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
zhl112@mail.harvard.edu

Contents

1 Problem Formulation

We consider the estimation of switching linear dynamical systems (SLDS) 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 an 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:

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

$$y_t|x_t \sim \mathcal{N}(Cx_t, R) \tag{2}$$

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, Σ_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 $\mathbf{\Theta} = \mathcal{A} \times \mathcal{B} \times \{\Sigma_t\}$ indexed by $\mathcal{Z}$. Finally, if denote $z_t \in \mathcal{Z}$ the index of θ_t, SLDS assumes z_t follows an Markov process with transition matrix $\mathbf{\Pi}_{|\Theta|\times|\Theta|} = [\pi_1, \ldots, \pi_z, \ldots]^T$, 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 $\mathbf{y}_t \in \mathbb{R}^3$ the observed position of the maneuvering fighter aircraft, which comes from a latent $\mathbf{x}_t \in \mathbb{R}^9$ comprised of position, speed and momentum in the 3D space. We can learn how pilot executes different maneuvers by estimating $\mathbf{\Theta}$ (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 $\mathbf{\Pi}_{|\Theta|\times|\Theta|}$ 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 $\mathbf{\Pi}$ is $O(|\mathbf{\Theta}|^2)$ 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 $\mathbf{\Pi}$ through global shrinkage on each state-specific transition distributions.

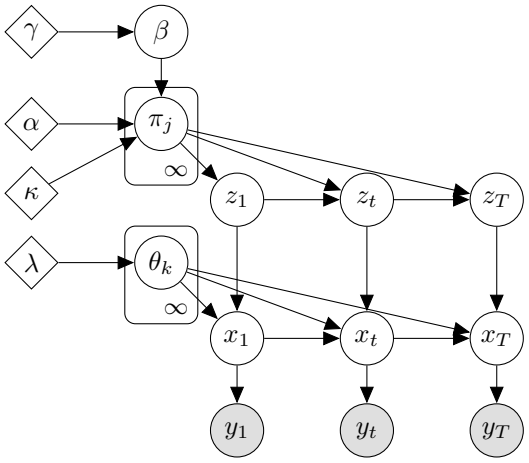


Figure 1: Graphical Model for Switching Linear Dynamics System

In the rest of this report, we discuss how to properly adapt HDP into the estimation of $\mathbf{\Pi}$ 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 $\mathbf{\Theta}$ and $\mathbf{\Pi}$. 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 $\boldsymbol{\theta}$ on a state-space model, we have to battle several technical difficulties such as (1) choosing proper prior for $\boldsymbol{\theta}$, (2) identify efficient algorithm for computing $p(\mathbf{z}|\mathbf{y}, \mathbf{\Theta}, \mathbf{\Pi})$ on Markov time series, and (3) explicit sampling for

infinite dimensional π_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 $\boldsymbol{\theta}$ are not of interest in classical HDP and therefore $\boldsymbol{\theta}$, π_j can be integrated out in their Gibbs sampler. In this section, we discuss how we have overcame the three aforementioned difficulties and constructed our Gibbs sampler for efficient and regularized joint inference of $\boldsymbol{\Theta}$ and $\boldsymbol{\Pi}$.

2.1 Sampling Dynamical Parameters θ 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[1]. 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[2] lead to closed form full conditionals, see Appendix E for derivations. Hyperparameters can be set to regularize estimates toward stable dynamics.

2.2 Sampling for Hidden Variables (x, 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 D.

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 $\boldsymbol{\Pi}$, $\boldsymbol{\beta}$ using HDP Prior

Finally, we perform estimation of π_k, the rows of the transition matrix $\boldsymbol{\Pi}$. We notice that the high-dimensional nature of $\boldsymbol{\Pi}$ ($dim(\boldsymbol{\Pi}) = O(|\mathcal{Z}|^2)$) calls for regularization not only on the individual values of π_{kj}'s, but more importantly, also on the global dimension of $\boldsymbol{\Pi}$ (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 $\boldsymbol{\Theta}_k$'s.

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

$$Dirichlet(\frac{\alpha}{K}, \dots, \frac{\alpha}{K}) \to GEM(\alpha)$$

and effectively upperbounding the number of possible clusters by K.

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 κ, i.e. using the sticky HDP prior:

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

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

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

[2] $\mathcal{IW}$=Inverse Wishart

Algorithm 1 MKL PFBS algorithm

procedure `rls_dual_mkl_pfbs`($\mathbf{K}$, $\mathbf{y}$, (τ, μ, σ))
  $\boldsymbol{\alpha}^0 = \mathbf{0}$
  for $p = 1$ to `MAX_ITER` **do**
    $\boldsymbol{\alpha}_0^p = (1 - \frac{\mu}{\sigma})\boldsymbol{\alpha}^{p-1} - \frac{1}{\sigma n}(\mathbf{K}\boldsymbol{\alpha}^{p-1} - \mathbf{y})$
    $\boldsymbol{\alpha}^p = \mathbf{S}_{\frac{\tau}{\sigma}}(K, \boldsymbol{\alpha}_0^p)$
  end for
  return $f^{\texttt{MAX\_ITER}} = (\boldsymbol{\alpha}^{\texttt{MAX\_ITER}})^T\mathbf{k}$
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 Θ and $x_{1:T}$. We

3.3 HDP-SLDS

4 Conclusion and Future Direction

4.1 Sampling Hyperparameters

4.2 Automatic Relevance Determination

A Hierarchical Dirichlet Process

A.1 Model

Classic view:

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

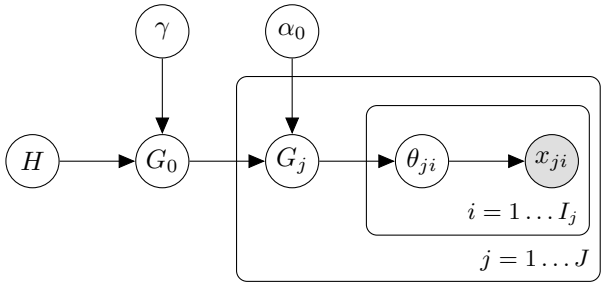
where $P \sim DP(\alpha, G)$ adopts the stick breaking representation w.p. 1:

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

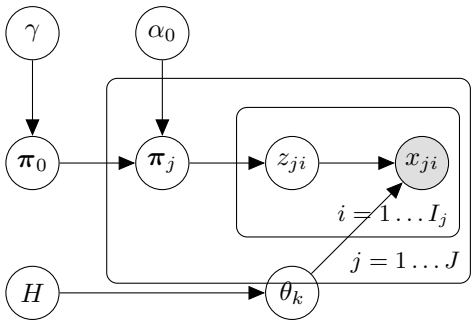
Alternatively, one may describe the generative processes of π_k and θ_k separately as:

$$\begin{aligned} \boldsymbol{\pi}_0|\gamma &\sim GEM(\gamma) & \theta_k|H \sim H \\ \boldsymbol{\pi}_j|\alpha_0, \boldsymbol{\pi}_0 &\sim DP(\alpha_0, \boldsymbol{\pi}_0) \end{aligned}$$

$$\begin{aligned} z_{ji}|\boldsymbol{\pi}_j &\sim \boldsymbol{\pi}_j \\ x_{ji}|z_{ji}, (\theta_k)_{k=1}^{\infty} &\sim F(\theta_{z_{ji}}) \end{aligned}$$



(a) Hierarchical Dirichlet Process



(b) Hierarchical Dirichlet Process

Figure 2: Hierarchical Dirichlet Process

A.2 Inference

Assuming conjugacy between H and F and holding (γ, α_0) fixed,

we now describe a simplied Gibbs approach to sample parameters $(z_{ji}, m_{jk}, \boldsymbol{\pi}_0)$ from the Chinese Restaurant Franchise (see Appendix C) representation of the posterior, where the parameter z_{ji} are refered to respectively as customer-specific dish assignment, m_{jk} as dish-specific table count, and π_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 $(z_{ji}, m_{jk}, \boldsymbol{\pi}_0)$ iteratively as:

1. Sample $z_{ji} = k|\mathbf{z}_{-(ji)}, \mathbf{m}, \boldsymbol{\pi}_0$ from the distribution:

$$z_{ji} = k|\mathbf{z}_{-(ji)}, \mathbf{m}, \boldsymbol{\pi}_0 \propto \begin{cases} f_k^{-x_{ji}}(x_{ji}) * (n_{jk}^{-(ji)} + \alpha_0\pi_{0,k}) & k \leq K \\ f_{K+1}^{-x_{ji}}(x_{ji}) * \alpha_0\pi_{0,u} & k = K+1 \end{cases}$$

2. Sample $m_{jk} = m|\mathbf{z}, \mathbf{m}_{-(jk)}, \boldsymbol{\pi}_0$, by setting $m_{jk} = \sum_i I(t_{ji} = t_{new}|k_{jt_{new}} = k)$, we can sample t_{ji} from:

$$t_{ji} = t|k_{jt} = k, \mathbf{t}_{-(ji)}, \boldsymbol{\pi}_0 \propto \begin{cases} n_{jt}^{-(ji)} & t \leq T \\ \alpha_0\pi_{0,k} & t = T+1 \end{cases}$$

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

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

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

$$\boldsymbol{\pi}_0 \sim Dir(m_1, \ldots, m_K, \gamma)$$

A.3 Application: Clustering Hierarchical Gaussian Data

Consider mixture of Gaussian data $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$ with $\mathbf{x}_k \overset{iid}{\sim} MVN(\boldsymbol{\theta}_{k,2\times 1}, \mathbf{I}_{2\times 2})$ 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:

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

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

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

B HDP for Hidden Markov Model

B.1 Hidden Markov Model

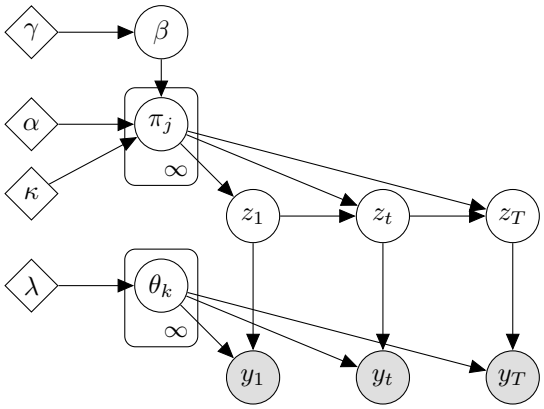


Figure 3: Hidden Markov Model

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

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

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

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 π_k is augmented *a priori* from β 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 β and π follows:

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

Define $\boldsymbol{\theta}_k$ as emission parameter for state k, we sample $(\mathbf{z},\mathbf{m},\boldsymbol{\pi}_0,\boldsymbol{\theta})$ 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:

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

   (b) Sample override variable:

$$w_j \sim Binom\Big(m'_{jj}, \frac{\kappa}{\kappa + \alpha\boldsymbol{\pi}_{0,j}}\Big)$$

   (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:

$$\boldsymbol{\pi}_0 \sim Dir(\frac{\gamma}{L} + m_1, \ldots, \frac{\gamma}{L} + m_K)$$

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:

$$\begin{aligned} \text{Backward Message}: &\quad \beta_t(z_t) = p(\mathbf{y}_{T>t}|z_t) \\ \text{Forward Message}: &\quad \alpha_t(z_t) = p(\mathbf{y}_{T\leq t}, z_t) \\ \text{Joint Message}: &\quad \alpha_t(z_t)\beta_t(z_t) = p(\mathbf{y}, z_t) \end{aligned}$$

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

$$\begin{aligned} \text{Backward Message}: &\quad \beta_t(z_t) = p(\mathbf{y}_{T>t}|z_t) = m_{t+1,t}(z_t) \\ \text{Forward Message}: &\quad \alpha_t(z_t) = p(y_t|z_t)p(\mathbf{y}_{T<t}, z_t) = p(y_t|z_t)m_{t-1,t}(z_t) \end{aligned}$$

.

These two types of messages can be computed β_t backward and α_t forward in time as:

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

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

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 θ'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:

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

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 Block Sampling for Linear Dynamical System

D.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 Σ_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

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

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

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

The combined density in the integrand is then given by

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

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

$$\theta_{t,t+1} = \Sigma^{-1}B + \Sigma^{-1}A(A'\Sigma^{-1}A+\Lambda^f_{t|t})^{-1}(\theta^f_{t|t}-A'\Sigma^{-1}B)$$
$$\Lambda_{t,t+1} = \Sigma^{-1} - \Sigma^{-1}A(A'\Sigma^{-1}A+\Lambda^f_{t|t})^{-1}A'\Sigma^{-1}$$

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 $M_t = (A')^{-1}\Lambda^f_{t|t}A^{-1}$ and $J_t = M_t(M_t+\Sigma^{-1})^{-1}$. Note that $M_t' = M_t$. Then,

$$\Lambda_{t,t+1} = \Sigma^{-1}\left(I-\left(\Sigma^{-1}+(A')^{-1}\Lambda^f_{t|t}A^{-1}\right)^{-1}\Sigma^{-1}\right) = \Sigma^{-1}\left(I-\left(\Sigma^{-1}+M_t\right)^{-1}\Sigma^{-1}\right)$$
$$= \Sigma^{-1}\left(\Sigma^{-1}+M_t\right)^{-1}M_t = \Sigma^{-1}J_t'$$

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:

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

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

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

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

$$\theta^f_{t+1|t+1} = \theta_{t,t+1} + C'R^{-1}y_{t+1}$$
$$\Lambda^f_{t+1|t+1} = \Lambda_{t,t+1} + C'R^{-1}C$$

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

D.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 β_{t-1} messages that would be computed in the backward part of the forward-backward algorithm. If we already know $m_{t+1,t}(x_t) \sim \mathcal{N}^{-1}(x_t;\ \theta^b_{t+1,t},\ \Lambda^b_{t+1,t})$, 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:

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

$$p(y_t|x_t) \propto \exp\left\{-\frac{1}{2}\begin{pmatrix}x_{t-1}\\x_t\end{pmatrix}'\begin{pmatrix}0 & 0\\ 0 & C'R^{-1}C\end{pmatrix}\begin{pmatrix}x_{t-1}\\x_t\end{pmatrix} + \begin{pmatrix}x_{t-1}\\x_t\end{pmatrix}'\begin{pmatrix}0\\ C'R^{-1}y_t\end{pmatrix}\right\}$$

$$m_{t+1,t}(x_t) \propto \exp\left\{-\frac{1}{2}\begin{pmatrix}x_{t-1}\\x_t\end{pmatrix}'\begin{pmatrix}0 & 0\\ 0 & \Lambda^b_{t+1,t}\end{pmatrix}\begin{pmatrix}x_{t-1}\\x_t\end{pmatrix} + \begin{pmatrix}x_{t-1}\\x_t\end{pmatrix}'\begin{pmatrix}0\\ \theta^b_{t+1,t}\end{pmatrix}\right\}$$

Combining these together, the integrand becomes:

$$p(x_t|x_{t-1})p(y_t|x_t)m_{t+1,t}(x_t) \propto \exp\left\{-\frac{1}{2}\begin{pmatrix}x_{t-1}\\x_t\end{pmatrix}'\begin{pmatrix}A'\Sigma^{-1}A & -A'\Sigma^{-1}\\ -\Sigma^{-1}A & \Sigma^{-1} + C'R^{-1}C + \Lambda^b_{t+1,t}\end{pmatrix}\begin{pmatrix}x_{t-1}\\x_t\end{pmatrix}\cdots\right.$$
$$\left.\cdots + \begin{pmatrix}x_{t-1}\\x_t\end{pmatrix}'\begin{pmatrix}-A'\Sigma^{-1}B\\ \Sigma^{-1}B + C'R^{-1}y_t + \theta^b_{t+1,t}\end{pmatrix}\right\}$$

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

$$m_{t,t-1} \propto \mathcal{N}^{-1}(x_{t-1}\ ;\ \theta^b_{t,t-1}\ ,\ \Lambda^b_{t,t-1})$$

where

$$\Lambda^b_{t,t-1} = A'\Sigma^{-1}A - A'\Sigma^{-1}(\Sigma^{-1} + C'R^{-1}C + \Lambda^b_{t+1,t})^{-1}\Sigma^{-1}A$$
$$\theta^b_{t,t-1} = -A'\Sigma^{-1}B + A'\Sigma^{-1}(\Sigma^{-1} + C'R^{-1}C + \Lambda^b_{t+1,t})^{-1}(\Sigma^{-1}B + C'R^{-1}y_t + \theta^b_{t+1,t})$$

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

$$\Lambda^b_{t|t} = \Lambda^b_{t+1,t} + C'R^{-1}C$$
$$\theta^b_{t|t} = \theta^b_{t+1,t} + C'R^{-1}y_t$$

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:

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

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

E Unknown Dynamical Parameters

Up to this point, we have assumed known dynamical parameters A_t, Σ_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, Σ_t are independent of R. Focusing on A_t,Σ_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[3]. Our likelihood is a more general model than that presented by Fox, who sets $B = \mathbf{0}$. 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:

$$\begin{aligned}\Sigma &\sim \mathcal{IW}(\nu, \Delta)\\ A|\Sigma &\sim \mathcal{MN}(M_A, \Sigma, K_A)\\ B|\Sigma &\sim \mathcal{N}(M_B, \Sigma/\kappa_0)\end{aligned}$$

The hyperparameter κ_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 are specified as independent conditional on Σ, 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,

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

with slightly modified sufficient statistics

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

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:

$$\begin{aligned}p(B|\Sigma, D, A) &\propto \exp\left\{-\frac{1}{2}\text{tr}\left[\Sigma^{-1}\left((Y - AX - B1')(Y - AX - B1')' + \kappa_0(B - M_B)(B - M_B)'\right)\right]\right\}\\ &\propto \exp\left\{-\frac{1}{2}\text{tr}\left[\Sigma^{-1}\left(B1'1B' - (Y - AX)1B' - B1'(Y - AX)' \ + \dots\right.\right.\right.\\ &\left.\left.\left. \dots \ +\kappa_0(BB' - M_BB' - BM_B')\right)\right]\right\}\\ &\propto \exp\left\{-\frac{1}{2}\text{tr}\left[\Sigma^{-1}(n + \kappa_0)\left(BB' - \frac{1}{n + \kappa_0}\left((Y - AX)1 + \kappa_0M_B\right)B' \ + \dots\right.\right.\right.\\ &\left.\left.\left. \dots \ -B\frac{1}{n + \kappa_0}\left((Y - AX)1 + \kappa_0M_B\right)'\right)\right]\right\}\end{aligned}$$

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

$$\frac{1}{n + \kappa_0}\left((Y - AX)1 + \kappa_0M_B\right) = \frac{1}{n + \kappa_0}\left(Y1 - AX1 + \kappa_0M_B\right)$$

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

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

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

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

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

F References

References

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`.

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`.