma_V^{-1} B + dg\,(\beta)\, B^\mathsf{T}\Sigma_V^{-1}B\right] \\
&= vc\,(B)^\mathsf{T}\left(\underbrace{\sum_{n=1}^{N} q(z^n=k)\sum_{t=1}^{T}\left\langle h_t^n (h_t^n)^\mathsf{T}\right\rangle \otimes W_t^n + dg\,(\beta)\otimes I_V}_{H_{BM}}\right)\left(I_H \otimes \Sigma_V^{-1}\right) vc\,(B),
\end{aligned}
$$

that is the covariance of $q(vc\,(B)\,|\Sigma_V^{-1})$ is given by:

$$\Sigma_B = (I_H \otimes \Sigma_V)\, H_{BM}^{-1}.$$

The linear term in B of $\mathcal{E}$ is given by:

$$\begin{aligned}
&- 2tr\left[\sum_{n=1}^{N} q(z^n = k)\sum_{t=1}^{T}\langle h_t^n\rangle\,(v_t^n)^{\mathsf{T}}\Sigma_V^{-1}W_t^n B + dg\,(\beta)\,\hat{B}^{\mathsf{T}}\Sigma_V^{-1}B\right] \\
&= -2vc\left(\Sigma_V^{-1}\left(\underbrace{W_t^n\sum_{n=1}^{N} q(z^n = k)\sum_{t=1}^{T} v_t^n\,\langle h_t^n\rangle^{\mathsf{T}} + \hat{B}dg\,(\beta)}_{N_B}\right)\right)^{\mathsf{T}} vc\,(B)\,,
\end{aligned}$$

that is the mean of $q(vc\,(B)\,|\Sigma_V^{-1})$ is given by:

$$\mu_B = \Sigma_B vc\left(\Sigma_V^{-1}N_B\right) = H_{BM}^{-1}vc\,(N_B)\,.$$

In the case in which there are not missing observations ($W_t^n = I_V$), the formula for the covariance reduces to:

$$\Sigma_B = \left(\underbrace{\sum_{n=1}^{N} q(z^n = k)\sum_{t=1}^{T}\left\langle h_t^n (h_t^n)^{\mathsf{T}}\right\rangle + dg\,(\beta)}_{H_B}\right)^{-1}\otimes\Sigma_V,$$

and the mean becomes:

$$\mu_B = vc\left(N_B H_B^{-1}\right).$$

B.2.2 Determining $q\left(\left[\Sigma_V^k\right]^{-1}\right)$

In Section B.2.1, we have shown that $q\left(B|\Sigma_V^{-1}\right)$ is Gaussian with exponent:

$$-\frac{1}{2}\left(vc\,(B) - \mu_B\right)^{\mathsf{T}}\Sigma_B^{-1}\left(vc\,(B) - \mu_B\right). \tag{12}$$

The part of Eq. (12) which is not explicitly present in Eq. (11) is given by:

$$-\frac{1}{2}\mu_B^{\mathsf{T}}\Sigma_B^{-1}\mu_B = -\frac{1}{2}tr\left[M_B N_B^{\mathsf{T}}\Sigma_V^{-1}\right].$$

where $M_B = N_B H_B^{-1}$ for the case in which there are not missing observations, while M_B is the $V\times H$ matrix formed by the vector $H_{BM}^{-1}vc\,(N_B)$ for the case of missing observations. The negative of this term, together with the part in the exponent of Eq. (11) which contains a dependency Σ_V, gives as exponent for $q(\Sigma_V^{-1})$ for the case in which Σ_V^{-1} follows a Wishart distribution $\mathcal{W}\,(\nu_V, S_V)$ (which we permit when there are no missing observations):

$$\frac{1}{2}tr\left[M_B N_B^{\mathsf{T}}\Sigma_V^{-1}\right] - \frac{1}{2}\sum_{n=1}^{N} q(z^n = k)\sum_{t=1}^{T}(v_t^n)^{\mathsf{T}}\Sigma_V^{-1}v_t^n - \frac{1}{2}B_j^{\mathsf{T}}\Sigma_V^{-1}B_j - \frac{1}{2}\mathrm{tr}\left[S_V^{-1}\Sigma_V^{-1}\right].$$

The terms that contain dependency on $|\Sigma_V|$ are given by:

$$\frac{|\Sigma_B|^{1/2}|dg\,(\beta)\,|^{V/2}\left|\Sigma_V^{-1}\right|^{(\nu_V - V - 1)/2}}{|\Sigma_V|^{\frac{T}{2}\sum_{n=1}^{N} q(z^n=k)}|\Sigma_V|^{H/2}} = \frac{|H_B^{-1}|^{V/2}|dg\,(\beta)\,|^{V/2}\left|\Sigma_V^{-1}\right|^{(\nu_V - V - 1)/2}}{|\Sigma_V|^{\frac{T}{2}\sum_{n=1}^{N} q(z^n=k)}}.$$

That means that if Σ_V^{-1} follows a Wishart distribution $\mathcal{W}\,(\nu_V, S_V)$ the updates are:

$$q(\Sigma_V^{-1}) = \mathcal{W}\left(\nu_V + T\sum_{n=1}^{N} q(z^n = k), \left(S_V^{-1} + \sum_{n=1}^{N} q(z^n = k)\sum_{t=1}^{T} v_t^n (v_t^n)^{\mathsf{T}} - G_B + \hat{B}dg\,(\beta)\,\hat{B}^{\mathsf{T}}\right)^{-1}\right),$$

where $G_B \equiv N_B H_B^{-1} N_B^{\mathsf{T}}$. Instead, for the constraint $\Sigma_V^{-1} = dg(\rho)$, where each diagonal element ρ_i follows a Gamma prior $\mathcal{G}(b_1^i, b_2^i)$, the optimal updates are:

$$q(\rho_i) = \mathcal{G}\left(b_1^i + \frac{T}{2}\sum_{n=1}^{N} q(z^n = k), b_2^i + \frac{1}{2}\left(\sum_{n=1}^{N} q(z^n = k)\sum_{t=1}^{T}[v_t^n]_i^2 - [G_B]_{ii} + \sum_j \beta_j \hat{B}_{ij}^2\right)\right).$$

Updates for $q\left(A^k, \left[\Sigma_H^k\right]^{-1}\right)$

The optimal $q\left(A, \Sigma_H^{-1}\right)$ is a Gaussian-Gamma distribution given by:

$$\frac{1}{|2\pi\Sigma_H|^{\frac{T-1}{2}\sum_{n=1}^{N} q(z^n=k)}} e^{-\frac{1}{2}\sum_{n=1}^{N} q(z^n=k)\sum_{t=2}^{T}\left\langle (h_t^n - A h_{t-1}^n)^{\mathsf{T}}\Sigma_H^{-1}(h_t^n - A h_{t-1}^n)\right\rangle_{q\left(h_{t-1:t}^n | z^n=k\right)}} p(A|\alpha, \Sigma_H^{-1}) p(\Sigma_H^{-1}|\hat{\Theta}) \tag{13}$$

B.2.3 Determining $q\left(A|\Sigma_H\right)$

Optimally, $q\left(A|\Sigma_H^{-1}\right)$ is a Gaussian. In order to obtain independence of the mean and other quantities from Σ_H^{-1}, here we have to assume that Σ_H^{-1} is diagonal (this will not be the case for other choices of the prior $p(A|\alpha, \Sigma_H^{-1})$).

The quadratic term in A of the exponent of Eq. (13) is given by:

$$\begin{aligned}
&\sum_{n=1}^{N} q(z^n = k)\sum_{t=2}^{T}\left\langle (h_{t-1}^n)^{\mathsf{T}} A^{\mathsf{T}}\Sigma_H^{-1} A h_{t-1}^n\right\rangle + \sum_{ij}\alpha_{ij} A_{ij}^{\mathsf{T}}\left[\Sigma_H^{-1}\right]_{ii} A_{ij} \\
&= tr\left[\sum_{n=1}^{N} q(z^n = k)\sum_{t=2}^{T}\left\langle h_{t-1}^n (h_{t-1}^n)^{\mathsf{T}}\right\rangle A^{\mathsf{T}}\Sigma_H^{-1} A\right] + vr\left(A\right)^{\mathsf{T}} \underbrace{bdg\left(\left[\Sigma_H^{-1}\right]_{11} dg\left(\alpha_{1'}\right), \ldots, \left[\Sigma_H^{-1}\right]_{HH} dg\left(\alpha_{H'}\right)\right)}_{D_A} vr\left(A\right) \\
&= vr\left(A\right)^{\mathsf{T}}\left(\left[\Sigma_H^{-1}\otimes\sum_{n=1}^{N} q(z^n = k)\sum_{t=2}^{T}\left\langle h_{t-1}^n (h_{t-1}^n)^{\mathsf{T}}\right\rangle\right] + D_A\right) vr\left(A\right) \\
&= vr\left(A\right)^{\mathsf{T}} bdg\left(\left[\Sigma_H^{-1}\right]_{11} H_{1A}, \ldots, \left[\Sigma_H^{-1}\right]_{HH} H_{HA}\right) vr\left(A\right),
\end{aligned}$$

where $[H_{iA}]_{jl} \equiv \sum_{n=1}^{N} q(z^n = k)\sum_{t=2}^{T}\left\langle \left[h_{t-1}^n\right]_j \left[h_{t-1}^n\right]_l\right\rangle_{q(h_{t-1}^n)} + \alpha_{ij}\delta_{jl}$.

This means that the covariance of $q(vr\left(A\right)|\Sigma_H^{-1})$ is given by

$$\Sigma_A = bdg\left([\Sigma_H]_{11} H_{1A}^{-1}, \ldots, [\Sigma_H]_{HH} H_{HA}^{-1}\right).$$

The linear term is given by:

$$\begin{aligned}
&\sum_{n=1}^{N} q(z^n = k)\sum_{t=2}^{T}\left\langle (h_t^n)^{\mathsf{T}}\Sigma_H^{-1} A h_{t-1}^n\right\rangle + \sum_{ij}\alpha_{ij}\hat{A}_{ij}^{\mathsf{T}}\left[\Sigma_H^{-1}\right]_{ii} A_{ij} \\
&= tr\left[\sum_{n=1}^{N} q(z^n = k)\sum_{t=2}^{T}\left\langle h_{t-1}^n (h_t^n)^{\mathsf{T}}\right\rangle \Sigma_H^{-1} A\right] + vr\left(\hat{A}\right)^{\mathsf{T}} D_A vr\left(A\right) \\
&= vr\left(\Sigma_H^{-1}\sum_{n=1}^{N} q(z^n = k)\sum_{t=2}^{T}\left\langle h_t^n (h_{t-1}^n)^{\mathsf{T}}\right\rangle\right)^{\mathsf{T}} vr\left(A\right) + vr\left(\hat{A}\right)^{\mathsf{T}} D_A vr\left(A\right),
\end{aligned}$$

that is the mean of $q(vr\left(A\right)|\Sigma_H^{-1})$ is given by:

$$\begin{aligned}
\mu_A &= \Sigma_A\left(vr\left(\Sigma_H^{-1}\sum_{n=1}^{N} q(z^n = k)\sum_{t=2}^{T}\left\langle h_t^n (h_{t-1}^n)^{\mathsf{T}}\right\rangle\right) + D_A vr\left(\hat{A}\right)\right) \\
&= vert\left(\left([N_A]_{1'} H_{1A}^{-1}\right)^{\mathsf{T}}, \ldots, \left([N_A]_{H'} H_{HA}^{-1}\right)^{\mathsf{T}}\right).
\end{aligned}$$

where $[N_A]_{ij} = \sum_{n=1}^{N} q(z^n = k)\sum_{t=2}^{T}\left\langle \left[h_{t-1}^n\right]_j \left[h_t^n\right]_i\right\rangle_{q(h_{t-1:t}^n)} + \alpha_{ij}\hat{A}_{ij}$

B.2.4 Determining $q\left(\left[\Sigma_H^k\right]^{-1}\right)$

The missing part of $(vr\,(A)-\mu_A)\Sigma_A^{-1}(vr\,(A)-\mu_A)$ in the exponent of $q(A,\Sigma_H^{-1})$ (Eq. (13)) is given by:

$$\sum_i \left[\Sigma_H^{-1}\right]_{ii} [N_A]_{i'} \, H_{iA}^{-1} \, [N_A]_{i'}^{\mathsf{T}} .$$

For $\Sigma_H^{-1} = dg\,(\tau)$, where each element τ_i follows a Gamma prior $\mathcal{G}(a_1^i, a_2^i)$, the updates are:

$$q(\tau_i) = \mathcal{G}\left(a_1^i + \frac{T-1}{2}\sum_{n=1}^N q(z^n=k), a_2^i + \frac{1}{2}\left(\sum_{n=1}^N q(z^n=k)\sum_{t=2}^T \left\langle [h_t^n]_i^2 \right\rangle - [G_A]_i + \sum_j \alpha_{ij}\hat{A}_{ij}^2\right)\right),$$

where $[G_A]_i \equiv [N_A]_{i'} \, H_{iA}^{-1} \, [N_A]_{i'}^{\mathsf{T}}$.

Updates for $q\left(\mu^k, \left[\Sigma^k\right]^{-1}\right)$

The optimal $q\left(\mu, \Sigma^{-1}\right)$ is a Gaussian-Wishart(Gamma) distribution given by:

$$\frac{1}{|2\pi\Sigma|^{\frac{1}{2}\sum_{n=1}^N q(z^n=k)}} e^{-\frac{1}{2}\sum_{n=1}^N q(z^n=k)\left\langle (h_1^n-\mu)^{\mathsf{T}}\Sigma^{-1}(h_1^n-\mu)\right\rangle_{q(h_1|z^n=k)}} p(\mu|\Sigma^{-1})p(\Sigma^{-1}|\hat{\Theta})$$

The exponent (excluding $p(\Sigma^{-1})|\hat{\Theta})$ is given by $-\frac{1}{2}\mathcal{E}$, where:

$$\mathcal{E} = \sum_{n=1}^N q(z^n=k)\left(\left\langle (h_1^n)^{\mathsf{T}}\Sigma^{-1}h_1^n\right\rangle - 2\mu^{\mathsf{T}}\Sigma^{-1}\left\langle h_1^n\right\rangle + \mu^{\mathsf{T}}\Sigma^{-1}\mu\right) + (\mu-\mu_\mu)^{\mathsf{T}}\,\Sigma^{-1}\Sigma_\mu^{-1}\,(\mu-\mu_\mu)\,.$$

B.2.5 Determining $q\left(\mu^k | \left[\Sigma^k\right]^{-1}\right)$

Optimally, $q\left(\mu|\Sigma^{-1}\right)$ is a Gaussian. We can write the quadratic term in μ of $\mathcal{E}$ as $\sum_{n=1}^N q(z^n=k)\mu^{\mathsf{T}}\Sigma^{-1}\mu + \mu^{\mathsf{T}}\Sigma^{-1}\Sigma_\mu^{-1}\mu$, that is the covariance of $q(\mu|\Sigma^{-1})$ is given by:

$$\Sigma_h = \left(\sum_{n=1}^N q(z^n=k)I_H + \Sigma_\mu^{-1}\right)^{-1}\Sigma.$$

The linear term in μ of $\mathcal{E}$ is given by $-2\mu^{\mathsf{T}}\Sigma^{-1}\left(\underbrace{\sum_{n=1}^N q(z^n=k)\left\langle h_1^n\right\rangle + \Sigma_\mu^{-1}\mu_\mu}_{m}\right)$, that is the mean of $q(\mu|\Sigma^{-1})$ is given by:

$$\mu_h = \left(\sum_{n=1}^N q(z^n=k)I_H + \Sigma_\mu^{-1}\right)^{-1} m$$

B.2.6 Determining $q\left(\left[\Sigma^k\right]^{-1}\right)$

For the case in which $p\left(\Sigma^{-1}\right) = \mathcal{W}\,(\nu, S)$, the exponent of $q(\Sigma^{-1})$ is given by:

$$\frac{1}{2}\mu_h^{\mathsf{T}}\Sigma^{-1}m - \frac{1}{2}\sum_{n=1}^N q(z^n=k)\left\langle (h_1^n)^{\mathsf{T}}\Sigma^{-1}h_1^n\right\rangle - \frac{1}{2}\mu_\mu^{\mathsf{T}}\Sigma^{-1}\Sigma_\mu^{-1}\mu_\mu - \frac{1}{2}tr\left[S^{-1}\Sigma^{-1}\right]$$
$$= \frac{1}{2}tr\left[-\left(m\mu_h^{\mathsf{T}} + \sum_{n=1}^N q(z^n=k)\left\langle h_1^n(h_1^n)^{\mathsf{T}}\right\rangle + \Sigma_\mu^{-1}\mu_\mu\mu_\mu^{\mathsf{T}} + S^{-1}\right)\Sigma^{-1}\right]$$

The terms that containing $|\Sigma|$ are given by $\left|\Sigma^{-1}\right|^{(\nu-H-1+\sum_{n=1}^{N} q(z^n=k))/2}$. That means that the updates are:

$$\mathcal{W}\left(\nu+\sum_{n=1}^{N} q(z^n=k), S^{-1}-m\mu_h^{\mathsf{T}}+\sum_{n=1}^{N} q(z^n=k)\left\langle h_1^n (h_1^n)^{\mathsf{T}}\right\rangle+\Sigma_\mu^{-1}\mu_\mu\mu_\mu^{\mathsf{T}}\right)$$

In the case in which each element λ_i of Σ^{-1} follows a gamma distribution $p\left(\Sigma^{-1}\right)=\mathcal{G}\left(\sigma_1^i,\sigma_2^i\right)$, we have:

$$q(\lambda_i)=\mathcal{G}\left(\sigma_1^i+\frac{1}{2}\sum_{n=1}^{N} q(z^n=k), \sigma_2^i+\frac{1}{2}\left(-\left[m\mu_h^{\mathsf{T}}\right]_{ii}+\sum_{n=1}^{N} q(z^n=k)\left\langle\left[h_1^n\right]_i^2\right\rangle+\left[\Sigma_\mu^{-1}\mu_\mu\mu_\mu^{\mathsf{T}}\right]_{ii}\right)\right)$$

B.3 Updates for $q(z)$

The average $\langle\log p(z^n=k|z^{\neg n},\gamma)\rangle_{\prod_{m\neq n} q(z^m)}$ is approximated using a second order Taylor expansion. More specifically:

$$p(z^n=k|z^{\neg n},\gamma)=\frac{N_{k,\neg n}+\gamma/K}{N-1+\gamma}\equiv f(N_{k,\neg n}), \tag{14}$$

where $N_{k,\neg n}\equiv N_k-I[z^n=k]$ is the number of times z is in state k, excluding z^n. The quantities $N_{k,\neg n}$ are sums of Bernoulli variables and may be approximated with a Gaussian with mean and variance given by:

$$M_{k,\neg n}\equiv\sum_{m=1,m\neq n}^{N} q(z^m=k), \qquad S_{k,\neg n}\equiv\sum_{m=1,m\neq n}^{N} q(z^m=k)(1-q(z^m=k)).$$

We then approximate $\langle f(N_{k,\neg n})\rangle$ in Eq. (14) by using a second order Taylor expansion[16]:

$$\langle f(N_{k,\neg n})\rangle=f(M_{k,\neg n})+\frac{1}{2}f''(M_{k,\neg n})S_{k,\neg n}.$$

B.4 Inference on $q(h_{1:T})$

In this section we describe a standard algorithm the standard predictor-corrector form of the Kalman Filter, together with the Rauch-Tung-Striebel Smoother from the LGSSM literature, which can be used for performing inference on terms such as:

$$q(h_{1:T})\propto e^{\langle\log p(v_{1:T},h_{1:T}|\Theta)\rangle_{q(\Theta)}}.$$

This requires defining a new set of $\tilde{A},\tilde{B},\tilde{\Sigma}_H,\tilde{\Sigma}_V,\tilde{\mu},\tilde{\Sigma}$ parameters of the type described in Section 4.1. We also give a slight modification of the predictor-corrector algorithm which obviates the need to introduce fictitious outputs.

Algorithm 1 describe the standard predictor-corrector form of the Kalman Filter, together with the Rauch-Tung-Striebel Smoother [9] for computing $q(h_t|v_{1:T})=\tilde{q}(h_t|\tilde{v}_{1:T})$.

There are two variants of the FORWARD pass. Either we may call procedure FORWARD in Algorithm 1 with parameters $\tilde{A},\tilde{B},\tilde{\Sigma}_H,\tilde{\Sigma}_V,\tilde{\mu},\tilde{\Sigma}$ and the augmented visible variables $\tilde{v}_t$ in which we use steps 1a, 2a, 5a and 6a. This is exactly the predictor-corrector form of a Kalman Filter [9]. Otherwise, in order to reduce the computational cost, we may call procedure FORWARD with the parameters $\tilde{A},\langle B\rangle,\tilde{\Sigma}_H,\left\langle\Sigma_V^{-1}\right\rangle^{-1},\tilde{\mu},\tilde{\Sigma}$ and the original visible variable v_t[17] in which we use steps 1b (where $U_{AB}^{\mathsf{T}}U_{AB}\equiv S_A+S_B$[18]), 2b, 5b and 6b. The two algorithms are mathematically equivalent, as shown below. Computing $q(h_t|v_{1:T})=\tilde{q}(h_t|\tilde{v}_{1:T})$ is then completed by calling the common BACKWARD pass.

The important point here is that the reader may supply any standard Kalman Filtering/Smoothing routine, and simply call it with the appropriate parameters. In some parameter regimes, or in very long time-series, numerical stability may be a serious concern, for which several stabilized algorithms have been developed over the years, for example the square-root forms [9, 19, 20].

[16] The potentially more accurate procedure of using Quadrature fails in this case, since the arguments under Gaussian Quadrature take the function out of defined regions.

[17] At time 1, we need to use $\tilde{B}\equiv vert\left(\langle B\rangle,\mathbf{0}_H^{\mathsf{T}}\right)$, $\tilde{v}_1^n\equiv vert\left(v_1^n,\left(\left\langle\mu^{\mathsf{T}}\Sigma^{-1}\mu\right\rangle-\langle\mu\rangle^{\mathsf{T}}\left\langle\Sigma^{-1}\right\rangle\langle\mu\rangle\right)^{\frac{1}{2}}\right)$, $\tilde{\Sigma}_V\equiv bdg\left(\left\langle\Sigma_V^{-1}\right\rangle^{-1},1\right)$.

[18] At time T, we need to define U_{AB} such that $U_{AB}^{\mathsf{T}}U_{AB}\equiv S_B$.

Algorithm 1 LGSSM: Forward and backward recursive updates. The smoothed posterior $p(h_t|v_{1:T})$ is returned in the mean $\hat{h}_t^T$ and covariance P_t^T.

procedure FORWARD
- 1a: $P \leftarrow \Sigma$
- 1b: $P \leftarrow D\Sigma$, where $D \equiv I - \Sigma U_{AB}^{\mathsf{T}} \left(I + U_{AB}\Sigma U_{AB}^{\mathsf{T}}\right)^{-1} U_{AB}$
- 2a: $\hat{h}_1^0 \leftarrow \mu$
- 2b: $\hat{h}_1^0 \leftarrow D\mu$
- 3: $K \leftarrow PB^{\mathsf{T}}(BPB^{\mathsf{T}} + \Sigma_V)^{-1}$, $P_1^1 \leftarrow (I - KB)P$, $\hat{h}_1^1 \leftarrow \hat{h}_1^0 + K(v_t - B\hat{h}_1^0)$
- **for** $t \leftarrow 2, T$ **do**
  - 4: $P_t^{t-1} \leftarrow AP_{t-1}^{t-1}A^T + \Sigma_H$
  - 5a: $P \leftarrow P_t^{t-1}$
  - 5b: $P \leftarrow D_t P_t^{t-1}$, where $D_t \equiv I - P_t^{t-1}U_{AB}^{\mathsf{T}} \left(I + U_{AB}P_t^{t-1}U_{AB}^{\mathsf{T}}\right)^{-1} U_{AB}$
  - 6a: $\hat{h}_t^{t-1} \leftarrow A\hat{h}_{t-1}^{t-1}$
  - 6b: $\hat{h}_t^{t-1} \leftarrow D_t A\hat{h}_{t-1}^{t-1}$
  - 7: $K \leftarrow PB^{\mathsf{T}}(BPB^{\mathsf{T}} + \Sigma_V)^{-1}$, $P_t^t \leftarrow (I - KB)P$, $\hat{h}_t^t \leftarrow \hat{h}_t^{t-1} + K(v_t - B\hat{h}_t^{t-1})$
- **end for**

end procedure

procedure BACKWARD
- **for** $t \leftarrow T-1, 1$ **do**
  - $\overleftarrow{A}_t \leftarrow P_t^t A^{\mathsf{T}}(P_{t+1}^t)^{-1}$
  - $P_t^T \leftarrow P_t^t + \overleftarrow{A}_t(P_{t+1}^T - P_{t+1}^t)\overleftarrow{A}_t^{\mathsf{T}}$
  - $\hat{h}_t^T \leftarrow \hat{h}_t^t + \overleftarrow{A}_t(\hat{h}_{t+1}^T - A\hat{h}_t^t)$
- **end for**

end procedure

Equivalence of Algorithm 1 a and b

The filtered covariance P_t^t obtained from Algorithms 1 a and b are equivalents. Indeed, let suppose that we have demonstrated the equivalence at time $t-1$, then by repetitive application of the matrix inversion lemma[19] we obtain:

$$
\begin{aligned}
P_t^t &= P_t^{t-1} - P_t^{t-1}\tilde{B}^{\mathsf{T}}\left(\tilde{B}P_t^{t-1}\tilde{B}^{\mathsf{T}} + \tilde{\Sigma}_V\right)^{-1}P_t^{t-1} \\
&= \left((P_t^{t-1})^{-1} + \tilde{B}^{\mathsf{T}}\tilde{\Sigma}_V^{-1}\tilde{B}\right)^{-1} \\
&= \Big(\underbrace{(P_t^{t-1})^{-1} + S_A + S_B}_{P^{-1}} + \langle B\rangle^{\mathsf{T}}\left\langle\Sigma_V^{-1}\right\rangle\langle B\rangle\Big)^{-1} \\
&= P - P\langle B\rangle^{\mathsf{T}}\left(\langle B\rangle P\langle B\rangle^{\mathsf{T}} + \left\langle\Sigma_V^{-1}\right\rangle^{-1}\right)^{-1}\langle B\rangle P
\end{aligned}
$$

where P can be written as $P = P_t^{t-1} - P_t^{t-1}U_{AB}^{\mathsf{T}}(U_{AB}P_t^{t-1}U_{AB}^{\mathsf{T}} + I)^{-1}U_{AB}P_t^{t-1}$.

[19]Matrix inversion lemma: if the matrices A, B, C, D satisfy $B^{-1} = A^{-1} + C^{\mathsf{T}}D^{-1}C$, where all inverses are assumed to exist, then $B = A - AC^{\mathsf{T}}(CAC^{\mathsf{T}} + D)^{-1}CA$.

The contribution from the observations to form the mean $\hat{h}_t^t$ is equivalent in the two algorithms, indeed:

$$\begin{aligned}
P_t^{t-1}\tilde{B}^{\mathsf{T}}\big(\tilde{B}P_t^{t-1}\tilde{B}^{\mathsf{T}}+\tilde{\Sigma}_V\big)^{-1} &= P_t^{t-1}\tilde{B}^{\mathsf{T}}\Big(\tilde{\Sigma}_V^{-1}-\tilde{\Sigma}_V^{-1}\tilde{B}\big(\tilde{B}^{\mathsf{T}}\tilde{\Sigma}_V^{-1}\tilde{B}+(P_t^{t-1})^{-1}\big)^{-1}\tilde{B}^{\mathsf{T}}\tilde{\Sigma}_V^{-1}\Big)\\
&= \Big(P_t^{t-1}-P_t^{t-1}\tilde{B}^{\mathsf{T}}\tilde{\Sigma}_V^{-1}\tilde{B}\big(\tilde{B}^{\mathsf{T}}\tilde{\Sigma}_V^{-1}\tilde{B}+(P_t^{t-1})^{-1}\big)^{-1}\Big)\tilde{B}^{\mathsf{T}}\tilde{\Sigma}_V^{-1}\\
&= \Big(P_t^{t-1}-P_t^{t-1}\big(I+(P_t^{t-1})^{-1}(\tilde{B}^{\mathsf{T}}\tilde{\Sigma}_V^{-1}\tilde{B})^{-1}\big)^{-1}\Big)\tilde{B}^{\mathsf{T}}\tilde{\Sigma}_V^{-1}\\
&= \Big(P_t^{t-1}-P_t^{t-1}\big(P_t^{t-1}+(\tilde{B}^{\mathsf{T}}\tilde{\Sigma}_V^{-1}\tilde{B})^{-1}\big)^{-1}P_t^{t-1}\Big)\tilde{B}^{\mathsf{T}}\tilde{\Sigma}_V^{-1}\\
&= \big((P_t^{t-1})^{-1}+\tilde{B}^{\mathsf{T}}\tilde{\Sigma}_V^{-1}\tilde{B}\big)^{-1}\tilde{B}^{\mathsf{T}}\tilde{\Sigma}_V^{-1}\\
&= P_t^t\tilde{B}^{\mathsf{T}}\tilde{\Sigma}_V^{-1}\\
&= vert\left(P_t^t\left\langle B\right\rangle^{\mathsf{T}}\left\langle\Sigma_V^{-1}\right\rangle, P_t^tU_A^{\mathsf{T}}, P_t^tU_B^{\mathsf{T}}\right)\\
&= vert\left(P\left\langle B\right\rangle^{\mathsf{T}}\left(\left\langle B\right\rangle P\left\langle B\right\rangle^{\mathsf{T}}+\left\langle\Sigma_V^{-1}\right\rangle^{-1}\right)^{-1}, P_t^tU_A^{\mathsf{T}}, P_t^tU_B^{\mathsf{T}}\right)
\end{aligned}$$

The first element in vert is equivalent to the contribution of Algorithm 1 b, while the second and third do not contribute given that the corresponding elements in $\tilde{v}_t$ are zeros. Finally, the contribution from $\hat{A}h_{t-1}^{t-1}$ to form the mean $\hat{h}_t^t$ is

$$\begin{aligned}
\big(I-P_t^{t-1}\tilde{B}^{\mathsf{T}}(\tilde{B}P_t^{t-1}\tilde{B}^{\mathsf{T}}+\tilde{\Sigma}_V)^{-1}\tilde{B}\big)P_t^t(P_{t-1}^t)^{-1} &= \Big(I-P\left\langle B\right\rangle^{\mathsf{T}}\left(\left\langle B\right\rangle P\left\langle B\right\rangle^{\mathsf{T}}+\left\langle\Sigma_V^{-1}\right\rangle^{-1}\right)^{-1}\left\langle B\right\rangle\Big)P(P_{t-1}^t)^{-1}\\
&= \Big(I-P\left\langle B\right\rangle^{\mathsf{T}}\left(\left\langle B\right\rangle P\left\langle B\right\rangle^{\mathsf{T}}+\left\langle\Sigma_V^{-1}\right\rangle^{-1}\right)^{-1}\left\langle B\right\rangle\Big)D_tP_{t-1}^t(P_{t-1}^t)^{-1}
\end{aligned}$$

B.5 Parameter Covariance

The parameter covariance S_B introduced in Section 4.1 is given by:

$$\begin{aligned}
\left[(S_B)_t^n\right]_{jl} &= tr\left[\left\langle\Sigma_V^{-1}W_t^n(B_l-\left\langle B_l\right\rangle)(B_j-\left\langle B_j\right\rangle)^{\mathsf{T}}\right\rangle_{q(B,\Sigma_V^{-1})}\right]\\
&= \sum_{i,o,p=1}^{V}\left[\left\langle\left[\Sigma_V^{-1}\right]_{io}W_{op}^n\left[\Sigma_B\right]_{p+H(l-1),i+H(j-1)}\right\rangle_{q(\Sigma_V^{-1})}\right]\\
&= \sum_{i,o,p=1}^{V}\left\langle\left[\Sigma_V^{-1}\right]_{io}W_{op}^n\sum_{qr}\left[I_H\right]_{lr}\left[\Sigma_V\right]_{pq}\left[H_{BM}^{-1}\right]_{q+H(r-1),i+H(j-1)}\right\rangle\\
&= \sum_{i}^{V}W_{ii}^n\left[H_{BM}^{-1}\right]_{i+H(l-1),i+H(j-1)}.
\end{aligned}$$

In the particular case in which there are not missing observations this reduces to:

$$S_B = VH_B^{-1}.$$

Analogously $S_A=\sum_i H_{iA}$. Indeed:

$$\begin{aligned}
\left[S_A\right]_{jl} &= tr\left[\left\langle\Sigma_H^{-1}(A_j-\left\langle A_j\right\rangle)(A_l-\left\langle A_l\right\rangle)^{\mathsf{T}}\right\rangle\right]\\
&= \sum_{i,k}\left\langle\left[\Sigma_H^{-1}\right]_{ik}\left[\Sigma_H\right]_{ki}\left[H_{kA}\right]_{lj}\right\rangle\\
&= \sum_i\left[H_{iA}\right]_{jl}.
\end{aligned}$$

B.6 Hyperparameter Updates

Updates for β_j^k

If we compute the derivative of Eq. (6) with respect to the hyperparameter β_j, $j = 1, \ldots, H$ and set it to zero we obtain:

$$\beta_j = \frac{V}{\left\langle (B_j - \hat{B}_j)^\mathsf{T} \Sigma_V^{-1} (B_j - \hat{B}_j) \right\rangle},$$

where

$$\begin{aligned}\left\langle (B_j - \hat{B}_j)^\mathsf{T} \Sigma_V^{-1} (B_j - \hat{B}_j) \right\rangle &= \underbrace{\left\langle (B_j - \langle B_j \rangle)^\mathsf{T} \Sigma_V^{-1} (B_j - \langle B_j \rangle) \right\rangle}_{[S_B]_{jj}} \\ &\quad + \langle B_j \rangle^\mathsf{T} \left\langle \Sigma_V^{-1} \right\rangle \langle B_j \rangle - 2 \left\langle B_j^\mathsf{T} \right\rangle \left\langle \Sigma_V^{-1} \right\rangle \hat{B}_j + \hat{B}_j^\mathsf{T} \left\langle \Sigma_V^{-1} \right\rangle \hat{B}_j.\end{aligned}$$

Updates for α_{ij}^k

If we compute the derivative of Eq. (6) with respect to the hyperparameter α_{ij}, $i, j = 1, \ldots, H$, and set it to zero we obtain:

$$\alpha_{ij} = \frac{1}{\left\langle \left[\Sigma_H^{-1}\right]_{ii} (A_{ij} - \hat{A}_{ij})^2 \right\rangle},$$

where

$$\begin{aligned}\left\langle \left[\Sigma_H^{-1}\right]_{ii} (A_{ij} - \hat{A}_{ij})^2 \right\rangle &= \underbrace{\left\langle \left[\Sigma_H^{-1}\right]_{ii} (A_{ij} - \langle A_{ij} \rangle)^2 \right\rangle}_{\left[H_{iA}^{-1}\right]_{jj}} \\ &\quad + \left\langle \left[\Sigma_H^{-1}\right]_{ii} \right\rangle \langle A_{ij} \rangle^2 - 2 \left\langle \left[\Sigma_H^{-1}\right]_{ii} \right\rangle \langle A_{ij} \rangle \hat{A}_{ij} + \left\langle \left[\Sigma_H^{-1}\right]_{ii} \right\rangle \hat{A}_{ij}^2.\end{aligned}$$

Updates for b_1^k, b_2^k

For the constraint $\Sigma_V^{-1} = dg\,(\rho)$, each diagonal element ρ_i follows a Gamma prior $\mathcal{G}(b_1^i, b_2^i)$.
In order to constrain b_1^i to be positive, we set $b_1^i = b^2$. The derivative of Eq. (6) with respect to b is given by:

$$2b \log b_2^i - 2b\psi(b^2) + 2b \left\langle \log \rho_i \right\rangle_{q(\rho_i)},$$

where

$$\begin{aligned}\left\langle \log \rho_i \right\rangle_{q(\rho_i)} &= \frac{q_2^{q_1}}{\Gamma(q_1)} \int_{\rho_i} \rho_i^{q_1 - 1} e^{-q_2 \rho_i} \log \rho_i \\ &= \frac{q_2^{q_1}}{\Gamma(q_1)} \frac{\partial}{\partial q_1} \frac{\Gamma(q_1)}{q_2^{q_1}} \\ &= \psi(q_1) - \log q_2.\end{aligned}$$

Given that we cannot obtain a closed form update for b_1^i, we have to use some optimization method. Setting to zero the derivative of Eq. (6) with respect to b_2^i, we obtain:

$$b_2^i = \frac{b_1^i}{\langle \rho_i \rangle_{q(\rho_i)}}.$$

Similar updates can be obtained for a_1^k and a_2^k.

Updates for ν_V^k, S_V^k

If Σ_V^{-1} follows a Wishart prior distribution $p(\Sigma_V^{-1}|\nu_V, S_V) = \mathcal{W}(\nu_V, S_V)$, the derivative of Eq. (6) with respect to ν_V is given by:

$$-\frac{V}{2}\log 2 - \frac{1}{2}\log|S_V| - \sum_{i=1}^{V}\psi\left(\frac{\nu_V + 1 - i}{2}\right) + \left\langle \log|\Sigma_V^{-1}|\right\rangle_{q(\Sigma_V^{-1})},$$

where

$$\begin{aligned}\left\langle \log|\Sigma_V^{-1}|\right\rangle_{q(\Sigma_V^{-1})} &= \frac{1}{Z_q}\int \log|\Sigma_V^{-1}||\Sigma_V^{-1}|^{\frac{\nu_q - V - 1}{2}} e^{-\frac{1}{2}tr\left[S_q^{-1}\Sigma_V^{-1}\right]} \\ &= \frac{2}{Z_q}\frac{\partial Z_q}{\partial \nu_q} \\ &= \sum_i \psi\left(\frac{\nu_q + 1 - i}{2}\right) + V\log 2 + \log|S_q|.\end{aligned}$$

Setting to zero the derivative of Eq. (6) with respect to S_V, we obtain:

$$S_V = \frac{1}{\nu_V}\Sigma_V^{-1}.$$

Updates for μ_μ^k, Σ_μ^k

Setting to zero the derivative of Eq. (6) with respect to μ_μ and Σ_μ we obtain:

$$\begin{aligned}\mu_\mu &= \langle\mu\rangle \\ \Sigma_\mu &= \left\langle \mu\mu^{\mathsf{T}}\Sigma^{-1}\right\rangle - \langle\mu\rangle\mu_\mu^{\mathsf{T}}\left\langle\Sigma^{-1}\right\rangle - \mu_\mu\langle\mu\rangle^{\mathsf{T}}\left\langle\Sigma^{-1}\right\rangle + \mu_\mu\mu_\mu^{\mathsf{T}}\left\langle\Sigma^{-1}\right\rangle \\ &= \left(\sum_{n=1}^{N} q(z^n = k)I_H + \Sigma_\mu^{-1}\right)^{-1} + (\mu_h - \mu_\mu)(\mu_h - \mu_\mu)^{\mathsf{T}}\left\langle\Sigma^{-1}\right\rangle\end{aligned}$$

Updates for σ_1, σ_2

The updates are similar to the updates for b_1 and b_2 described above.

Updates for ν, S

The updates are similar to the updates for ν_V and S_V described above.

Updates for γ

In order to constrain γ to be positive we set $\gamma = \delta^2$. The derivative of Eq. (6) with respect to δ is given by:

$$-2\delta\left(\psi(\delta^2) - \psi(N + \delta^2) - \psi\left(\frac{\delta^2}{K}\right)\right) + 2\frac{\delta}{K}\sum_{k=1}^{K}\left\langle\psi(N_k + \delta^2/K)\right\rangle_{\prod_{n=1}^{N} q(z^n)},$$

which does not give a close form update for γ.

B.7 Computing the Log-likelihood Bound

The log-likelihood bound (Eq. (6)) can be rewritten as:

$$\begin{aligned}\mathcal{F} = &\sum_{n=1}^{N}\sum_{k=1}^{K} q(z^n = k)H_q(h_{1:T}^n|z^n = k) + \sum_{n=1}^{N} H_q(z^n) \\ &- \sum_{k=1}^{K}\left\langle \log\frac{q\left(A|\Sigma_H^{-1}\right)}{p\left(A|\Sigma_H^{-1}\right)}\right\rangle_{q\left(A,\Sigma_H^{-1}\right)} - \sum_{k=1}^{K}\left\langle \log\frac{q\left(B|\Sigma_V^{-1}\right)}{p\left(B|\Sigma_V^{-1}\right)}\right\rangle_{q\left(B,\Sigma_V^{-1}\right)} - \sum_{k=1}^{K}\left\langle \log\frac{q\left(\mu|\Sigma^{-1}\right)}{p\left(\mu|\Sigma^{-1}\right)}\right\rangle_{q(\mu,\Sigma^{-1})} \\ &- \sum_{k=1}^{K}\left\langle \log\frac{q\left(\Sigma_H^{-1}\right)}{p\left(\Sigma_H^{-1}\right)}\right\rangle_{q\left(\Sigma_H^{-1}\right)} - \sum_{k=1}^{K}\left\langle \log\frac{q\left(\Sigma_V^{-1}\right)}{p\left(\Sigma_V^{-1}\right)}\right\rangle_{q\left(\Sigma_V^{-1}\right)} - \sum_{k=1}^{K}\left\langle \log\frac{q\left(\Sigma^{-1}\right)}{p\left(\Sigma^{-1}\right)}\right\rangle_{q(\Sigma^{-1})} \\ &+ \left\langle \log p\left(z^{1:N}\right)\right\rangle_{\prod_{n=1}^{N} q(z^n)} + \sum_{n=1}^{N}\sum_{k=1}^{K} q(z^n = k)\left\langle \log p\left(v_{1:T}^n, h_{1:T}^n|\Theta^k\right)\right\rangle_{q(\Theta)q\left(h_{1:T}^n|z^n=k\right)}.\end{aligned}$$

Each entropic element in the first term can be computed as:

$$-\left\langle\sum_{t=1}^{T-1}\log\frac{q\left(h_t^n,h_{t+1}^n|v_{1:T}\right)}{q\left(h_{t+1}^n|v_{1:T}\right)}+\log q(h_T^n|v_{1:T})\right\rangle_{q\left(h_{1:T}^n|v_{1:T},z^n=k\right)}=\sum_{t=1}^{T-1}\frac{1}{2}\log\det(P_t^T-\overleftarrow{A}_tP_{t+1}^T\overleftarrow{A}_t^{\mathsf{T}})$$
$$+\frac{1}{2}\log\det(P_T^T)+\frac{T}{2}H(1+\log(2\pi)),$$

where P_t^T is the covariance of $q\left(h_t^n|v_{1:T}\right)$, $\overleftarrow{A}_t=P_t^tA^{\mathsf{T}}\left(AP_t^tA^{\mathsf{T}}+\Sigma_H\right)^{-1}$, and where we have used the property

$$\det\begin{pmatrix}A & B\\ C & D\end{pmatrix}=\det\left(AD-BD^{-1}CD\right).$$

Terms such as $\left\langle\log q\left(A|\Sigma_H^{-1}\right)/p\left(A|\Sigma_H^{-1}\right)\right\rangle_{q\left(A,\Sigma_H^{-1}\right)}$ can be computed as:

$$\frac{1}{2}\left(\log\left(\frac{\det\Sigma_p}{\det\Sigma_q}\right)+\mathrm{tr}\Sigma_p^{-1}\Sigma_q+(\mu_p-\mu_q)^{\mathsf{T}}\left\langle\Sigma_p^{-1}\right\rangle_{q(\Sigma_H)}(\mu_p-\mu_q)-H^2\right),$$

where μ_q and Σ_q are the mean and covariance of $q(A|\Sigma_H)$, and μ_p and Σ_p are the mean and covariance of $p(A|\Sigma_H)$. Notice that this simple formula comes form our choice for the prior, which makes Σ_q and Σ_p to have a common dependency on Σ_H and the means μ_q and μ_p not dependent on Σ_H.

Notice also that when $q(z^n=k)=0$ for a given k and $n=1,\ldots,N$, that is when no sequences are assigned to component k, we have $q\left(A|\Sigma_H^{-1}\right)=p\left(A|\Sigma_H^{-1}\right)$, $q\left(B|\Sigma_V^{-1}\right)=p\left(B|\Sigma_V^{-1}\right)$, $q\left(\Sigma_H^{-1}\right)=p\left(\Sigma_H^{-1}\right)$ and $q\left(\Sigma_V^{-1}\right)=p\left(\Sigma_V^{-1}\right)$. As a consequence $\alpha_{ij}^k,\beta_j^k\to\infty$. This can be seen from the updates in Appendix B.2 and Appendix B.6.

C

C.1 Uncollapsed Model

Consider the model described in Section 4.1. If we do not integrate out π from the model, we have the approximation:

$$p\left(z^{1:N},h_{1:T}^{1:N},\Theta^{1:K},\pi|v_{1:T}^{1:N},\hat\Theta^{1:K}\right)\approx\left\{\prod_{n=1}^{N}q\left(z^n\right)q\left(h_{1:T}^n|z^n\right)\right\}\prod_{k=1}^{K}q\left(\Theta^k\right)q(\pi),$$

which gives the following lower bound on the log-likelihood:

$$\mathcal{F}\equiv\sum_{k=1}^{K}H_q(\Theta^k)+\sum_{n=1}^{N}\sum_{k=1}^{K}q\left(z^n=k\right)H_q\left(h_{1:T}^n|z^n=k\right)+\sum_{n=1}^{N}H_q\left(z^n\right)+H_q\left(\pi\right)+\sum_{k=1}^{K}\left\langle\log p\left(\Theta^k|\hat\Theta^k\right)\right\rangle_{q(\Theta^k)}$$
$$+\left\langle\log p\left(\pi\right)\right\rangle_{q(\pi)}+\left\langle\log p\left(z^{1:N}|\pi\right)\right\rangle_{\prod_n q(z^n)q(\pi)}+\sum_{n=1}^{N}\sum_{k=1}^{K}q(z^n=k)\left\langle\log p\left(v_{1:T}^n,h_{1:T}^n|\Theta^k\right)\right\rangle_{q(\Theta^k)q\left(h_{1:T}^n|z^n=k\right)}.$$

The updates for the q distribution are:

$$q\left(\Theta^k\right)\propto p\left(\Theta^k|\hat\Theta^k\right)e^{\sum_{n=1}^{N}q(z^n=k)\left\langle\log p\left(v_{1:T}^n,h_{1:T}^n|\Theta^k\right)\right\rangle_{q\left(h_{1:T}^n|z^n=k\right)}}$$
$$q\left(z^n=k\right)\propto e^{H_q\left(h_{1:T}^n|z^n=k\right)+\left\langle\log p(z^n=k|\pi)\right\rangle_{q(\pi)}+\left\langle\log p\left(v_{1:T}^n,h_{1:T}^n|\Theta^k\right)\right\rangle_{q\left(h_{1:T}^n|z^n=k\right)q(\Theta^k)}}$$
$$q\left(h_{1:T}^n|z^n=k\right)\propto e^{\left\langle\log p\left(v_{1:T}^n,h_{1:T}^n|\Theta^k\right)\right\rangle_{q(\Theta^k)}}$$
$$q(\pi)\propto p\left(\pi\right)e^{\left\langle\log p\left(z^{1:N}|\pi\right)\right\rangle_{\prod_n q(z^n)}}$$

The update for π gives a Dirichlet distribution with parameters $\lambda_k=\gamma/K+\sum_{n=1}^{N}q\left(z^n=k\right)$. The different terms in the bound can be computed as follows:

$$\left\langle \log p\left(z^{1:N}|\pi\right)\right\rangle_{q(z^{1:N})q(\pi)} = \sum_{n=1}^{N}\sum_{k} q\left(z^n=k\right)\left\langle \log p\left(z^n=k|\pi\right)\right\rangle_{q(\pi)}$$
$$= \sum_{n=1}^{N}\left(\sum_{k} q\left(z^n=k\right)\left(\psi(\lambda_k)-\psi\left(\sum_i \lambda_i\right)\right)\right)$$

$$\left\langle \log p\left(\pi\right)\right\rangle_{q(\pi)} = \log\frac{\Gamma\left(\gamma\right)}{\Gamma\left(\gamma/K\right)^K} + \sum_j \left\langle \log\left(\pi_j\right)^{\gamma/K-1}\right\rangle_{q(\pi)}$$
$$= \log\frac{\Gamma\left(\gamma\right)}{\Gamma\left(\gamma/K\right)^K} + \frac{\Gamma\left(\sum_i\lambda_i\right)}{\prod_i\Gamma\left(\lambda_i\right)}\sum_j\int_\pi \log\left(\pi_j\right)^{\gamma/K-1}\prod_i\left(\pi_i\right)^{\lambda_i-1}$$
$$= \log\frac{\Gamma\left(\gamma\right)}{\Gamma\left(\gamma/K\right)^K} + \left(\gamma/K-1\right)\frac{\Gamma\left(\sum_i\lambda_i\right)}{\prod_i\Gamma\left(\lambda_i\right)}\sum_j\frac{\partial}{\partial\lambda_j}\frac{\prod_i\Gamma\left(\lambda_i\right)}{\Gamma\left(\sum_i\lambda_i\right)}$$
$$= \log\frac{\Gamma\left(\gamma\right)}{\Gamma\left(\gamma/K\right)^K} + \left(\gamma/K-1\right)\sum_j\left(\psi(\lambda_j)-\psi\left(\sum_i\lambda_i\right)\right)$$

$$-\left\langle \log q\left(\pi\right)\right\rangle_{q(\pi)} = -\log\frac{\Gamma\left(\sum_i\lambda_i\right)}{\prod_i\Gamma\left(\lambda_i\right)} - \sum_i\left\langle \log\left(\pi_i\right)^{\lambda_i-1}\right\rangle_{q(\pi)}$$
$$= -\log\frac{\Gamma\left(\sum_i\lambda_i\right)}{\prod_i\Gamma\left(\lambda_i\right)} - \frac{\Gamma\left(\sum_i\lambda_i\right)}{\prod_i\Gamma\left(\lambda_i\right)}\sum_j\int_\pi \log\left(\pi_j\right)^{\lambda_j-1}\prod_i\left(\pi_i\right)^{\lambda_i-1}$$
$$= -\log\frac{\Gamma\left(\sum_i\lambda_i\right)}{\prod_i\Gamma\left(\lambda_i\right)} - \sum_j\left(\lambda_j-1\right)\left(\psi(\lambda_j)-\psi\left(\sum_i\lambda_i\right)\right)$$

C.2 Priors on the Hyperparameters

Consider the model described in Section 4.1. Instead of learning the hyperparameters with the ML-II approach, we could put a prior distribution. Under the assumption $q(\Theta^{1:K},\hat{\Theta}^{1:K}) \equiv q(\hat{\Theta}^{1:K})q(\Theta^{1:K})$, the lower bound on the log-likelihood is given by:

$$\mathcal{F} \equiv \sum_{k=1}^{K} H_q(\Theta^k) + \sum_{k=1}^{K} H_q(\hat{\Theta}^k) + \sum_{n=1}^{N}\sum_{k=1}^{K} q\left(z^n=k\right) H_q\left(h^n_{1:T}|z^n=k\right) + \sum_{n=1}^{N} H_q\left(z^n\right)$$
$$+ \sum_{k=1}^{K}\left\langle \log p\left(\Theta^k|\hat{\Theta}^k\right)\right\rangle_{q(\Theta^k)q\left(\hat{\Theta}^k\right)} + \sum_{k=1}^{K}\left\langle \log p\left(\hat{\Theta}^k\right)\right\rangle_{q\left(\hat{\Theta}^k\right)} + \left\langle \log p\left(z^{1:N}|\gamma\right)\right\rangle_{\prod_n q(z^n)}$$
$$+ \sum_{n=1}^{N}\sum_{k=1}^{K} q(z^n=k)\left\langle \log p\left(v^n_{1:T},h^n_{1:T}|\Theta^k\right)\right\rangle_{q(\Theta^k)q\left(h^n_{1:T}|z^n=k\right)}.$$

The updates for q are given by:

$$q\left(\hat{\Theta}^k\right) \propto p\left(\hat{\Theta}^k\right) e^{\left\langle \log p\left(\Theta^k|\hat{\Theta}^k\right)\right\rangle_{q(\Theta^k)}}$$
$$q\left(\Theta^k\right) \propto e^{\left\langle \log p\left(\Theta^k|\hat{\Theta}^k\right)\right\rangle_{q(\hat{\Theta}^k)}} e^{\sum_{n=1}^{N} q(z^n=k)\left\langle \log p\left(v^n_{1:T},h^n_{1:T}|\Theta^k\right)\right\rangle_{q\left(h^n_{1:T}|z^n=k\right)}}$$
$$q\left(z^n=k\right) \propto e^{H_q(h^n_{1:T}|z^n=k) + \langle \log p(z^n=k|z^{\neg n},\gamma)\rangle_{\prod_{m\neq n} q(z^m)} + \left\langle \log p\left(v^n_{1:T},h^n_{1:T}|\Theta^k\right)\right\rangle_{q\left(h^n_{1:T}|z^n=k\right)q(\Theta^k)}}$$
$$q\left(h^n_{1:T}|z^n=k\right) \propto e^{\left\langle \log p\left(v^n_{1:T},h^n_{1:T}|\Theta^k\right)\right\rangle_{q(\Theta^k)}}$$

This approach can be taken for α_{ij}^k and β_j^k. However, for $\nu_V^k, S_V^k, \nu_H^k, S_H^k, \nu^k, S^k$ $(a_1^k, a_2^k, b_1^k, b_2^k, \sigma_1^k, \sigma_2^k)$ and γ we need to use the ML-II approach, since $q\left(\hat{\Theta}^k\right)$ would be a unknown distribution[20]. The ML-II approach has to be used also for μ_μ^k and Σ_μ^k, since, unless heavy constraints are imposed, we cannot find a prior $p(\mu_\mu^k, \Sigma_\mu^k)$ which results in a known distribution $q(\mu_\mu^k, \Sigma_\mu^k)$.

Determining $q(\beta_j^k)$
$p(\beta_j) = \mathcal{G}(\beta 1, \beta 2)$, where $\beta 1$ and $\beta 2$ are fixed to certain values.

$$\begin{aligned}
q(\beta_j) &\propto e^{\left\langle \log p\left(B|\Sigma_V^{-1}, \beta_j\right)\right\rangle_{q\left(B, \Sigma_V^{-1}\right)}} p(\beta_j) \\
&\propto \beta_j^{\frac{V}{2}} e^{-\frac{\beta_j}{2}\left\langle \left(B_j - \hat{B}_j\right)^\intercal \Sigma_V^{-1}\left(B_j - \hat{B}_j\right)\right\rangle} \beta_j^{\beta 1 - 1} e^{-\beta 2 \beta_j} \\
&= \beta_j^{\beta 1 - 1 + \frac{V}{2}} e^{-\left(\beta 2 + \frac{1}{2}\left\langle \left(B_j - \hat{B}_j\right)^\intercal \Sigma_V^{-1}\left(B_j - \hat{B}_j\right)\right\rangle\right)\beta_j} \\
&= \mathcal{G}\left(\beta 1 + \frac{V}{2}, \beta 2 + \frac{1}{2}\left\langle \left(B_j - \hat{B}_j\right)^\intercal \Sigma_V^{-1}\left(B_j - \hat{B}_j\right)\right\rangle\right)
\end{aligned}$$

Determining $q(\alpha_{ij}^k)$
Similarly,

$$q(\alpha_{ij}) = \mathcal{G}\left(\alpha 1 + \frac{1}{2}, \alpha 2 + \frac{1}{2}\left\langle \left[\Sigma_H^{-1}\right]_{ii}\left(A_{ij} - \hat{A}_{ij}\right)^2\right\rangle\right)$$

For $q\left(B^k | \left[\Sigma_V^k\right]^{-1}\right)$ and $q\left(A^k | \left[\Sigma_H^k\right]^{-1}\right)$, the updates are similar to the ML-II case, with the difference that the optimal values for the hyperparameters are replaced by the mean values $\frac{\beta 1 + V/2}{\beta 2 + \frac{1}{2}\left\langle \left(B_j - \hat{B}_j\right)^\intercal \Sigma_V^{-1}\left(B_j - \hat{B}_j\right)\right\rangle}$ and $\frac{\alpha 1 + 1/2}{\alpha 2 + \frac{1}{2}\left\langle \left[\Sigma_H^{-1}\right]_{ii}\left(A_{ij} - \hat{A}_{ij}\right)^2\right\rangle}$. Notice that for $\beta 1, \beta 2, \alpha 1, \alpha 2 \to 0$ the optimal values for the hyperparameters obtained by ML-II and these mean values are the same. That is using a ML-II approach is equivalent to defining Gamma priors on β_j^k and α_{ij}^k with $\beta 1, \beta 2, \alpha 1, \alpha 2 = 0$.

[20]For example, using an inverse-Gamma and Gamma priors on a_1 and a_2 respectively, would result in a Gamma distribution $q(a_2)$ and in an unknown distribution $q(a_1)$.

References

[1] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley, New York, 2000.

[2] Z. Ghahramani and M. J. Beal. Variational inference for Bayesian mixtures of factor analysers. In *Advances in Neural Information Processing Systems 12 (NIPS)*, pages 449–455, 2000.

[3] D. Görür. *Nonparametric Bayesian Discrete Latent Variable Models for Unsupervised Learning*. Phd thesis, Technischen Universität Berlin, 2007.

[4] A. Kottas. Dirichlet Process Mixtures of Beta Distributions, with Applications to Density and Intensity Estimation. In *Workshop on Learning with Nonparametric Bayesian Methods, 23rd International Conference on Machine Learning (ICML)*, 2006.

[5] C. E. Rasmussen. The infinite Gaussian mixture model. In *Advances in Neural Information Processing Systems 12 (NIPS)*, pages 554–560, 2000.

[6] K. Kalpakis, S. Gada, and V. Puttagunta. Distance measures for effective clustering of ARIMA time-series. pages 273–280, 2001.

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

[8] Y. Bar-Shalom and X.-R. Li. *Estimation and Tracking: Principles, Techniques and Software*. Artech House, 1998.

[9] M. S. Grewal and A. P. Andrews. *Kalman Filtering: Theory and Practice Using MATLAB*. John Wiley and Sons, Inc., 2001.

[10] R. H. Shumway and D. S. Stoffer. *Time Series Analysis and its Applications*. Springer, 2000.

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

[12] M. J. Beal. *Variational Algorithms for Approximate Bayesian Inference*. Ph.D. thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

[13] D. J. C. MacKay. Ensemble learning and evidence maximisation. Unpublished manuscipt: www.variational-bayes.org, 1995.

[14] H. Valpola and J. Karhunen. An unsupervised ensemble learning method for nonlinear dynamic state-space models. *Neural Computation*, 14:2647–2692, 2002.

[15] D. Barber and S. Chiappa. Unified inference for variational Bayesian linear Gaussian state-space models. In *Advances in Neural Information Processing Systems 20 (NIPS)*, 2006.

[16] M. J. Beal, F. Falciani, Z. Ghahramani, C. Rangel, and D. L. Wild. A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics*, 21:349–356, 2005.

[17] A. T. Cemgil and S. J. Godsill. Probabilistic phase vocoder and its application to interpolation of missing values in audio signals. In *13th European Signal Processing Conference*, 2005.

[18] S. Chiappa. Inference in the variational linear Gaussian state-space model using belief propagation and the junction tree algorithm, 2007. Technical Notes.

[19] M. Morf and T. Kailath. Square-root algorithms for least-squares estimation. *IEEE Transactions on Automatic Control*, 20:487–497, 1975.

[20] P. Park and T. Kailath. New square-root smoothing algorithms. *IEEE Transactions on Automatic Control*, 41:727–732, 1996.

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

[22] C. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2:1152–1174, 1974.

[23] K. Kurihara, M. Welling, and Y. W. Teh. Collapsed variational Dirichlet process mixture models. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2007.

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

[25] D. J. C. MacKay. Local minima, symmetry-breaking, and model pruning in variational free energy minimization. Inference Group, Cavendish Laboratory, Cambridge, U.K., 2001.

List of Notations