Output Grouping using Dirichlet Mixtures of Linear Gaussian State-Space Models

Silvia Chiappa
MPI for Biological Cybernetics
Spemannstrasse 38,
Tübingen, Germany
silvia.chiappa@tuebingen.mpg.de

David Barber
Department of Computer Science
University College London
Gower Street, London, U.K.
d.barber@cs.ucl.ac.uk

Abstract

We consider a model to cluster the components of a vector time-series. The task is to assign each component of the vector time-series to a single cluster, basing this assignment on the simultaneous dynamical similarity of the component to other components in the cluster. This is in contrast to the more familiar task of clustering a set of time-series based on global measures of their similarity. The model is based on a Dirichlet Mixture of Linear Gaussian State-Space models (LGSSMs), in which each LGSSM is treated with a prior to encourage the simplest explanation. The resulting model is approximated using a 'collapsed' variational Bayes implementation.

1 Introduction

Consider a V-dimensional time-series $v \equiv \{v_1, \ldots, v_T\}$ where, at each time t, v_t denotes a vector having components v_t^i, $i = 1, \ldots, V$. This paper addresses the task of clustering the component time-series[1] v^i based on their simultaneous dynamical similarity, see Fig 1. We are interested in the case in which the number of clusters is not known in 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.

To achieve these desiderate, we use a form of Dirichlet Mixture of Bayesian Linear Gaussian State-Space models. A Gaussian prior is used to encourage the model to have the smallest parameterization consistent with the data, and a Polya distribution is used on the assignments to determine an appropriate number of clusters. In our model, output components are assigned to the same cluster if generated by the *same* realization of a linear Gaussian dynamical process.

[1]For ease of notation, we will consider only a single time-series, although the generalization carries over naturally to a set of time-series. Similarly, we consider grouping only scalar outputs, although splitting the vector v into a set of sub-vectors and clustering these is a straightforward extension.

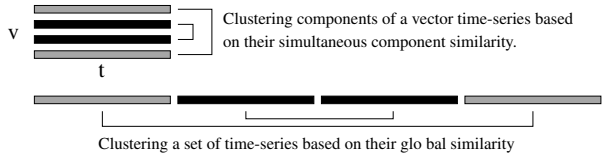

Figure 1. (Top) Time-series clustering based on simultaneous activity, versus (Below) clustering based on global similarity.

An alternative, and perhaps more familiar clustering task, is to assign two outputs to the same cluster when generated by a *different* realization of the same linear Gaussian dynamical process. In such a viewpoint, clustering would not be based on the simultaneous behavior of the components of the vector, but rather on a measure of the global similarity of the components. For a depiction of this key difference, see Fig 1.

The paper is organized as follows: In the next section we recall the basic theory of the LGSSM, which will be married with the Dirichlet Mixture Model in Section 3. A toy demonstration of the method is given in Section 4. Fuller technical details and derivations are to be found in [1], which discusses in addition alternative clustering models.

2 Linear Gaussian State-Space Models

In a Linear Gaussian State-Space Model[2][2] a sequence of observations $v \equiv \{v_1, \ldots, v_T\}$, $v_t \in \Re^V$, is generated according to a latent Markovian linear dynamical system on states $h \equiv \{h_1, \ldots h_T\}$, $h_t \in \Re^H$:

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

where $\mathcal{N}(x|\mu, \Sigma)$ denotes a Gaussian in variable x with mean μ and covariance Σ, and $\mathbf{0}_X$ denotes an X-

[2]This model is also called a Linear Dynamical System. We prefer not to use the terminology Kalman Filter/Smoother since this refers to a particular kind of inference on a LGSSM.

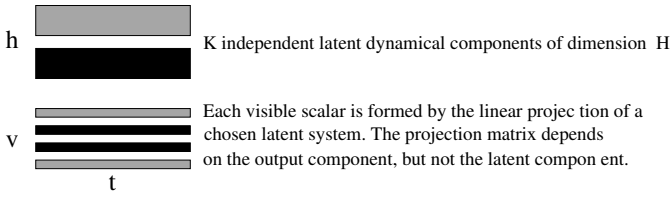

Figure 2. Output clustering based on independent latent dynamical systems.

dimensional zero vector. A probabilistic description of this model is

$$p(v, h|\Theta) = \prod_{t=1}^{T} p(v_t|h_t, \Theta)p(h_t|h_{t-1}, \Theta),$$

where we define the transitions

$$p(h_t|h_{t-1}, \Theta) \equiv \mathcal{N}(h_t|Ah_{t-1}, \Sigma_H) \tag{2}$$

and emissions

$$p(v_t|h_t, \Theta) \equiv \mathcal{N}(v_t|Bh_t, \Sigma_V).$$

The dynamics is initialized with $p(h_1|h_0) \equiv \mathcal{N}(h_1|\mu, \Sigma)$. The combined set of parameters of the model is denoted with $\Theta = \{A, B, \Sigma_H, \Sigma_V, \mu, \Sigma\}$. Since the model is a simple pairwise Markov Gaussian chain, most quantities of interest, such as the posterior density $p(h_t|v, \Theta)$, posterior entropy of h and likelihood $p(v|\Theta) = \int_h p(v, h|\Theta)$ can be computed efficiently in $O(T)$ operations[2].

3 Dirichlet Mixture of Bayesian LGSSMs

Our model assumes the presence of a set of K latent dynamical systems $h^1, \ldots, h^K$, each precessing independently according to

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

as in Eq. (2). Dynamical system k projects to a subset of the visible components, forming a linear mixing of the states h_t^k to produce the outputs associated with component k. In this way, we form a partition of the outputs into groups: within group k the outputs are dependent and can be explained by the k^{th} linear dynamical system.

The key ingredient is to introduce an indicator variable $z^i \in \{1, \ldots, K\}$ which assigns each output v^i to a particular cluster latent dynamics h^k (see Fig 2). This will be achieved by

$$p(v^i|h, z^i = k) = p(v^i|h^k).$$

Each output component is therefore assigned to a single cluster, although each h^k is potentially responsible for several outputs.

In a standard mixture model, the indicators z^i are independent. However, in the Dirichlet mixture case, we specify a joint distribution on the variables $z = \{z^1, \ldots, z^V\}$ to encourage the components of the mixture to be used in a parsimonious fashion, enabling the model to automatically identify a reasonable number of clusters[3]. Furthermore, we place a Bayesian prior on the parameters of each LGSSM to bias each dynamical system into its simplest form.

The joint density on all variables (dropping the hyperparameters from the notation) is given by:

$$p(v, z, h, \Theta) = p(z)p(\Theta)p(v|h, z, \Theta)p(h|\Theta).$$

This is composed of the LGSSM emission $p(v|h, z, \Theta)$, transitions $p(h|\Theta)$, parameter prior $p(\Theta)$ and joint indicator prior $p(z)$. We will consider each term in more detail below.

Indicator Prior $p(z)$

To model the joint cluster allocations we define

$$p(z) = \int_\pi \left\{\prod_i p(z^i|\pi)\right\} p(\pi), \tag{3}$$

where $p(z^i = k|\pi) \equiv \pi_k$ is a multinomial distribution. Using the Dirichlet as the multinomial conjugate prior[3]

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

the integral in Eq. (3) gives rise to the Polya distribution:

$$p(z) = \frac{\Gamma(\gamma)}{\Gamma(V+\gamma)} \prod_{k=1}^{K} \frac{\Gamma(V_k + \gamma/K)}{\Gamma(\gamma/K)}, \tag{4}$$

where

$$V_k \equiv \sum_{i=1}^{V} I[z^i = k] \tag{5}$$

counts the number of times that state k occurs in the indicators[4]. In the limit of infinite K, the prior expected number of clusters for a set of V outputs is [4]

$$\sum_{i=1}^{V} \frac{\gamma}{\gamma + i - 1} \approx \gamma \log\left(\frac{V}{\gamma} + 1\right).$$

Whilst it is possible to optimize the marginal likelihood with respect to γ, in our experiments we typically set γ to around the maximum number of possible clusters[5].

In our work, we will consider K to be finite. This is in contrast to Dirichlet *Process* Mixture Models[3, 5, 6], in which the $K \to \infty$ limit is formally taken. This can be

[3] The scaling γ/K ensures a sensible limit as $K \to \infty$.

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

[5] The distribution of the number of clusters is heavily skewed, which creates a bias towards using few clusters.

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 will generate at times new clusters until sufficiently many are present to explain the data well. This is the origin of the Chinese Restaurant Process interpretation of the Dirichlet Process (see, for example [7]). 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.

LGSSM Emissions $p(v|z,h,\Theta)$

The emission term, which is central to clustering, is

$$p(v|z,h,\Theta) = \prod_{i=1}^{V} p(v^i|h,z^i,\Theta).$$

When z^i is in state k, the dynamics of the k^{th} latent state is projected to the observation:

$$p(v_t^i|h_t, z^i = k, \Theta) \equiv \mathcal{N}\left(v_t^i|B^i h_t^k, [\Sigma_V]_{ii}\right),$$

where $B^i \equiv B_{i,\cdot}$ is a $1 \times H$ vector. Each time point contributes independently, giving:

$$p(v^i|h, z^i = k, \Theta) = \prod_t \mathcal{N}\left(v_t^i|B^i h_t^k, [\Sigma_V]_{ii}\right).$$

Hence $z^i = k$ has the effect of assigning output i to dynamical system k.

LGSSM Transitions $p(h|\Theta)$

The transitions term is given by

$$p(h|\Theta) = \prod_{k=1}^{K} p(h^k|\Theta^k).$$

Each of the K linear dynamical systems proceeds independently of the rest under the usual LGSSM Markovian dynamics (see Eq. (1))

$$p(h^k|\Theta^k) = \prod_t p(h_t^k|h_{t-1}^k, \Theta^k).$$

LGSSM Parameter Prior $p(\Theta)$

The covariances are taken to be diagonal and parameterized, for convenience, via their inverses

$$\Sigma_H^k = diag(\left[\tau^k\right]^{-1}), \qquad \Sigma_V = diag(\rho^{-1}),$$

where each diagonal element follows a Gamma distribution

$$p(\tau_i^k) = \text{Gamma}(\tau_i^k|a_1, a_2), \qquad p(\rho_i) = \text{Gamma}(\rho_i|b_1, b_2).$$

We fix a_1, a_2, b_1, b_2 to achieve broad priors. To bias the transition parameters to preferred values, we use a set of Gaussian priors

$$p(A^k|\alpha^k, \tau^k) \propto \prod_{i,j=1}^{H} e^{-\frac{\alpha_{ij}^k \tau_i^k}{2}\left(A_{ij}^k - \bar{A}_{ij}^k\right)^2}$$

and for the emissions

$$p(B|\beta, \rho) \propto \prod_{i,j=1}^{V,H} e^{-\frac{\beta_{ij}\rho_i}{2}\left(B_{ij} - \bar{B}_{ij}\right)^2}.$$

Here $\bar{A}^k$ and $\bar{B}$ are the preferred values of A^k and B, and α^k and β are the corresponding matrices of hyperparameters.

3.1 Variational Bayes Approximation

The goal of learning is to find the optimal hyperparameters $\hat{\Theta} = \{\alpha, \beta\}$ with respect to the marginal likelihood

$$p(v|\hat{\Theta}) = \int_{z,h,\Theta} p(v|h,z,\Theta)p(h|\Theta)p(z)p(\Theta|\hat{\Theta}).$$

In addition, once optimized, we wish to examine the marginal posterior $p(z^i|v, \hat{\Theta})$ to assess the cluster assignments.

However, computing the marginal likelihood $p(v|\hat{\Theta})$ is computationally intractable, and we resort to a deterministic approximation, a form of 'collapsed' Variational Bayes (VB) procedure[6][8]. We assume that the posterior has an approximate factorization

$$p(h, z, \Theta|v) \approx q(h)q(z)q(\Theta)$$

where we further assume the factorized forms

$$q(z) \equiv \prod_i q(z_i), q(h) \equiv \prod_k q(h^k), q(\Theta) \equiv \prod_k q(\Theta^k).$$

Taking the Kullback-Leibler divergence[9]

$$KL(q(h)q(z)q(\Theta)||p(h,z,\Theta|v))$$

yields a lower bound on $\log p(v|\hat{\Theta})$ given by

$$\begin{aligned}
&\sum_k H_q(h^k) + \sum_i H_q(z^i) + \sum_k H_q(\Theta^k) \\
&+ \sum_{i,t} \left\langle \log p(v_t^i|h_t, z^i, \Theta) \right\rangle_{q(z^i)q(h_t)q(\Theta)} \\
&+ \sum_{t,k} \left\langle \log p(h_t^k|h_{t-1}^k, \Theta^k) \right\rangle_{q(h^k)q(\Theta^k)} \\
&+ \left\langle \log p(z) \right\rangle_{\prod_i q(z^i)} + \sum_k \left\langle \log p(\Theta^k|\hat{\Theta}^k) \right\rangle_{q(\Theta^k)},
\end{aligned}$$

[6] In the 'uncollapsed' procedure, one retains the Dirichlet variable π as part of the joint distribution and introduces an additional factorization $q(z)q(\pi)$. This is seductive since it renders an approximation easy to compute. However, the explicit decoupling of z and π in the approximation makes the method practically too inaccurate[8].

where $H_q(x)$ denotes the entropy of the distribution $q(x)$ and $\langle\phi\rangle_q$ denotes expectation of ϕ with respect to q. VB then proceeds by iteratively maximizing the lower bound with respect to the q distributions for fixed hyperparameters $\hat{\Theta}$ and vice-versa until no further improvement is found. The forms of the resulting updates for q and $\hat{\Theta}$ are sketched in the appendix. Full details are to be found in [1]. At convergence, we may read off approximations to the marginal indicator posteriors $p(z^i|v,\hat{\Theta}) \approx q(z^i)$ to assess which outputs are clustered together.

This results in a general algorithm that can be used to cluster time-series outputs based on their simultaneous dynamical similarity. We will demonstrate an application of this technique in Section 4.

Relation to Previous Work

Variational Bayes has been applied to LGSSMs in a variety of contexts, ranging from acoustics[10] to gene-expression analysis[11]. Particularly in the analysis of short sequences of gene-expression profiles, the use of strong prior information to sparsify the model is crucial in obtaining plausible results[11]. In [12], a procedure for making the implementation of VB to the LGSSM straightforward and numerically stable was discussed. This has the advantage that off-the-shelf inference procedures such as the standard predictor-corrector algorithm[2] can be used directly with the Bayesian LGSSM.

The model discussed in Section 3 extends the Bayesian LGSSM to a mixture model. Whilst the setting of output clustering is special, other works have addressed the more frequently considered case of clustering a set of time-series (see Fig 1) based on mixtures of linear dynamical systems. Works using the simpler autoregressive models include [13], which uses a mixture of ARMA models, with the number of mixtures determined by the BIC criterion. In [14] specially constrained LGSSMs were used to form components in a Dirichlet mixture; the authors used a Bayesian prior to encourage simplicity of each LGSSM. This model is similar to ours – however this is a form of clustering a set of time-series, and not the outputs (see the distinction in Fig 1). Furthermore, sampling in this model is slow, since computing the likelihood of the LGSSM requires $O(T)$ operations, so a single MCMC update is $O(T)$. This model can be seen as an extension of [5], which discusses a sampling approach for a Dirichlet Process Mixture of Factor Analyzers. Sampling in this model is also computationally expensive, so that the method is prohibitive for large datasets and also large observation vectors.

4 Demonstration

As a simple illustration of our output clustering method, in Fig 3 we plot a set of $V = 6$ output sequences which were generated by projecting from two independent linear systems of dimension $H = 6$. We trained our model on this data, assuming $K = 4$ latent linear dynamical systems, each of dimension $H = 8$. We set the parameter biases $\bar{A}$ and $\bar{B}$ to zero in order to encourage the simplest latent transition and emissions to be discovered[15]. The Polya parameter γ in Eq. (4) was set to 10.

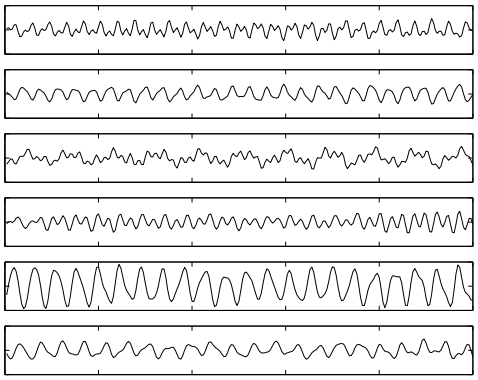

Figure 3. Output clustering. Our model correctly identifies two clusters, assigning the top three output sequences to one cluster and the bottom three to another.

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. In addition, the initial over parameterization of the latent systems was reduced from 8 to 6, as can be seen in Fig 4, where each emission B^i (corresponding to a row in the matrix) has at least two zeros, indicating that the effective dimensionality that contributes to the model is at most 6 for each of the latent systems.

As in all mixture model clustering techniques, some care is required in using sensible initializations. Whilst this is a highly problem specific issue, we generally found that initializing the mean transition matrices to the identity and the transition noise Σ_H to be small helped the model more rapidly learn reasonable latent representations h^k. This initialization renders the model similar to a Mixture of Factor Analyzers[5] in the initial training phase, after which the

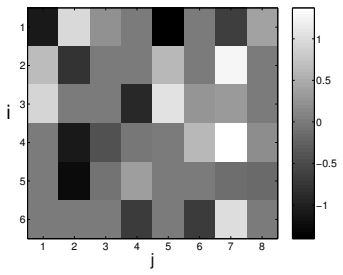

Figure 4. Posterior mean of B_{ij} ($i = 1,\ldots,V$, $j = 1,\ldots,H$). The Bayesian prior has reduced each latent dimension H from 8 to 6.

dynamics of the latent representations emerge.

Throughout our experiments, we found that the marginal likelihood bound is generally a reliable measure of clustering quality, provided that we consider models from the same class. However, for two models with different parameterizations (K, H values), one cannot always rely on their corresponding likelihood bounds to determine which model performed best. There are potentially several reasons why this may be the case, bearing in mind that the bound is only an approximation of the true marginal likelihood. Indeed, a point often overlooked in the literature is that the likelihood approximation, since it will typically approximate a single mode of the posterior, will miss $K!$ modes in the equivalence class defined by permuting the cluster labels. However, this correction alone cannot always account for the sometimes poor quality of the bound as a relative performance criterion across models. It might be that in difficult cases the factorization between parameters and latent states explicit in the Variational Bayes approximation is too crude to accurately capture the mass of the posterior. Similar potential difficulties with the Variational Bayes method have previously been reported[16].

5 Conclusion

We introduced a method to cluster output sequences based on their simultaneous dynamical activity. Our method is based on the assumption that the data clusters have an underlying latent dynamical representation. Whilst this is certainly not always the case, we believe that this may be reasonable in a large variety of applications in the physical sciences. We are currently applying our method to sequence clustering in biosignal analysis, refining the model to encode more specific prior knowledge of the system dynamics.

Full details and code are available from [1].

Acknowledgements

This work was supported in part by the European Community, under the PASCAL Network of Excellence, IST-2002-506778.

A Variational Updates

Variational Bayes iterates between updating the variational approximating distributions $q(h), q(z), q(\Theta)$ and the hyperparameters $\hat{\Theta}$. Below we simply sketch the structure of the these updates. Full details and derivations are given in [1].

Indicator Updates

$$q(z^i = k) \propto e^{\left\langle \log p(z^i = k|z^{\neg i}) \right\rangle_{q(z^{\neg i})}} \times e^{\sum_t \left\langle \log p(v_t^i|h_t^k, \Theta) \right\rangle_{q(h_t^k)q(\Theta)}}.$$

The first term in the exponent $p(z^i = k|z^{\neg i})$ is the probability of $z^i = k$ conditioned on all remaining variables excluding z^i. The average $\left\langle \log p(z^i = k|z^{\neg i}) \right\rangle_{q(z^{\neg i})}$ needs attention since, naively, $p(z^i = k|z^{\neg i})$ 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 [8], as follows. Explicitly,

$$p(z^i = k|z^{\neg i}) = \frac{V_{k,\neg i} + \gamma/K}{V - 1 + \gamma} \equiv f(V_{k,\neg i}), \qquad (6)$$

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

$$M_{k,\neg i} \equiv \sum_{j=1, j\neq i}^{V} q(z^j = k)$$

$$S_{k,\neg i} \equiv \sum_{j=1, j\neq i}^{V} q(z^j = k)(1 - q(z^j = k)).$$

We then approximate $\langle f(V_{k,\neg i}) \rangle$ in Eq. (6) by using a second order Taylor expansion[7]:

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

Latent Posterior $q(h^k)$ Updates

$$q(h^k) \propto e^{\sum_{i,t} q(z^i=k) \left\langle \log p(v_t^i|h_t, z^i=k, \Theta) \right\rangle_{q(\Theta)}} \times e^{\sum_t \left\langle \log p(h_t^k|h_{t-1}^k, \Theta^k) \right\rangle_{q(\Theta^k)}}. \qquad (7)$$

This term is closely related to a standard VB approximation to a Bayesian LGSSM [10, 11, 12]. Clearly the structure of $q(h^k)$ is a pairwise Markov chain, and inference algorithms such as Belief Propagation[11, 10] can be used. However, we take the approach discussed in [12] 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 write terms such as

$$-2q(z^i = k) \left\langle \log p(v_t^i|h_t, z^i = k, \Theta) \right\rangle_{q(\Theta)} = q(z^i = k) \left\langle \left(v_t^i - B_i h_t^k\right)^{\mathsf{T}} \rho_i \left(v_t^i - B_i h_t^k\right) \right\rangle + const.$$

as a decomposition consisting of a 'mean' term

$$q(z^i = k) \left(v_t^i - \langle B_i \rangle h_t^k\right)^{\mathsf{T}} \langle \rho_i \rangle \left(v_t^i - \langle B_i \rangle h_t^k\right) \qquad (8)$$

and 'fluctuation' term

$$\left(h_t^k\right)^{\mathsf{T}} S_{B_i}^k h_t^k,$$

where $S_{B_i}^k$ is the covariance of $q(z^i = k) B_i{}^{\mathsf{T}} \rho_i B_i$. The analytical expression for this covariance is given in [1] and,

[7] 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.

crucially, does not explicitly involve ρ_i. The mean term Eq. (8) represents the contribution of a standard LGSSM with parameters B replaced by their average values and a change to the emission covariance. Similarly, we can apply this decomposition to the transition terms $\left\langle \log p(h_t^k | h_{t-1}^k, \Theta^k) \right\rangle$ in Eq. (7).

The key observation is to consider the extra 'fluctuation' terms as having been generated from fictitious zero-valued observations $(0 - (\sum_i S_{B_i}^k)^{\frac{1}{2}} h_t^k)^\mathsf{T} (0 - (\sum_i S_{B_i}^k)^{\frac{1}{2}} h_t^k)$. Hence, by augmenting the LGSSM with fictitious outputs and adjusting the emissions, we can reformulate Eq. (7) as the posterior of a standard LGSSM, for which any of the standard algorithms in the literature[2] may be applied to perform inference of $q(h_t^k)$ and related quantities. A slight modification of the standard algorithm produces a more efficient procedure obviating the need to introduce fictitious outputs[12].

LGSSM Parameter Updates

The parameter updates are straightforward. For example, the prior on $p(A^k, \tau^k | \hat{\alpha}^k)$ is a multivariate Normal-Gamma distribution which gives rise to a Normal-Gamma posterior approximation $q(A^k, \tau^k)$ of the form:

$$q(\Theta^k) \propto p(\Theta^k | \hat{\Theta}^k) e^{\sum_t \left\langle \log p(h_t^k | h_{t-1}^k, \Theta^k) \right\rangle_{q(h^k)}}.$$

A similar update occurs for B and ρ.

Hyperparameter Updates

Assuming a hyperparameter β_{ij} for each element of the matrix B_{ij} to bias it towards a desired value $\bar{B}_{ij}$, and taking the derivative of the bound with respect to β_{ij}, we obtain the fixed point condition

$$\frac{1}{\beta_{ij}} = \left\langle \rho_i (B_{ij} - \bar{B}_{ij})^2 \right\rangle_{q(B,\rho)}.$$

The averages are analytically available from [1]. Similarly, we have a bias for each A_{ij}^k towards a desired $\bar{A}_{ij}^k$. The fixed point condition is then given by

$$\frac{1}{\alpha_{ij}^k} = \left\langle \tau_i^k (A_{ij}^k - \bar{A}_{ij}^k)^2 \right\rangle_{q(A^k, \tau^k)}.$$

References

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

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

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

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

[5] D. Görür. *Nonparametric Bayesian Discrete Latent Variable Models for Unsupervised Learning*. Phd thesis, Technical University of Berlin, 2007.

[6] 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.

[7] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

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

[9] M. T. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 1991.

[10] A. T. Cemgil and S. J. Godsill. Efficient variational inference for the dynamic harmonic model. In *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, 2005.

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

[12] 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.

[13] 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.

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

[15] D. J. C. MacKay. Probable networks and plausisble predictions – a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, 6(3), 1995.

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