

---

# Bayesian Nonparametric Estimation of Switching Linear Dynamic System

---

**Will Townes**  
Department of Biostatistics  
Harvard University  
Boston, MA 02115  
ftownes@g.harvard.edu

**(Jeremiah) Zhe Liu**  
Department of Biostatistics  
Harvard University  
Boston, MA 02115  
zh1112@mail.harvard.edu

## Contents

<b>1</b>	<b>Problem Formulation</b>	<b>1</b>
<b>2</b>	<b>Methods</b>	<b>2</b>
2.1	Sampling Dynamical Parameters $\theta$ using Matrix Normal Prior . . . . .	2
2.2	Sampling for Hidden Variables $(\mathbf{x}, \mathbf{z})$ using Message Passing . . . . .	2
2.3	Sampling Transition Probabilities $\Pi, \beta$ using HDP Prior . . . . .	2
<b>3</b>	<b>Results</b>	<b>4</b>
3.1	HDP-HMM . . . . .	4
3.2	SLDS . . . . .	4
3.3	HDP-SLDS . . . . .	7
<b>4</b>	<b>Conclusion and Future Direction</b>	<b>7</b>

Appendix8

## 1 Problem Formulation

Dynamical systems consist of objects changing over time. These systems are often modeled deterministically using differential equations. Given noisy observations from such a system (a time series), one may wish to recover underlying latent dynamics. Every high order ordinary linear differential equation (ODE) can be represented as a system of first order linear ODEs. The popular linear dynamical system (LDS) model is a probabilistic analog to discretized systems of first order linear ODEs. We consider the estimation of switching linear dynamical systems (SLDS), which allow changes between multiple LDS regimes, and attempt to replicate results from Fox [2009]. SLDS is a state-space model in which at time  $t$ , an agent's observed state  $y_t \in \mathbb{R}^{d_y}$  is a noisy and censored version of the underlying state  $x_t \in \mathbb{R}^{d_x}$ , whose movement is governed by an time-varying linear dynamic system. The underlying state trajectory is subject to its own noise as well. Namely:

$$x_t|x_{t-1} \sim \mathcal{N}(A_t x_{t-1} + B_t, \Sigma_t) \quad (1)$$

$$y_t|x_t \sim \mathcal{N}(C x_t, R) \quad (2)$$

where  $C_{d_y, d_x} = [\mathbf{I}_{d_y} \ \mathbf{0}_{d_x-d_y}]$  is a fixed "censoring matrix" that selects the first  $d_y$  elements of  $x_t$ ,  $\Sigma_t$  is the transition model noise matrix, and  $R$  is the observation noise matrix. Further, SLDS assumes the set of time-specific dynamics  $\theta_t = \{A_t, B_t, \Sigma_t\}$  arise from a countable set  $\Theta = \mathcal{A} \times \mathcal{B} \times \{\Sigma_t\}$  indexed by  $\mathcal{Z}$ . Finally, if denote  $z_t \in \mathcal{Z}$  the index of  $\theta_t$ , SLDS assumes  $z_t$  follows a Markov process with transition matrix  $\Pi_{|\Theta| \times |\Theta|} = [\pi_1, \dots, \pi_z, \dots]^T$ , such that:

$$z_t|z_{t-1} \sim \pi_{z_{t-1}}$$

Despite the Markovian assumption, SLDS is capable of modeling a diverse collection of phenomenon with complex temporal dependencies from maneuvering aircraft trajectory to financial time-series. For example, in order to use SLDS to analyze fighter pilot's combat style, we may denote  $y_t \in \mathbb{R}^3$  the observed position of the maneuvering fighter aircraft, which comes from a latent  $x_t \in \mathbb{R}^9$  comprised of position, speed and momentum in the 3D space. We can learn how pilot executes different maneuvers by estimating  $\Theta$  (a countably finite set) containing dynamics that describe a range of offensive and defensive maneuvers (e.g. "barrel roll attack", "lag roll", "break", "last-ditch", etc). We can also learn the pilots' habit of "maneuver combo's" by estimating  $\Pi_{|\Theta| \times |\Theta|}$  the transition matrix describing how pilots moves from one maneuver to the other.

However, the flexible nature of SLDS caused also considerable difficulty in estimation, in particular the dimension of transition matrix  $\Pi$  is  $O(|\Theta|^2)$  and can theoretically grow to infinity. To this regard, Bayesian nonparametric methods, in particular Hierarchical Dirichlet Process (HDP), achieves efficient inference and sparse solution to  $\Pi$  through global shrinkage on each state-specific transition distributions.

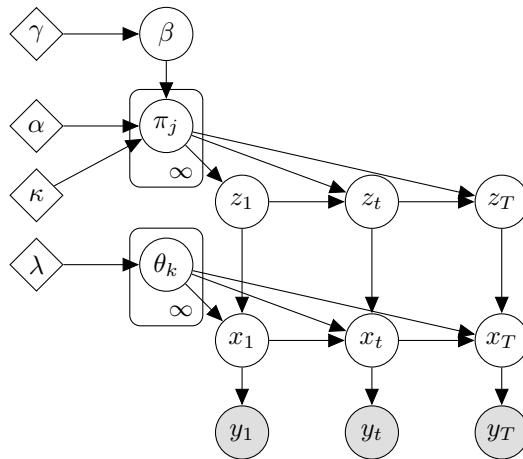


Figure 1: Graphical Model for Switching Linear Dynamics System

In the rest of this report, we discuss how to properly adapt HDP into the estimation of  $\Pi$  matrix, and further how to integrate this method into the entire estimation process for the SLDS under the Bayesian framework by choosing appropriate priors. We outline a Gibbs Sampling procedure for joint inference on  $\Theta$  and  $\Pi$ . Finally, we show results for each compartment of the overall model, as well as preliminary results for an integrated model.

## 2 Methods

As shown in the graphical model, the full conditional for the hidden continuous states  $x_{1:T}$  is a time-varying (switching) linear dynamical system (SLDS), while the full conditional for the hidden categorical modes  $z_{1:T}$  is a hidden markov model (HMM) for finite  $\mathbf{\Pi}$ . Due to our goal of sampling dynamic parameters  $\theta$  on a state-space model, we have to battle several technical difficulties such as (1) choosing proper prior for  $\theta$ , (2) identify efficient algorithm for computing  $p(\mathbf{z}|\mathbf{y}, \Theta, \mathbf{\Pi})$  on Markov time series, and (3) explicit sampling for infinite dimensional  $\pi_j$ . These problems are not encountered in classical HDP [Teh et al., 2006] due to the lack of time-series dependency between observations, and the fact that  $\theta$  are not of interest in classical HDP and therefore  $\theta, \pi_j$  can be integrated out in their Gibbs sampler. In this section, we discuss how we have overcome the three aforementioned difficulties and constructed our Gibbs sampler for efficient and regularized joint inference of  $\Theta$  and  $\mathbf{\Pi}$ . We implemented the methods in R, code is available at <https://github.com/willtownes/mit6882>.

### 2.1 Sampling Dynamical Parameters $\theta$ using Matrix Normal Prior

First note that while  $|\Theta|$  may be infinite in the probability model, for a finite dataset, at any given Gibbs iteration only a finite number  $K$  of modes are instantiated. Hence, conditional on  $z_{1:T}$ , we can assume  $\mathcal{Z} = \{1 \dots K\}$ .  $z_t$  acts as a selector for the dynamical parameters used at time  $t$ . Hence, the  $k^{th}$  unique set of dynamical parameters  $\theta^{(k)} = (A^{(k)}, B^{(k)}, \Sigma^{(k)})$  appears in the likelihood only for indices  $\{t, t-1 : z_t = k\}$ . We collect all the column vectors  $x_{t:z_t=k}$  into the matrix  $\Psi^{(k)}$  and similarly form  $\bar{\Psi}^{(k)}$  from  $x_{\{(t-1):z_t=k\}}$ . Then model (1) implies that

$$\Psi^{(k)} \sim \mathcal{MN}(A^{(k)}\bar{\Psi}^{(k)} + B^{(k)}\mathbf{1}', \Sigma^{(k)}, \mathbb{I})$$

This is a multivariate version of linear regression, where  $\mathcal{MN}$  denotes the *matrix normal* distribution<sup>1</sup>. The (conditionally) conjugate priors are given by  $A^{(k)} \sim \mathcal{MN}(M_A, \Sigma^{(k)}, K_A^{-1})$ ,  $B^{(k)} \sim \mathcal{N}(M_B, \Sigma^{(k)}/\kappa_0)$ , and  $\Sigma^{(k)} \sim \mathcal{IW}(\nu_0, \Delta_0)$ . These<sup>2</sup> lead to closed form full conditionals, see Appendix F for derivations. Hyperparameters can be set to regularize estimates toward stable dynamics.

### 2.2 Sampling for Hidden Variables $(\mathbf{x}, \mathbf{z})$ using Message Passing

Conditional on other unknowns, sampling the hidden states  $x_{1:T}$  can be performed simultaneously rather than sequentially. This leads to faster mixing by reducing the dimensionality of the parameter space. The full conditional distribution can be obtained in closed form by running a Kalman smoother, which uses forward and backward message passing over the sequential observations. However, since we only want to sample from this distribution, we can equivalently pass messages in one direction and recursively sample in the other direction. This saves computation time since message passing requires marginalizing over each  $x_t$ , which involves inverting  $d_x \times d_x$  matrices. We refer to this procedure as a ‘‘Kalman Sampler’’. Full derivations are provided in Appendix E.

Similarly, hidden categories  $\mathbf{z}$  can also be sampled jointly in the forward fashion, due to the Markovian decomposition of  $p(\mathbf{z}|\mathbf{y}) = \left[ \prod_{t=2}^T p(z_t|z_{t-1}, \mathbf{y}) \right] * p(z_1|\mathbf{y})$ .

### 2.3 Sampling Transition Probabilities $\mathbf{\Pi}, \beta$ using HDP Prior

Finally, we perform estimation of  $\pi_k$ , the rows of the transition matrix  $\mathbf{\Pi}$ . We notice that the high-dimensional nature of  $\mathbf{\Pi}$  ( $\dim(\mathbf{\Pi}) = O(|\mathcal{Z}|^2)$ ) calls for regularization not only on the individual values of  $\pi_{kj}$ ’s, but more importantly, also on the global dimension of  $\mathbf{\Pi}$  (equivalently, the cardinality of  $\mathcal{Z}$ ) such that the states corresponding to the same dynamics are gathered into the same cluster, which is crucial for stable estimation of  $\Theta_k$ ’s.

Operationally, we regularize individual  $\pi_{jk}$ ’s using a global base measure  $\beta$  by assuming  $\pi_k$ ’s are conditionally exchangeable given a HDP prior, and we further regularize the dimension of  $\mathbf{\Pi}$  using an degree  $K$  weak limit approximation to the Dirichlet process:

$$\text{Dirichlet}\left(\frac{\alpha}{K}, \dots, \frac{\alpha}{K}\right) \rightarrow \text{GEM}(\alpha)$$

and effectively upperbounding the number of possible clusters by  $K$ .

<sup>1</sup>Let  $Z$  be a matrix with iid entries  $z_{ij} \sim \mathcal{N}(0, 1)$  and  $X = M + AXB$ , then  $X \sim \mathcal{MN}(M, U, V)$ .  $M$  is the mean matrix,  $U = AA'$  is the row covariance parameter and  $V = B'B$  is the column covariance parameter. See Appendix F for more details.

<sup>2</sup> $\mathcal{IW}$ =Inverse Wishart

Furthermore, the excessive flexibility of HDP makes it easy to overfit data and produce hidden states estimates  $\{z_t\}_{t=1}^T$  implying unrealistically fast state switching (e.g. a fight jet switching between offensive diving and defensive spiral every other millisecond). We tackle this issue by augmenting the self-transition probability for all state  $k \in \mathcal{Z}$  by a non-negative value  $\kappa$ , i.e. using the sticky HDP prior:

$$\pi_{kj} \stackrel{iid}{\sim} DP(\alpha + \kappa, \frac{\alpha\beta + \kappa\delta_k(j)}{\alpha + \kappa}) + \kappa\delta_k(j) \quad \text{or equivalently}$$

$$E(\pi_{kj}) = \begin{cases} \beta_j & j \neq k \\ \beta_k + \kappa & j = k \end{cases}$$

---

**Algorithm 1** MKL PFBS algorithm

---

```

1: procedure rls_dual_mkl_pfbs(K, y, ( $\tau, \mu, \sigma$ ))
2:    $\alpha^0 = \mathbf{0}$ 
3:   for  $p = 1$  to MAX_ITER do
4:      $\alpha_0^p = (1 - \frac{\mu}{\sigma})\alpha^{p-1} - \frac{1}{\sigma n}(\mathbf{K}\alpha^{p-1} - \mathbf{y})$ 
5:      $\alpha^p = \mathbf{S}_{\frac{\tau}{\sigma}}(K, \alpha_0^p)$ 
6:   end for
7:   return  $f^{\text{MAX\_ITER}} = (\alpha^{\text{MAX\_ITER}})^T \mathbf{k}$ 
8: end procedure

```

---

### 3 Results

#### 3.1 HDP-HMM

#### 3.2 SLDS

In this section we assume known observation noise  $R$  and hidden mode sequence  $z_{1:T}$  and focus on inference of  $\Theta$  and  $x_{1:T}$ . We simulated several datasets corresponding to physical systems, including projectile motion, linear motion, and harmonic motion (mass on a spring). Note that since any high order linear differential equation can be reformulated as a system of first order differential equations, we can simulate highly nonlinear trajectories with a sufficiently high latent state space. For example, projectile motion and harmonic motion have second order dynamics. For two-dimensional projectile motion, we used a latent space of dimension 4. For one dimensional harmonic motion, we used a latent space of dimension two. Details of converting from differential equations to LDS parameters are given in Appendix D. Projectile and linear motion are non-stationary, while harmonic motion is stationary. Fox [2009] considered only stationary motion. To test our implementations, we considered the following scenarios:

1. Two dimensional, two mode SLDS with projectile motion followed by linear motion
2. One dimensional, three mode SLDS with three different kinds of damped harmonic motion.

For each scenario, we first used the true dynamical parameters and mode sequence to learn the posterior over the hidden states  $x_{1:T}$  using the Kalman sampler. Then, we tried learning both the hidden states and the dynamical parameters by alternating between Kalman sampling and the multivariate linear regression posterior described above.

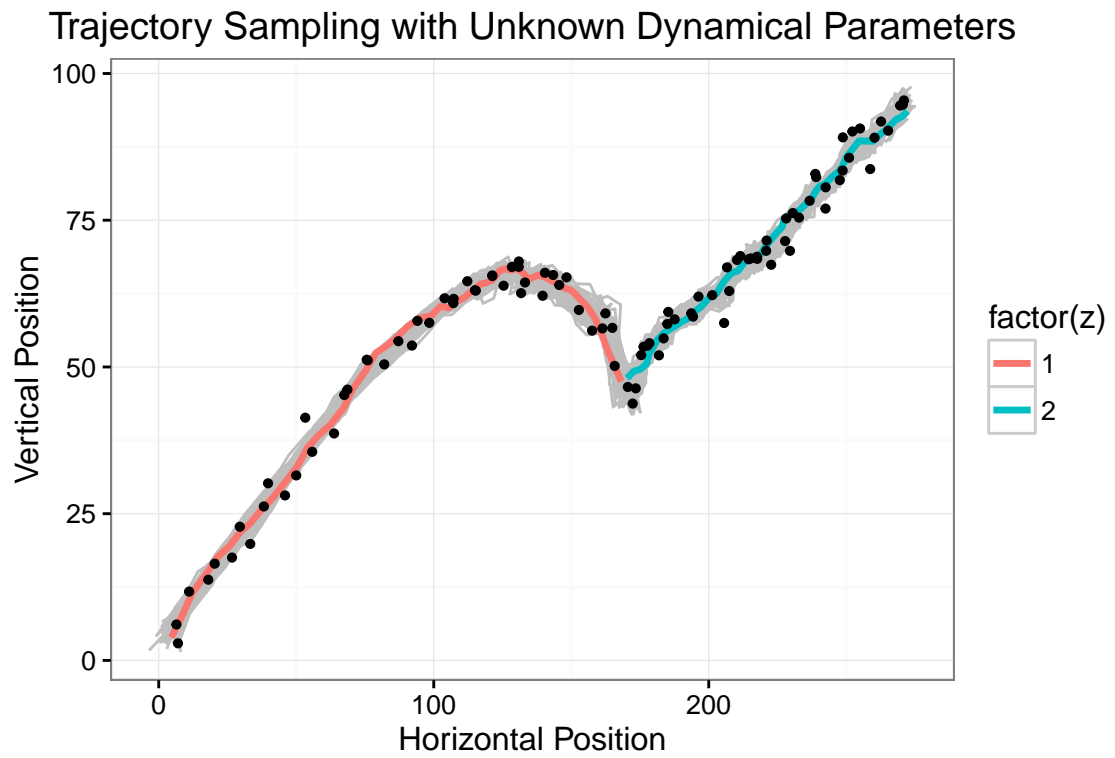
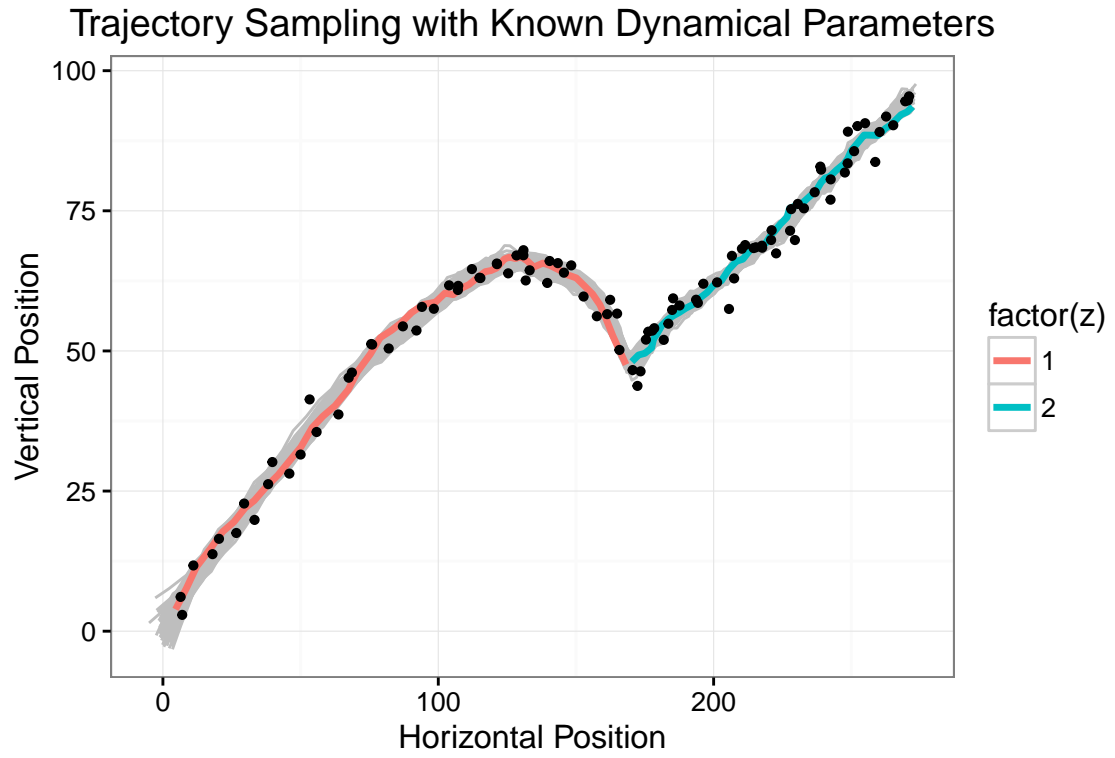


Figure 2: Dark points=observed positions. Coloured lines= true latent trajectory. Gray lines= samples from posterior. Colours=dynamical modes

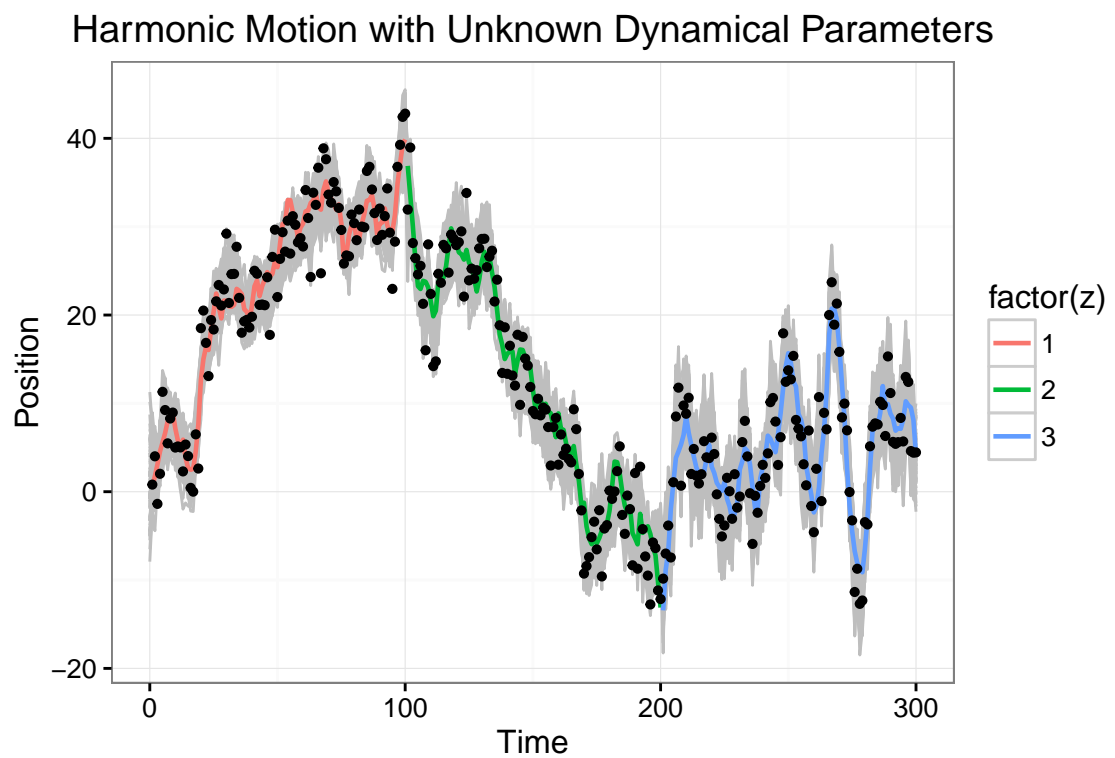
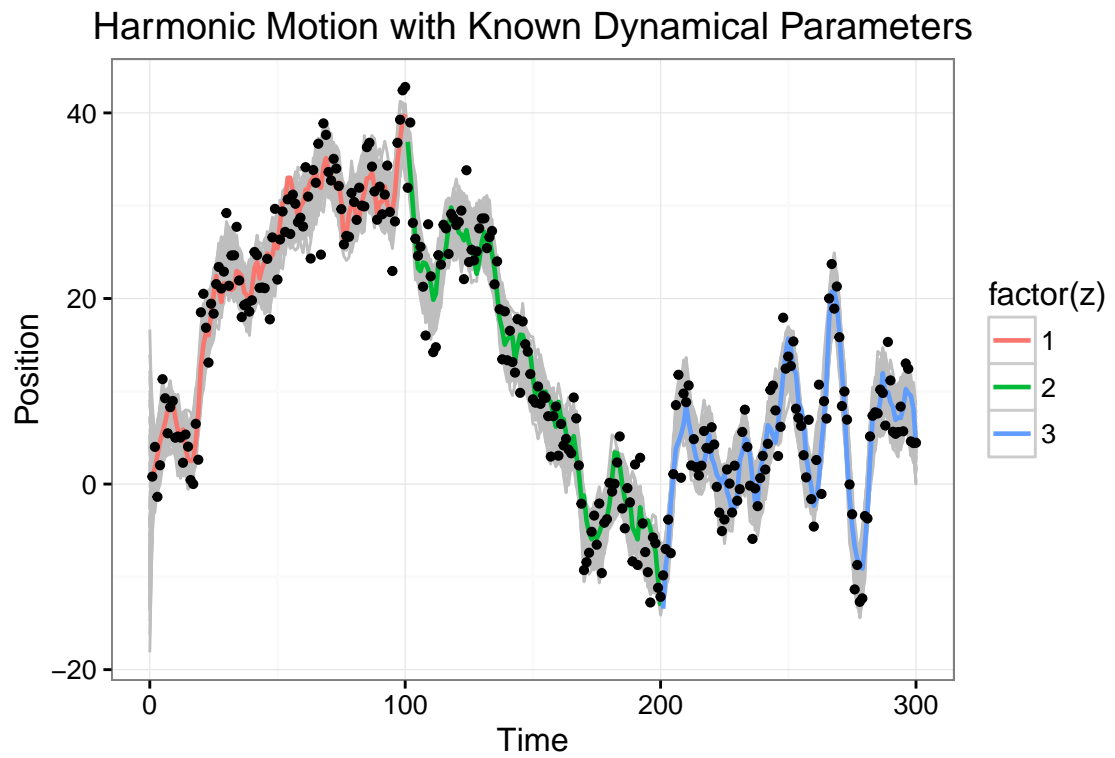


Figure 3: Dark points=observed positions. Coloured lines= true latent trajectory. Gray lines= samples from posterior. Colours=dynamical modes

Table 1: Posterior Mean Accuracy for Scenario 1: Projectile/Linear Motion

mode	param	MSE (vs truth)	MSE (vs prior)
1	A	0.06	0.06
1	B	2.31	1.56
1	$\Sigma$	9.69	9.18
2	A	0.12	0.12
2	B	0.93	0.04
2	$\Sigma$	9.55	0.01

Table 2: Posterior Mean Accuracy for Scenario 2: Harmonic Motion

mode	param	MSE (vs truth)	MSE (vs prior)
1	A	1.13	0.31
1	B	0.16	0.12
1	$\Sigma$	1.26	1.20
2	A	0.12	0.72
2	B	0.05	0.02
2	$\Sigma$	0.10	2.08
3	A	18.00	14.00
3	B	3.53	3.24
3	$\Sigma$	0.16	2.54

As shown in the figures, we found that in many cases the posterior over trajectories inferred by the sampler with unknown dynamical parameters was similar to what was inferred by specifying the true parameters. Individual trajectories sampled with unknown dynamical parameters tended to be noisier, especially in regions of sharp change. Interestingly, switch points between modes (indicated by colour) did not have greater uncertainty than sharp changes within a mode.

We found that the choice of hyperparameters was crucial to stable sampling. Specifically, using weak priors caused samplers to diverge or run into numerical problems, particularly in the case of nonstationary dynamics and/or high dimensional latent states (scenario 1). To remedy this, we used a heuristic of first estimating the latent trajectory roughly by computing local regressions (LOESS) over each observed dimension versus time, then setting the hyperprior scale matrix for  $\Sigma$  based on the empirical covariance of the observations versus the LOESS prediction at that point. If  $y_t \in \mathbb{R}^n$  and  $x_t \in \mathbb{R}^d$ , this procedure accounts for the first  $n$  dimensions of  $x_t$ . The remaining  $d - n$  elements of the scale matrix were set to a diagonal with determinant equal to the first  $n$  block. This is a generalization of the procedure given by Fox [2009], which can be considered a special case where the LOESS bandwidth is set very wide. We were able to get away with weaker priors for the stationary dynamical system (scenario 2).

We also compared our estimates for the dynamical parameters to their true values using mean squared error as shown in tables 1 and 2. Interestingly, in most cases the MSE between the posterior means (empirically computed from samples) and the priors was less than MSE between posterior means and true values. This reinforces the fact that the prior held a great influence over the results in these models, especially under the nonstationary scenario 1. Furthermore, it is clear that estimation of mean parameters  $A$  and  $B$  was more successful than covariance parameters  $\Sigma$ .

### 3.3 HDP-SLDS

## 4 Conclusion and Future Direction

Based on our results that hyperparameter settings are crucial to numerical stability of SLDS inference, we suspect that using less general models such as autoregressive structures, or employing sparsity promoting priors such as automatic relevance determination (ARD) could have advantages over the conjugate prior we employed. Indeed Fox [2009] considered such alternative priors and found many benefits. This makes sense intuitively, in that the matrix normal/ inverse wishart prior induces a larger parameter space which is only partially alleviated by strongly regularizing hyperparameters. Furthermore, conjugacy of a prior, while mathematically convenient, does not imply appropriateness to the model at hand. Finally, we note that in modeling dynamical systems, it is not unrealistic to assume strong prior information may be available, for example if the trajectory is a physical system like we considered. Future studies should consider how to design priors that reflect this kind of information in the model.



## A Hierarchical Dirichlet Process

### A.1 Model

Classic view:

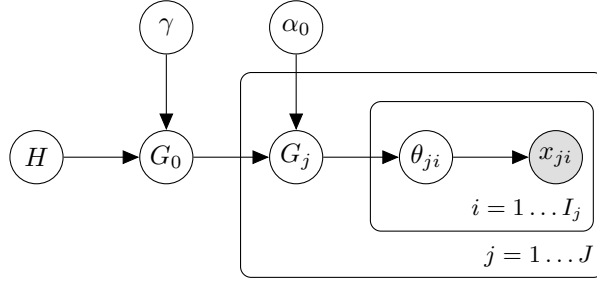
$$\begin{aligned} G_0 | \gamma, H &\sim DP(\gamma, H) \\ G_j | \alpha_0, G_0 &\sim DP(\alpha_0, G_0) \\ \theta_{ji} | G_j &\sim G_j \\ x_{ji} | \theta_{ji} &\sim F(\theta_{ji}) \end{aligned}$$

where  $P \sim DP(\alpha, G)$  adopts the stick breaking representation w.p. 1:

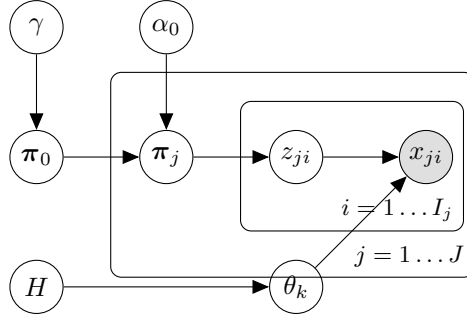
$$P = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k} \quad \text{where:} \quad \pi_k \sim GEM(\alpha), \quad \phi_k \sim G$$

Alternatively, one may describe the generative processes of  $\pi_k$  and  $\theta_k$  separately as:

$$\begin{aligned} \pi_0 | \gamma &\sim GEM(\gamma) & \theta_k | H &\sim H \\ \pi_j | \alpha_0, \pi_0 &\sim DP(\alpha_0, \pi_0) \\ z_{ji} | \pi_j &\sim \pi_j \\ x_{ji} | z_{ji}, (\theta_k)_{k=1}^{\infty} &\sim F(\theta_{z_{ji}}) \end{aligned}$$



(a) Hierarchical Dirichlet Process



(b) Hierarchical Dirichlet Process

Figure 4: Hierarchical Dirichlet Process

### A.2 Inference

Assuming conjugacy between  $H$  and  $F$  and holding  $(\gamma, \alpha_0)$  fixed,

we now describe a simplified Gibbs approach to sample parameters  $(z_{ji}, m_{jk}, \pi_0)$  from the Chinese Restaurant Franchise (see Appendix C) representation of the posterior, where the parameter  $z_{ji}$  are referred to respectively as customer-specific dish assignment,  $m_{jk}$  as dish-specific table count, and  $\pi_0$  as global dish distribution. This particular method is referred to as "direct assignment" in Teh et al. [2006] since it circumvented the issue of bookkeeping for every  $t_{ij}$  (customer-specific table assignment) and  $k_{jt}$  (table-specific dish assignment) variables.

In each Gibbs iteration, denote  $f_k^{-x_{ji}}(x_{ji}) = \frac{\int f(\mathbf{x}|\theta_k)h(\theta_k)d\theta_k}{\int f(\mathbf{x}_{-(ji)}|\theta_k)h(\theta_k)d\theta_k}$  the conditional distribution  $x_{ji}|\mathbf{x}_{-(ji)}$  under  $\theta = \theta_k$ , and assume there are currently  $K$  dishes and  $T$  tables, we sample  $(z_{ji}, m_{jk}, \pi_0)$  iteratively as:

1. Sample  $z_{ji} = k|\mathbf{z}_{-(ji)}, \mathbf{m}, \pi_0$  from the distribution:

$$z_{ji} = k|\mathbf{z}_{-(ji)}, \mathbf{m}, \pi_0 \propto \begin{cases} f_k^{-x_{ji}}(x_{ji}) * (n_{jk}^{-(ji)} + \alpha_0 \pi_{0,k}) & k \leq K \\ f_{K+1}^{-x_{ji}}(x_{ji}) * \alpha_0 \pi_{0,u} & k = K + 1 \end{cases}$$

2. Sample  $m_{jk} = m|\mathbf{z}, \mathbf{m}_{-(jk)}, \pi_0$ , by setting  $m_{jk} = \sum_i I(t_{ji} = t_{new}|k_{jt_{new}} = k)$ , we can sample  $t_{ji}$  from:

$$t_{ji} = t|k_{jt} = k, \mathbf{t}_{-(ji)}, \pi_0 \propto \begin{cases} n_{jt}^{-(ji)} & t \leq T \\ \alpha_0 \pi_{0,k} & t = T + 1 \end{cases}$$

and as in Fox [2009], sample  $I(t_{ji} = t_{new}|k_{jt_{new}} = k)$  directly from:

$$\text{Bern}\left(\frac{\alpha_0 \pi_{0,k}}{n_{jk} + \alpha_0 \pi_{0,k}}\right)$$

3. Sample  $\pi_0$  from distribution:

$$\pi_0 \sim \text{Dir}(m_1, \dots, m_K, \gamma)$$

### A.3 Application: Clustering Hierarchical Gaussian Data

Consider mixture of Gaussian data  $\mathbf{x} = \{\mathbf{x}_1, \dots, \mathbf{x}_K\}$  with  $\mathbf{x}_k \stackrel{iid}{\sim} MVN(\boldsymbol{\theta}_{k,2 \times 1}, \mathbf{I}_{2 \times 2})$  with unknown mean  $\boldsymbol{\theta}$ . Assuming diffused Gaussian prior  $\boldsymbol{\theta} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$ , the form of likelihood  $F$  and base measure  $H$  are:

$$\begin{aligned} f(x_{ji}|\boldsymbol{\theta}_k) &\propto \exp\left(-\frac{1}{2\sigma^2}(x_{ji} - \boldsymbol{\theta}_k)^T(x_{ji} - \boldsymbol{\theta}_k)\right) \\ h(\boldsymbol{\theta}_k) &\propto \exp\left(-\frac{1}{2\sigma_0^2}\boldsymbol{\theta}_k^T\boldsymbol{\theta}_k\right) \end{aligned}$$

Then  $f_k^{-x_{ji}}(x_{ji})$  should be:

$$f_k^{-x_{ji}}(x_{ji}) \sim N\left(\frac{n_k^{-(ji)}\sigma_0^2}{n_k^{-(ji)}\sigma_0^2 + \sigma^2}\bar{\mathbf{x}}_k^{-(ji)}, \left(1 + \frac{\sigma_0^2}{n_k^{-(ji)}\sigma_0^2 + \sigma^2}\right)\mathbf{I}\right)$$

## B HDP for Hidden Markov Model

### B.1 Hidden Markov Model

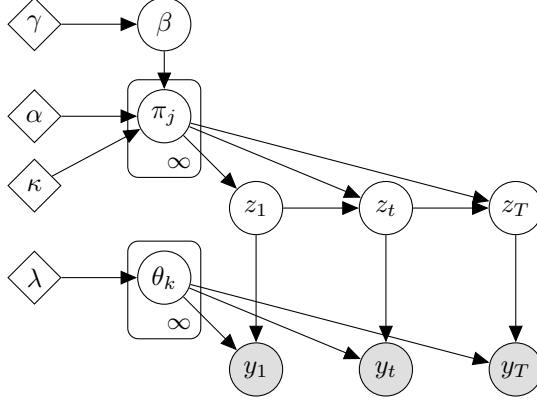


Figure 5: Hidden Markov Model

$$\begin{aligned}\beta|\gamma &\sim GEM(\gamma) \\ \pi_j|\beta, \alpha &\sim DP(\alpha, \beta) \\ \theta_k|\pi, \lambda &\sim H(\lambda)\end{aligned}$$

$$\begin{aligned}z_t|z_{t-1}, \pi &\sim \pi_{z_{t-1}} \\ y_t|z_t, \theta &\sim F(\theta_{z_t})\end{aligned}$$

$$f_k(y_t) = p(y_t|\theta_{z_t})p(z_t|z_{t-1})$$

### B.2 Sticky HDP

Though flexible, the fact that HDP-HMM is deploying  $\pi_k \sim DP(\alpha, \beta)$  leads to:

1. large posterior probability for unrealistically transition dynamics
2. once instantiated, the unrealistically transition dynamics will be reinforced by CRF

Sticky HDP address above issues by encouraging self-transition. More specifically, the base measure for  $\pi_k$  is augmented *a priori* from  $\beta$  to:

$$\pi_j \sim DP(\alpha + \kappa, \frac{\alpha\beta + \kappa\delta_j}{\alpha + \kappa})$$

### B.3 Inference

Inference for HMM with Sticky HDP prior follows the sticky extension of CRF. For a observation  $y_t$  at time  $t$ , "restaurant" corresponds to the state  $z_t$  that  $y_t$  is at, and dishes at restaurant  $z_t$  indicates the potential states that  $y_{t+1}$  can transit to. To improve mixing rate of state sequence  $\mathbf{z}$ , we deploy the blocked sampler which uses a weak limit approximation of the infinite-dimension DP prior. More specifically, we assume there are  $L$  states, and  $\beta$  and  $\pi$  follows:

$$\begin{aligned}\beta|\gamma &\sim Dir(\frac{\gamma}{L}, \dots, \frac{\gamma}{L}) \\ \pi_j|\alpha, \beta, \kappa &\sim Dir(\alpha\beta_1, \dots, \alpha\beta_j + \kappa, \dots, \alpha\beta_L)\end{aligned}$$

Define  $\theta_k$  as emission parameter for state  $k$ , we sample  $(\mathbf{z}, \mathbf{m}, \pi_0, \theta)$  as follows:

1. Sample  $z_t$  from the distribution:

$$z_t | \mathbf{z}_{-(ji)}, \mathbf{m}, \boldsymbol{\pi}_0, \boldsymbol{\theta} \sim f(z_t = k | \mathbf{y}, \mathbf{m}, \boldsymbol{\pi}_0, \boldsymbol{\theta})$$

where  $f(z_t = k | \mathbf{y})$  is calculated using the forward-backward message passing algorithm in B.3.1).

2. Sample  $m_{jk}$  through override correction:

- (a) Sample  $m'_{jk} = \sum_i I(t_{ji} = t_{new} | k_{jt_{new}} = k)$ , where:

$$I(t_{ji} = t_{new} | k_{jt_{new}} = k) \sim \text{Bern}\left(\frac{\alpha\pi_{0,k} + \kappa\delta_j(k)}{n_{jk} + \alpha\pi_{0,k} + \kappa\delta_j(k)}\right)$$

- (b) Sample override variable:

$$w_j \sim \text{Binom}\left(m'_{jj}, \frac{\kappa}{\kappa + \alpha\pi_{0,j}}\right)$$

- (c) Finally calculate  $m_{jk}$  as:

$$m_{jk} = \begin{cases} m'_{ij} & j \neq k \\ m'_{jj} - w_j & j = k \end{cases}$$

3. Sample  $\boldsymbol{\pi}_0$  from distribution:

$$\boldsymbol{\pi}_0 \sim \text{Dir}\left(\frac{\gamma}{L} + m_1, \dots, \frac{\gamma}{L} + m_K\right)$$

4. Sample  $\boldsymbol{\theta}$  from distribution:

$$\boldsymbol{\theta} \sim p(\boldsymbol{\theta} | \lambda, \mathbf{y})$$

### B.3.1 Forward-backward Message Passing

The forward-backward algorithm provide an efficient method for computing node marginals  $p(y_t)$ . Define:

$$\text{Backward Message : } \beta_t(z_t) = p(\mathbf{y}_{T>t} | z_t)$$

$$\text{Forward Message : } \alpha_t(z_t) = p(\mathbf{y}_{T\leq t}, z_t)$$

$$\text{Joint Message : } \alpha_t(z_t)\beta_t(z_t) = p(\mathbf{y}, z_t)$$

which can be alternatively defined using message  $m_{t_1, t_2}$

$$\text{Backward Message : } \beta_t(z_t) = p(\mathbf{y}_{T>t} | z_t) = m_{t+1, t}(z_t)$$

$$\text{Forward Message : } \alpha_t(z_t) = p(y_t | z_t)p(\mathbf{y}_{T< t}, z_t) = p(y_t | z_t)m_{t-1, t}(z_t)$$

.

These two types of messages can be computed  $\beta_t$  backward and  $\alpha_t$  forward in time as:

$$\beta_{t-1} = \sum_{z_t} p(y_t | z_t) p(z_t | z_{t-1}) \beta_t(z_t) \quad \text{with} \quad \beta_T(z_T) = 1$$

$$\alpha_{t+1} = \sum_{z_t} p(y_{t+1} | z_{t+1}) p(z_{t+1} | z_t) \alpha_t(z_t) \quad \text{with} \quad \alpha_1(z_1) = p(y_1, z_1) = p(y_1 | z_1) \pi^0(z_1)$$

Using the forward and backward messages, we can compute state assignment posterior as:

$$p(z_t | \mathbf{y}) = \frac{p(z_t, \mathbf{y})}{\sum_{z_t} p(z_t, \mathbf{y})} = \frac{\alpha_t(z_t)\beta_t(z_t) = p(\mathbf{y}, z_t)}{\sum_{z_t} \alpha_t(z_t)\beta_t(z_t) = p(\mathbf{y}, z_t)}$$

## C Chinese Restaurant Franchise

A hierarchical analogy of Chinese Restaurant Process, the Chinese Restaurant Franchise offers a convenient scheme to sample from the posterior of cluster-specific  $\theta$ 's in HDP. This process draw below analogy:

- $H$  as the dish distribution for all possible dishes in the world, with the types of possible dishes being  $(\theta_k)_{k=1}^{\infty}$ .
- $G_0 \sim DP(\gamma, H)$  as the dish distribution for the franchise
- $G_j \sim DP(\alpha_0, G_0)$  as the dish distribution for restaurant  $j$  in the franchise
- $\psi_{jt} \sim G_0$  as the dish served at table  $t$  in restaurant  $j$ .  
 $k_{jt} \sim \pi_0$  as the index of dish choice for this table.
- $\theta_{ji} \sim G_j$  as the dish will be enjoyed by customer  $i$  in restaurant  $j$ .  
 $t_{ji} \sim \pi_j$  as the index of table choice for this customer.

Integrating over  $G_j$ , the sampling scheme for subject-specific dish  $\theta_{ji} \sim G_j$  is:

$$\theta_{ji} | \theta_{j(-i)}, \alpha_0, G_0 \sim \sum_{k=1}^K \frac{n_{jk}}{\alpha_0 + n_{j..}} \delta_{\psi_{jk}} + \frac{\gamma}{n_{j..} + \gamma} G_0$$

Integrating over  $G_0$ , the sampling scheme for table-specific dish  $\psi_{jt} \sim G_0$  is:

$$\psi_{jk} | \Psi_{j(-k)}, \gamma, H \sim \sum_{k=1}^K \frac{m_{.k}}{\gamma + m_{..}} \delta_{\theta_k} + \frac{\gamma}{m_{..} + \gamma} H$$

## D Converting Differential Equations into Linear Dynamical Systems

### D.0.1 Linear and Projectile Motion

The linear dynamical system model is given by the following recursive equations:

$$\begin{aligned} x_t | x_{t-1} &\sim \mathcal{N}(A_t x_{t-1}, \Sigma_t) \\ y_t | x_t &\sim \mathcal{N}(C_t x_t, R_t) \end{aligned}$$

In Fox et al, they set  $C_t = C$  and  $R_t = R$  to be constant across time and without loss of generality force  $C = [I_d, 0]$  where  $d$  is the dimensionality of  $y_t$ . Hence the key parameters that switch are  $A_t, \Sigma_t$  and the key parameter that is constant for all time is  $R$ .

For the purposes of visualization, we produce data from a projectile trajectory. We follow the discussion in Section 1.3.2 in *Bayesian Time Series Models* by Barber et al. According to a discretization of Newton's Laws of motion, acceleration is constant. Let  $q(t)$  be the two-vector of position,  $v(t)$  be the velocity vector, and  $a = (0, -9.8)$  be the acceleration. If a projectile is launched from the origin  $q_0 = (0, 0)$  with initial velocity  $v_0 = (10, 10)$  (ie the initial angle is  $\theta = \pi/4$ , we expect it to follow a parabolic path to a maximum height of

$$y_{max} = \frac{\|v_0\|^2 \sin^2(\theta)}{2\|a\|}$$

A discretized version of the trajectory is:

$$\begin{aligned} q(t+k) &= q(t) + kv(t) + \frac{k^2}{2}a \\ v(t+k) &= v(t) + ka(t) \end{aligned}$$

Where  $k$  is the time gap between observations (assumed to be constant for simplicity). Let  $x(t) = (q(t), v(t))$  (a 4-vector). We can write  $x(t) = Ax(t-1) + B$  where

$$\begin{aligned} A &= \begin{pmatrix} I & kI \\ 0 & I \end{pmatrix} \\ B &= \begin{pmatrix} \frac{1}{2}k^2 Ia \\ kIa \end{pmatrix} \end{aligned}$$

Hence

$$(x_t | x_{t-1}) \sim \mathcal{N}(Ax_{t-1} + B, \Sigma)$$

with  $\Sigma$  representing random perturbations in motion due to the wind. We set  $\Sigma$  to a diagonal matrix with diagonal elements  $(\sigma_1^2, \sigma_1^2, \sigma_2^2, \sigma_2^2)$ . This indicates the noise in the velocity part is different than the position part, but the noise is the same in any spatial direction. This is a linear dynamical system. Assume we observe only a noisy version of the position,  $y(t)$ . Hence, the observation model is

$$(y_t | x_t) \sim \mathcal{N}(Cx_t, R)$$

Where  $C = (I, 0)$  (a 2x4 matrix) and  $R$  is measurement noise.

## E Block Sampling for Linear Dynamical System

### E.1 Forward Kalman Filter

Let  $x_t$  represent the hidden (continuous) state and let  $y_t$  represent the noisy observation. Then the linear dynamical system of interest has the following probability model:

$$\begin{aligned} p(x_t|x_{t-1}) &= \mathcal{N}(A_t x_{t-1} + B_t, \Sigma_t) \\ p(y_t|x_t) &= \mathcal{N}(C x_t, R) \end{aligned}$$

Following Fox (see thesis section 2.7.5), we set  $C = [I_d, 0]$  where  $d$  is the dimensionality of  $y_t$ . The key switching parameters are  $A_t, B_t$ , and  $\Sigma_t$  and the key constant parameter is  $R$ . Note that our model is more general than that of Fox since we allow the presence of the  $B_t$  parameter while she sets it to zero. Assuming these dynamical parameters are known, we can use a variant of the Kalman Filter to sample from the posterior of all the  $x_t$  states given the observed  $y_t$  states. The idea is to first compute backward messages and then sample in a forward pass. First, we derive the recursion for the forward messages.

$$\alpha_{t+1}(x_{t+1}) = \left[ \int p(x_{t+1}|x_t) \alpha_t(x_t) dx_t \right] p(y_{t+1}|x_{t+1})$$

Ignoring normalizing constants, the integrand depends on the following quantities

$$\begin{aligned} p(x_{t+1}|x_t) &\propto \exp \left\{ -\frac{1}{2} (x_{t+1} - A x_t - B)' \Sigma^{-1} (x_{t+1} - A x_t - B) \right\} \\ &\propto \exp \left\{ -\frac{1}{2} \begin{pmatrix} x_{t+1} \\ x_t \end{pmatrix}' \begin{pmatrix} \Sigma^{-1} & -\Sigma^{-1} A \\ -A' \Sigma^{-1} & A' \Sigma^{-1} A \end{pmatrix} \begin{pmatrix} x_{t+1} \\ x_t \end{pmatrix} + \begin{pmatrix} \Sigma^{-1} B \\ -A' \Sigma^{-1} B \end{pmatrix}' \begin{pmatrix} x_{t+1} \\ x_t \end{pmatrix} \right\} \end{aligned}$$

Assume  $\alpha_t(x_t)$  is a known Gaussian density function with offset  $\theta_{t|t}^f$  and information matrix  $\Lambda_{t|t}^f$ . Then Fox shows that

$$\alpha_t(x_t) \propto \exp \left\{ -\frac{1}{2} \begin{pmatrix} x_{t+1} \\ x_t \end{pmatrix}' \begin{pmatrix} 0 & 0 \\ 0 & \Lambda_{t|t}^f \end{pmatrix} \begin{pmatrix} x_{t+1} \\ x_t \end{pmatrix} + \begin{pmatrix} 0 \\ \theta_{t|t}^f \end{pmatrix}' \begin{pmatrix} x_{t+1} \\ x_t \end{pmatrix} \right\}$$

The combined density in the integrand is then given by

$$\begin{aligned} p(x_{t+1}|x_t) \alpha_t(x_t) &\propto \exp \left\{ -\frac{1}{2} \begin{pmatrix} x_{t+1} \\ x_t \end{pmatrix}' \begin{pmatrix} \Sigma^{-1} & -\Sigma^{-1} A \\ -A' \Sigma^{-1} & A' \Sigma^{-1} A + \Lambda_{t|t}^f \end{pmatrix} \begin{pmatrix} x_{t+1} \\ x_t \end{pmatrix} \right\} \cdots \\ &\quad \cdots + \begin{pmatrix} \Sigma^{-1} B \\ \theta_{t|t}^f - A' \Sigma^{-1} B \end{pmatrix}' \begin{pmatrix} x_{t+1} \\ x_t \end{pmatrix} \right\} \end{aligned}$$

We now have the joint distribution of  $(x_{t+1}, x_t)$  which is in the form of a blocked bivariate Gaussian. The marginal distribution of  $x_{t+1}$  is obtained by integrating out  $x_t$  using a standard identity:

$$\int \mathcal{N}^{-1} \left( \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}; \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix} \right) dx_2 = \mathcal{N}^{-1}(x_1; \theta_1 - \Lambda_{12} \Lambda_{22}^{-1} \theta_2, \Lambda_{11} - \Lambda_{12} \Lambda_{22}^{-1} \Lambda_{21})$$

Therefore,

$$\int p(x_{t+1}|x_t) \alpha_t(x_t) dx_t \propto \mathcal{N}^{-1}(x_{t+1}; \theta_{t,t+1}, \Lambda_{t,t+1})$$

where

$$\begin{aligned} \theta_{t,t+1} &= \Sigma^{-1} B + \Sigma^{-1} A (A' \Sigma^{-1} A + \Lambda_{t|t}^f)^{-1} (\theta_{t|t}^f - A' \Sigma^{-1} B) \\ \Lambda_{t,t+1} &= \Sigma^{-1} - \Sigma^{-1} A (A' \Sigma^{-1} A + \Lambda_{t|t}^f)^{-1} A' \Sigma^{-1} \end{aligned}$$

Notably, our offset term is different from Fox due to the nonzero  $B$  but the information matrix is the same. The above update equations can be simplified if  $A$  is invertible (cf Fox Algorithm 3). Additionally, it is desirable to enforce symmetry in computing  $\Lambda_{t,t+1}$ . Set  $M_t = (A')^{-1} \Lambda_{t|t}^f A^{-1}$  and  $J_t = M_t (M_t + \Sigma^{-1})^{-1}$ . Note that  $M_t' = M_t$ . Then,

$$\begin{aligned} \Lambda_{t,t+1} &= \Sigma^{-1} \left( I - \left( \Sigma^{-1} + (A')^{-1} \Lambda_{t|t}^f A^{-1} \right)^{-1} \Sigma^{-1} \right) = \Sigma^{-1} \left( I - (\Sigma^{-1} + M_t)^{-1} \Sigma^{-1} \right) \\ &= \Sigma^{-1} (\Sigma^{-1} + M_t)^{-1} M_t = \Sigma^{-1} J_t' \end{aligned}$$

This is equivalent to Fox's Algorithm 3 formula, as shown below:

$$\begin{aligned}
\Lambda_{t,t+1} &= (I - J_t)M_t(I - J_t)' + J_t\Sigma^{-1}J_t' \\
&= (I - J_t)M_t(I - J_t)' + J_t(\Lambda_{t,t+1}) \\
(I - J_t)\Lambda_{t,t+1} &= (I - J_t)M_t(I - J_t') \\
\Lambda_{t,t+1} &= M_t(I - J_t') = M_t - M_tJ_t' \\
\Sigma^{-1}J_t' &= M_t - M_tJ_t' \\
(\Sigma^{-1} + M_t)J_t' &= M_t \\
(\Sigma^{-1} + M_t)(\Sigma^{-1} + M_t)^{-1}M_t &= M_t
\end{aligned}$$

Fox's formula is better numerically since it automatically enforces symmetry. By a similar argument, we have a simplified version of the offset parameter:

$$\begin{aligned}
\theta_{t,t+1}^f &= \Sigma^{-1} \left( B + (\Sigma^{-1} + M_t)^{-1} \left( (A')^{-1}\theta_{t|t}^f - \Sigma^{-1}B \right) \right) \\
&= \Sigma^{-1} \left( (I - (\Sigma^{-1} + M_t)^{-1}\Sigma^{-1})B + (\Sigma^{-1} + M_t)^{-1}(A')^{-1}\theta_{t|t}^f \right) \\
&= \Sigma^{-1}J_t'B + \Sigma^{-1}(\Sigma^{-1} + M_t)^{-1}(A')^{-1}\theta_{t|t}^f \\
&= \Lambda_{t,t+1}B + (I - J_t)(A')^{-1}\theta_{t|t}^f
\end{aligned}$$

This reduces to Fox's formula when  $B = \mathbf{0}$  as expected.

The likelihood term is the same as in Fox:

$$p(y_{t+1}|x_{t+1}) \propto \exp \left\{ -\frac{1}{2}x_{t+1}'C'R^{-1}Cx_{t+1} + x_{t+1}'C'R^{-1}y_{t+1} \right\}$$

The combined density is then given by

$$\alpha_{t+1}(x_{t+1}) \propto \exp \left\{ -\frac{1}{2}x_{t+1}'(\Lambda_{t,t+1} + C'R^{-1}C)x_{t+1} + x_{t+1}'(\theta_{t,t+1} + C'R^{-1}y_{t+1}) \right\}$$

Therefore the updated filtering information and offset parameters at step (t+1) are:

$$\begin{aligned}
\theta_{t+1|t+1}^f &= \theta_{t,t+1} + C'R^{-1}y_{t+1} \\
\Lambda_{t+1|t+1}^f &= \Lambda_{t,t+1} + C'R^{-1}C
\end{aligned}$$

Filtered estimates of  $x_t|y_{1:t}$  can be obtained from  $E[x_t|y_{1:t}] = \hat{x}_{t|t} = (\Lambda_{t|t}^f)^{-1}\theta_{t|t}^f$  or sampled from the updated density  $x_t|y_{1:t} \sim \mathcal{N}(\hat{x}_{t|t}, (\Lambda_{t|t}^f)^{-1})$

## E.2 Backward Kalman Filter

This section follows closely with Fox Appendix D.2. To perform smoothing or sampling based on the joint distribution of the hidden states given the observed states (rather than just the filtered distribution), we also need to compute backward messages, defined as:

$$m_{t,t-1}(x_{t-1}) = p(y_{t:T}|x_{t-1})$$

These are similar to the  $\beta_{t-1}$  messages that would be computed in the backward part of the forward-backward algorithm. If we already know  $m_{t+1,t}(x_t) \sim \mathcal{N}^{-1}(x_t; \theta_{t+1,t}^b, \Lambda_{t+1,t}^b)$ , then

$$m_{t,t-1} \propto \int p(x_t|x_{t-1})p(y_t|x_t)m_{t+1,t}(x_t)dx_t$$

The components of the integrand can be expressed as:

$$\begin{aligned}
p(x_t|x_{t-1}) &\propto \exp \left\{ -\frac{1}{2} \begin{pmatrix} x_{t-1} \\ x_t \end{pmatrix}' \begin{pmatrix} A'\Sigma^{-1}A & -A'\Sigma^{-1} \\ -\Sigma^{-1}A & \Sigma^{-1} \end{pmatrix} \begin{pmatrix} x_{t-1} \\ x_t \end{pmatrix} + \begin{pmatrix} x_{t-1} \\ x_t \end{pmatrix}' \begin{pmatrix} -A'\Sigma^{-1}B \\ \Sigma^{-1}B \end{pmatrix} \right\} \\
p(y_t|x_t) &\propto \exp \left\{ -\frac{1}{2} \begin{pmatrix} x_{t-1} \\ x_t \end{pmatrix}' \begin{pmatrix} 0 & 0 \\ 0 & C'R^{-1}C \end{pmatrix} \begin{pmatrix} x_{t-1} \\ x_t \end{pmatrix} + \begin{pmatrix} x_{t-1} \\ x_t \end{pmatrix}' \begin{pmatrix} 0 \\ C'R^{-1}y_t \end{pmatrix} \right\} \\
m_{t+1,t}(x_t) &\propto \exp \left\{ -\frac{1}{2} \begin{pmatrix} x_{t-1} \\ x_t \end{pmatrix}' \begin{pmatrix} 0 & 0 \\ 0 & \Lambda_{t+1,t}^b \end{pmatrix} \begin{pmatrix} x_{t-1} \\ x_t \end{pmatrix} + \begin{pmatrix} x_{t-1} \\ x_t \end{pmatrix}' \begin{pmatrix} 0 \\ \theta_{t+1,t}^b \end{pmatrix} \right\}
\end{aligned}$$

Combining these together, the integrand becomes:

$$p(x_t|x_{t-1})p(y_t|x_t)m_{t+1,t}(x_t) \propto \exp \left\{ -\frac{1}{2} \begin{pmatrix} x_{t-1} \\ x_t \end{pmatrix}' \begin{pmatrix} A'\Sigma^{-1}A & -A'\Sigma^{-1} \\ -\Sigma^{-1}A & \Sigma^{-1} + C'R^{-1}C + \Lambda_{t+1,t}^b \end{pmatrix} \begin{pmatrix} x_{t-1} \\ x_t \end{pmatrix} \right\} \dots$$

$$\dots + \begin{pmatrix} x_{t-1} \\ x_t \end{pmatrix}' \begin{pmatrix} -A'\Sigma^{-1}B \\ \Sigma^{-1}B + C'R^{-1}y_t + \theta_{t+1,t}^b \end{pmatrix} \Bigg\}$$

Applying the Gaussian marginalization identity from the previous section to integrate out  $x_t$ , we obtain

$$m_{t,t-1} \propto \mathcal{N}^{-1}(x_{t-1}; \theta_{t,t-1}^b, \Lambda_{t,t-1}^b)$$

where

$$\Lambda_{t,t-1}^b = A'\Sigma^{-1}A - A'\Sigma^{-1}(\Sigma^{-1} + C'R^{-1}C + \Lambda_{t+1,t}^b)^{-1}\Sigma^{-1}A$$

$$\theta_{t,t-1}^b = -A'\Sigma^{-1}B + A'\Sigma^{-1}(\Sigma^{-1} + C'R^{-1}C + \Lambda_{t+1,t}^b)^{-1}(\Sigma^{-1}B + C'R^{-1}y_t + \theta_{t+1,t}^b)$$

We see that these recursions agree with Fox's equation D.12, except she uses  $\mu$  in place of  $B$ . She provides additional derivations to improve numerical stability (Algorithm 19). In particular, the messages are reparametrized using

$$\Lambda_{t|t}^b = \Lambda_{t+1,t}^b + C'R^{-1}C$$

$$\theta_{t|t}^b = \theta_{t+1,t}^b + C'R^{-1}y_t$$

Once the backward messages have been computed, forward sampling of the hidden states  $x_t$  is given from the recursion

$$p(x_t|x_{t-1}, y_{1:T}) \propto p(x_t|x_{t-1})p(y_t|x_t)m_{t+1,t}(x_t)$$

After some algebra, this yields the following distribution for sampling:

$$p(x_t|x_{t-1}, y_{1:T}) \propto \mathcal{N}^{-1}(x_t; \Sigma^{-1}(Ax_{t-1} + B) + \theta_{t|t}^b, \Sigma^{-1} + \Lambda_{t|t}^b)$$

(Fox's equation D.18). Note that generalization to a time-varying process with  $A_t, B_t$ , and  $\Sigma_t$  known is straightforward.

## F Unknown Dynamical Parameters

Up to this point, we have assumed known dynamical parameters  $A_t, \Sigma_t, B, C$ , and  $R$  and used these to sample from the state sequence  $x_{1:T}$ . We now reverse this procedure to sample from the dynamical parameters conditional on a state sequence. Throughout, we are conditioning on a fixed mode sequence  $z_{1:T}$  which is sampled from the sticky HDP-HMM procedure. We continue to assume  $C$  is fixed (for identifiability). Conditional on a known sequence of states computed by the Kalman smoother/sampler procedure,  $A_t, \Sigma_t$  are independent of  $R$ . Focusing on  $A_t, \Sigma_t$ , Fox shows in thesis section 2.4.4 that the problem reduces to a collection of multivariate linear regressions:

$$Y \sim \mathcal{MN}(AX + B1', \Sigma, I)$$

where  $1$  indicates a vector of ones, of length  $n$ . Suppose  $Y$  has  $d$  rows and  $n$  columns. The likelihood is that of a matrix normal distribution. If  $Y \sim \mathcal{MN}(M, V, K)$  then  $M$  is the mean parameter,  $V$  is the row covariance parameter and  $K^{-1}$  is the column covariance parameter<sup>3</sup>. Our likelihood is a more general model than that presented by Fox, who sets  $B = 0$ . We modify the notation from that used in other sections for simplicity. In the context of the overall model, the  $t^{th}$  column of  $Y$  would correspond to  $x_t$  and the  $t^{th}$  column of  $X$  would correspond to  $x_{t-1}$ . Also, as shown by Fox the model must also be split up based on the mode allocations  $z_{1:T}$  from the HMM, but it turns out that each mode has its own conditionally independent multivariate linear regression, so we can ignore  $z_{1:T}$  in the notation here.

The conjugate priors are:

$$\Sigma \sim \mathcal{IW}(\nu, \Delta)$$

$$A|\Sigma \sim \mathcal{MN}(M_A, \Sigma, K_A)$$

$$B|\Sigma \sim \mathcal{N}(M_B, \Sigma/\kappa_0)$$

The hyperparameter  $\kappa_0$  allows  $B$  to be assigned a more or less diffuse prior than  $A$ . Note that no such parameter can be included in the prior for  $A$  since it is unidentifiable with respect to  $K_A$ . While in the prior,  $A$  and  $B$

<sup>3</sup>Fox's parameterization differs from common usage (cf wikipedia) in that the right covariance matrix here is  $K^{-1}$  rather than  $K$



are specified as independent conditional on  $\Sigma$ , they become dependent in the posterior. Therefore, we modify Fox's result to specify full conditionals rather than a complete posterior. Let  $D = \{X, Y\}$  represent the data. For the full conditional of  $A$ , note that conditional on  $B$ , we can replace  $Y$  with  $Y - B1'$  and obtain exactly the same distribution as derived by Fox for the special case of  $B = 0$ . Therefore,

$$p(A|\Sigma, D, B) = \mathcal{MN}(A; S_{ayx}S_{axx}^{-1}, \Sigma, S_{axx})$$

with slightly modified sufficient statistics

$$\begin{aligned} S_{axx} &= XX' + K_A \\ S_{ayx} &= (Y - B1')X' + M_A K_A = YX' - B(X1)' + M_A K_A \end{aligned}$$

Note that post-multiplying a matrix by  $1$  is equivalent to computing row sums of the matrix. Also,  $1'1 = n$  and pre-multiplying by  $1'$  yields column sums.

Similarly, conditional on  $AX$ , we can replace  $Y$  with  $Y - AX$  in the likelihood and consider the “data” for  $B$  to be the row vector  $1'$  rather than the data matrix  $X$ . Since  $B$  is a vector, the distribution is just a reparameterized Gaussian likelihood. The full conditional is proportional to all the terms in the joint density involving  $B$ :

$$\begin{aligned} p(B|\Sigma, D, A) &\propto \exp \left\{ -\frac{1}{2} \text{tr} \left[ \Sigma^{-1} ((Y - AX - B1')(Y - AX - B1')' + \kappa_0(B - M_B)(B - M_B)') \right] \right\} \\ &\propto \exp \left\{ -\frac{1}{2} \text{tr} \left[ \Sigma^{-1} (B1'1B' - (Y - AX)1B' - B1'(Y - AX)' + \dots \right. \right. \\ &\quad \left. \left. \dots + \kappa_0(BB' - M_B B' - B M_B') \right) \right] \right\} \\ &\propto \exp \left\{ -\frac{1}{2} \text{tr} \left[ \Sigma^{-1} (n + \kappa_0) \left( BB' - \frac{1}{n + \kappa_0} ((Y - AX)1 + \kappa_0 M_B) B' + \dots \right. \right. \right. \\ &\quad \left. \left. \left. \dots - B \frac{1}{n + \kappa_0} ((Y - AX)1 + \kappa_0 M_B)' \right) \right] \right\} \end{aligned}$$

This is the kernel of a multivariate normal distribution with covariance  $\Sigma/(n + \kappa_0)$  and mean

$$\frac{1}{n + \kappa_0} ((Y - AX)1 + \kappa_0 M_B) = \frac{1}{n + \kappa_0} (Y1 - AX1 + \kappa_0 M_B)$$

Finally, the full conditional of  $\Sigma$  is proportional to all terms involving  $\Sigma$  in the joint density function:

$$\begin{aligned} p(\Sigma|D, A, B) &\propto |\Sigma|^{-\frac{n}{2}} \exp \left\{ -\frac{1}{2} \text{tr} \left[ \Sigma^{-1} ((Y - AX - B1')(Y - AX - B1')') \right] \right\} \times \dots \\ &\quad \dots \times |\Sigma|^{-\frac{\nu + d + 1}{2}} \exp \left\{ -\frac{1}{2} \text{tr} (\Sigma^{-1} \Delta) \right\} \end{aligned}$$

This is the kernel of an inverse Wishart distribution with degrees of freedom  $\nu + n$  and scale matrix

$$\Delta_n = \Delta + (Y - AX - B1')(Y - AX - B1)'$$

## G References

### References

- Emily Beth Fox. *Bayesian nonparametric learning of complex dynamical phenomena*. Thesis, Massachusetts Institute of Technology, 2009. URL <http://dspace.mit.edu/handle/1721.1/55111>.
- Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006. ISSN 0162-1459. URL <http://www.jstor.org/stable/27639773>.