Revista Colombiana de Estadística - Theoretical Statistics
January 2022, volume 45, issue 1, pp. 231–263
http://doi.org/10.15446/rce.v45n1.85385

Variable Selection in Switching Dynamic Regression Models

Selección de variables en modelos de regresión dinámicos de cambios de régimen

Dayna P. Saldaña-Zepeda[1,a], Ciro Velasco-Cruz[2,b], Víctor H. Torres-Preciado[3,c]

[1] Escuela de Mercadotecnia, Universidad de Colima, Colima, México

[2] Departamento de Estadística, Socioeconomía Estadística e Infomática, Colegio de Postgraduados, Texcoco, México

[3] Facultad de Economía, Universidad de Colima, Colima, México

Abstract

Complex dynamic phenomena in which dynamics is related to events (modes) that cause structural changes over time, are well described by the switching linear dynamical system (SLDS). We extend the SLDS by allowing the measurement noise to be mode-specific, a flexible way to model non stationary data. Additionally, for models that are functions of explanatory variables, we adapt a variable selection method to identify which of them are significant in each mode. Our proposed model is a flexible Bayesian nonparametric model that allows to learn about the number of modes and their location, and within each mode, it identifies the significant variables and estimates the regression coefficients. The model performance is evaluated by simulation and two application examples from a dataset of meteorological time series of Barranquilla, Colombia are presented.

Key words**:** bayesian filtering and smoothing; dirichlet process; hierarchical model; state-space model.

Resumen

Fenómenos dinámicos complejos en los que la dinámica está relacionada con eventos (modos) que provocan cambios estructurales a lo largo del tiempo, se aproximan mediante un sistema dinámico lineal de cambio de régimen (SDLR). Extendemos el SDLR al permitir que el error de medición

[a] Ph.D. E-mail: daynasz@ucol.mx
[b] Ph.D. E-mail: cvelasco@vt.edu
[c] Ph.D. E-mail: tpreciado04@gmail.com

sea específico del modo, una forma flexible de modelar datos no estacionarios. Además, para los modelos que son funciones de variables explicativas, adaptamos un método de selección de variables para identificar cuáles de ellas son significativas en cada modo. El modelo propuesto es un modelo bayesiano no paramétrico flexible que permite conocer el número de modos y su ubicación, y dentro de cada modo, identifica las variables significativas y estima los coeficientes de regresión. El desempeño del modelo se evalúa mediante simulación y se presentan dos ejemplos de aplicación de un conjunto de datos de series de tiempo meteorológicas de Barranquilla, Colombia.

***Palabras clave*:** filtrado y suavizamiento bayesianos; modelos de espacio-estado; modelos jerárquicos; procesos Dirichlet.

1. Introduction

The dynamic nature of many phenomena in fields such as finance and economics (Kim, 1994; Carvalho & Lopes, 2007; West, 2013; Zeng & Wu, 2013; McAlinn & West, 2016), human motion (Bregler, 1997, June; Pavlović et al., 2001), and environment (Lamon III et al., 1998; Huerta et al., 2004; Velasco-Cruz et al., 2012), has motivated the development of flexible methods in order to be able to handle changing conditions in the environment of the phenomena through time. For instance, several changes may occur in an endemic disease that may trigger an epidemic, or become highly contagious and thus trigger a pandemic; as a result, its rate of occurrence might change dramatically. The economic system of a country may change from a command-based economy to a market-based economy, or to a green economy, and then modify its economic activity. In modeling the rate of occurrence of a desease or the economic activity of a country through time, different *states* or *modes* account for changes the phenomena exhibit. Each *mode* is associated by an event (endemic, epidemic, pandemic; command economy, market economy, green economy), and the time series within each *mode* can be modeled by a linear dynamic system, such that the complex dynamics of the whole series can be approximated as switches among conditionally linear dynamical modes, known as Switching Linear Dynamical System (SLDS).

The SLDS can be viewed as a combination of the Hidden Markov Model (HMM) with a set of Linear Dynamical Systems (LDSs). In the HMM, the observations, which can be either discrete or continuous, are conditionally independent given a sequence of unobserved (*hidden*) discrete-valued *modes* that satisfy the *Markov property.* The LDS is considered an extension of the HMM with a linear structure on continuous latent (*hidden*) variables, most commonly Gaussian. The assumption of linearity and Gaussianity is specific to the LDS, but the dependence structure of the observations on the hidden variables is part of the definition of a general *state-space* model (Kalman, 1963, 1960). The SLDS is then formed by multiple Markov chains of continuous linear-Gaussian latent variables, each one associated with a value of the hidden discrete variable, such that the dynamics of a phenomenon is modeled by switching linear models, according to the underlying *mode*[1] sequence.

Inference in both, the HMM and the LDS, involves the quantification of the uncertainty associated with the unobserved variable at a particular time instance conditional on a sequence of data. To be explicit, let c be the *hidden* variable and y the observed variable, such that y_t is assumed conditionally independent of all other observations given the *state* c_t, for $t = 1, 2, \dots, T$. The sequence of states follows a first order Markov chain; if c is discrete-valued the process is a HMM, but if c is linear-Gaussian the process is a LDS. We use the notation $x_{1:t}$ to compactly represent the sequence $\{x_1, \dots, x_t\}$. The main inference problem resides in the computation of $p(c_{1:T}|y_{1:T})$. Since computing the full joint distribution of the states at all time steps is computationally very inefficient, it is more convenient to compute the marginal distribution $p(c_t|y_{1:l})$. Three different conditions are identified depending on l and t. If $l < t$ then it is known as *prediction*; if $l = t$ then it is referred to as *filtering*; and if $l > t$ then it is the *smoothing* problem. A recursive algorithm known as the *forward-backward* algorithm is commonly used to solve the inference problems. Since the LDS is a linear-Gaussian model, the joint distributions, as well as the marginals and conditionals, are Gaussians. It is well-known that the *filtering-smoothing* problems for the LDS are solved by the *Kalman filter* (Kalman, 1960) and Kalman smoother (Rauch et al., 1965).

The HMM can be viewed as a mixture model. The observations are regarded as being drawn independently from a mixture of distributions $F(\theta)$, such that $y_t|\theta_t \sim F(\theta_t)$, where θ_t are the parameters of the mixture component to which y_t belongs, and $F(\cdot)$ represents the distribution of the mixture. In this context, the discrete variable c_t indexes the parameters of the mixture component generating observation y_t as $y_t|\{\theta_k\}_{k=1}^{\infty}, c_t \sim F(\theta_{c_t})$. If two data points, y_t and y_l, belong to the same component (*cluster*), their component parameters will be identical, $\theta_t = \theta_l$. In the Bayesian nonparametric context, a Dirichlet Process (DP; Ferguson 1973) is used as a prior over the parameters of the mixture components. The resulting model is known as Dirichlet Process Mixture Model (DPMM; Antoniak 1974), which is used in a variety of clustering applications where the number of clusters is not known a priori.

In defining the SLDS, the DPMM in its hierarchical extension, proposed by Teh et al. (2006), and the LDS are fundamental. Fox et al. (2011*a*) develop a sampling algorithm that combines a truncated approximation of the DP (see Ishwaran & James, 2001, 2002; Ishwaran & Zarepour, 2002*b*) with an efficient joint sampling of the *mode* (due to the hidden Markov model) and *states* sequences (due to the linear dynamical system). The model is developed for time series exhibiting dynamical behaviors with the dynamics of the latent state process being mode-specific, only. Additionally, the model does not relate the time series to any explanatory variables other than intercept. In this paper we present an extension of the SLDS of Fox et al. (2011*a*), consisting of making the measurement model mode-specific. More precisely, the variance of the noise of this model is mode-

[1] The term *state* has a generic use in the state-space models. To avoid confusion, the continuous hidden variables are referred to as *states*, while the discrete hidden variables are referred to as *modes*.

specific. Also, the time series is described in terms of a regression model with explanatory variables. Furthermore, we include a variable selection method based on the formulation of Kuo & Mallick (1998). An important implication of it is that the design matrix is mode-specific, which is an element to distinguish between the distributions in the mixture. By employing a changing design matrix allows us to describe changing relationships between variables as time evolves. As a result, our proposed model is a flexible dynamic regression model for learning about the number of component distributions, for identifying the important variables, for determining where the mode changes originate, and for estimating the regression coefficients.

The outline of the paper is as follows. In Section 2 we provide some basic background on the *forward-backward* procedure, the DPMM, the hierarchical DPMM, and the *sticky* hierarchical DPMM (Fox et al., 2011*a*). The proposed model is derived in Section 3. We outline a Gibbs sampler and present results on synthetic datasets in Section 4. We use the proposed model on two applications by using a real dataset of meteorological series of Barranquilla, Colombia in Section 5. The paper concludes with a discussion.

2. Preliminaries

2.1. Forward-Backward Procedure

Consider the following general probabilistic *state-space* model:

$$\begin{aligned} &c_t \sim p(c_t|c_{t-1}) \\ &y_t \sim p(y_t|c_t), \quad t = 1, \dots, T. \end{aligned} \tag{1}$$

This dynamic model describes the system's dynamics and its uncertainties by a *Markov chain* on the state c, and a measurement model that describes y_t as a function of c_t.

The above *state-space* model is completely specified by the initial distribution $p(c_0)$ as

$$p(c_{0:t}, y_{1:t}) = p(c_0) \prod_{i=1}^{t} p(c_i|c_{i-1}) p(y_i|c_i). \tag{2}$$

From (2) one can derive any other distribution of interest. Because computing the full joint distribution of the states at all time steps is computationally very inefficient, the following marginal distributions for each $t = 1, \dots, T$ are considered instead: *filtering* distributions $p(c_t|y_{1:t})$; *prediction* distributions $p(c_{t+n}|y_{1:t})$, $n = 1, 2, \dots$; and *smoothing* distributions $p(c_t|y_{1:T})$. The two most important examples of *state-space* models are the HMM in which the hidden variables are discrete-valued random variables, and the LDS in which the state variables are Gaussians. Analytical derivations of the marginal distributions above are straightforward in both the HMM and the LDS. However, we only provide the resulting distributions for the LDS. For a review of the inference problems in both the HMM and the

LDS see Bishop (2006) and Barber (2012) (Meinhold & Singpurwalla, 1983 and Petris et al., 2009 are excellent references for beginners and practitioners in LDS).

Since the LDS is a linear-Gaussian model, the joint distributions over all latent and observed variables are Gaussian. The closed form solutions resulting from the Gaussian assumption allow us to have optimal algorithms for Bayesian filtering and smoothing. The filtering problem is solved by the well-known *Kalman filter* (Kalman, 1960); the corresponding smoothing problem is solved by the *Rauch-Tung-Striebel smoother* (RTSS) (Rauch et al., 1965). We briefly outline the algorithms.

2.1.1. Kalman filter

For the filtering problem, the data are supposed to arrive sequentially in time. We aim to estimate the current value of the *state* vector, based on the observations up to time t, in order to update our estimates and forecasts as new data become available at time $t+1$. For a general state-space model defined by (1), the filtering distribution can be computed as

$$p(c_t|y_{1:t}) = \frac{p(y_t|c_t, y_{1:t-1})p(c_t|y_{1:t-1})}{p(y_t|y_{1:t-1})} = \frac{p(y_t|c_t)p(c_t|y_{1:t-1})}{p(y_t|y_{1:t-1})}, \tag{3}$$

where $p(c_t|y_{1:t-1})$ and $p(y_t|y_{1:t-1})$ are the one-step-ahead predictive distributions for the states and for the measurements, respectively. Specifically for the LDS, consider the following model:

$$\begin{aligned} c_t &= A_t c_{t-1} + e_t, \quad e_t \sim N(0, \Sigma_t), \\ y_t &= X_t' c_t + w_t, \quad w_t \sim N(0, r_t), \end{aligned} \tag{4}$$

which is a simple linear-Gaussian case of (1) with c_t a vector of size p and y_t scalar. The matrix A_t is known as *evolution matrix*, and X_t is a vector of explanatory variables. The evolution error e_t and the measurement error w_t are internally and mutually independent random variables, such that for all (t, t'), $t \neq t'$, e_t and $e_{t'}$ are independent, w_t and $w_{t'}$ are independent, and e_t and $w_{t'}$ are independent (see West & Harrison, 1997 for an extensive review of LDSs). The model can readily be extended to the case where y_t is a vector.

Using standard results of the multivariate Gaussian distribution to the model in (4), it follows that the marginal and conditional distributions in (3) are also Gaussians, which are completely determined by their means and variances. Assuming that

$$p(c_{t-1}|y_{1:t-1}) = N(f_{t-1}, F_{t-1}),$$

the solution of the filtering problem for LDS is given by

$$\begin{aligned}
m_t &= \mathrm{E}(c_t|y_{1:t-1}) = \mathrm{E}(A_t c_{t-1} + e_t|y_{1:t-1}) = A_t f_{t-1}\\
S_t &= \mathrm{Var}(c_t|y_{1:t-1}) = \mathrm{Var}(A_t c_{t-1} + e_t|y_{1:t-1}) = A_t F_{t-1} A_t' + \Sigma_t\\
a_t &= \mathrm{E}(y_t|y_{1:t-1}) = \mathrm{E}(\mathrm{E}(y_t|c_t, y_{1:t-1})|y_{1:t-1}) = \mathrm{E}(X_t' c_t|y_{1:t-1}) = X_t' m_t\\
Q_t &= \mathrm{Var}(y_t|y_{1:t-1}) = \mathrm{E}(\mathrm{Var}(y_t|c_t, y_{1:t-1})|y_{1:t-1}) + \mathrm{Var}(\mathrm{E}(y_t|c_t, y_{1:t-1})|y_{1:t-1})\\
&= r_t + X_t' S_t X_t.
\end{aligned}$$

Then, it can be shown that $(c_t|y_{1:t}) \sim N(f_t, F_t)$, where:

$$\begin{aligned}
f_t &= F_t(X_t r_t^{-1} y_t + S_t^{-1} m_t) = m_t + S_t X_t (X_t' S_t X_t + r_t)^{-1}(y_t - X_t' m_t)\\
F_t &= (S_t^{-1} + X_t r_t^{-1} X_t')^{-1} = S_t - S_t X_t (X_t' S_t X_t + r_t)^{-1} X_t' S_t
\end{aligned}$$

One can learn about the state sequence $c_{1:T}$ by working sequentially *forwards* in time; starting with $c_0 \sim p(c_0)$ in order to compute $p(c_1|y_{1:1})$, and then estimating the filtering distribution $p(c_t|y_{1:t})$, $t = 2, \dots, T$, as new data become available (see Petris et al., 2009).

2.1.2. Kalman smoother

For the smoothing, we aim to retrospectively estimate the current value of the *state* vector. That is to say, we have observations of a time series up to time T to compute the conditional distributions of c_t given $y_{1:T}$, for any $t < T$, using the *backward*-recursive algorithm.

For a general *state-space* model defined in (1), the smoothing distributions can be computed as

$$p(c_t|y_{1:T}) = p(c_t|y_{1:t}) \int \frac{p(c_{t+1}|c_t)}{p(c_{t+1}|y_{1:t})} p(c_{t+1}|y_{1:T}) dc_{t+1}. \qquad (5)$$

Because the distributions in (5) are Gaussians, $p(c_t|y_{1:T})$ is Gaussian, with mean and variance given by

$$\begin{aligned}
b_t &= \mathrm{E}(c_t|y_{1:T}) = \mathrm{E}(\mathrm{E}(c_t|c_{t+1}, y_{1:T})|y_{1:T}) = f_t + F_t A_{t+1}' S_{t+1}^{-1}(b_{t+1} - m_{t+1}).\\
B_t &= \mathrm{Var}(c_t|y_{1:T}) = \mathrm{Var}(\mathrm{E}(c_t|c_{t+1}, y_{1:T})|y_{1:T}) + \mathrm{E}(\mathrm{Var}(c_t|c_{t+1}, y_{1:T})|y_{1:T})\\
&= F_t A_{t+1}' S_{t+1}^{-1} B_{t+1} S_{t+1}^{-1} A_{t+1} F_t + F_t - F_t A_{t+1}' S_{t+1}^{-1} A_{t+1} F_t\\
&= F_t + F_t A_{t+1}' S_{t+1}^{-1}(B_{t+1} - S_{t+1}) S_{t+1}^{-1} A_{t+1} F_t.
\end{aligned}$$

The algorithm begins with $(c_T|y_{1:T}) \sim N(b_T = f_T, B_T = F_T)$, and then it proceeds *backward* in time to compute $p(c_t|y_{1:T})$, $t = T-1, \dots, 1$ (see Petris et al., 2009).

2.2. Dirichlet Process Mixture Model

Let $(\mathcal{X}, \mathcal{A})$ be a measurable space such that $\mathcal{X}$ is a space and $\mathcal{A}$ is a σ-algebra of subsets of $\mathcal{X}$. Let $\tilde{\alpha} = \alpha G_0$ be a finite non-null measure on $(\mathcal{X}, \mathcal{A})$. Then, a stochastic process G, indexed by elements A of $\mathcal{A}$, is said to be a DP on $(\mathcal{X}, \mathcal{A})$ with parameter $\tilde{\alpha}$ if for any measurable partition $(A_1, \dots, A_k)$ of $\mathcal{X}$, the random vector $(G(A_1), \dots, G(A_k))$ has a Dirichlet distribution with parameter $(\tilde{\alpha}(A_1), \dots, \tilde{\alpha}(A_k))$ (Ferguson, 1973), or equivalently, $(\alpha G_0(A_1), \dots, \alpha G_0(A_k))$ (Rodríguez, 2007). G is a random probability measure on $(\mathcal{X}, \mathcal{A})$, $G(\emptyset)$ is degenerate at 0, $G(\mathcal{X})$ is degenerate at 1, and $G(A)$ takes values only in the interval $[0, 1]$.

An alternative definition of the DP, known as the *stick-breaking* construction, is provided in Sethuraman (1994). A random probability measure given by

$$G = \sum_{j=1}^{\infty} \pi_j \delta_{\theta_j} \tag{6}$$

has a $\mathrm{DP}(\alpha, G_0)$ prior if $\theta_j \sim G_0$ i.i.d and $\pi_j = w_j \prod_{k=1}^{j-1}(1 - w_k)$ with $w_j \sim \mathrm{Beta}(1, \alpha)$ i.i.d. δ_θ denotes a probability measure concentrated at θ, and $\sum_{j=1}^{\infty} \pi_j = 1$ with probability one. This constructive definition of the DP shows that G is discrete, even if G_0 is a continuous distribution. In the context of *mixture models*, a DP is used as prior on mixture components; this model is known as *Dirichlet Process Mixture Model* (Antoniak, 1974). The model applies to a sequence of data $y_1, \dots, y_n$ that are regarded as exchangeable. The distribution from which the y_i are drawn is a mixture of distributions of the form $F(\theta)$, with the mixing distribution over θ being G. The prior for this mixing distribution is a DP with concentration parameter α and baseline distribution G_0. Formally, the DPMM can be represented by the hierarchical model

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha, G_0) \\ \theta_i | G &\sim G \\ y_i | \theta_i &\sim F(\theta_i). \end{aligned} \tag{7}$$

Due to $\sum_{j=1}^{\infty} \pi_j = 1$ with probability one, $\boldsymbol{\pi} = \{\pi_j\}_{j=1}^{\infty}$ can be interpreted as a random probability measure on the positive integers. Let z_i be a indicator variable that specifies the mixture component associated with the observation y_i, such that z_i takes values on $\{1, 2, \dots\}$. Then, the model in (7) can be rewritten as

$$\begin{aligned} G = \sum_{j=1}^{\infty} \pi_j \delta_{\theta_j^\star} \qquad\qquad\qquad\qquad\qquad \boldsymbol{\pi} | \alpha &\sim \mathrm{GEM}(\alpha) \\ \theta_j^\star | G_0 &\sim G_0 \\ z_i | \boldsymbol{\pi} &\sim \boldsymbol{\pi} \\ y_i | z_i, \{\theta_j^\star\}_{j=1}^{\infty} &\sim F(\theta_{z_i}^\star), \end{aligned} \tag{8}$$

where $\mathrm{GEM}(\cdot)$ denotes a *stick-breaking* construction. Note that the parameters θ_i in the model (7) take on values $\theta_j^\star$ with probability π_j. Each value $\theta_j^\star$ defines a *state* of the system, and z_i specifies to which state y_i belongs. The variable z_i is also known as *state* variable.

Posterior inference of the DPMM is typically carried out using either the representation of the DP as a Pólya urn (Blackwell & MacQueen, 1973), or the Blocked Gibbs Sampler (BGS) (Ishwaran & James, 2001), which is based on approximations of the DP by K finite mixture models. The BGS avoids marginalizing over the prior, thus allowing the prior to be directly involved in the Gibbs sampling scheme.

2.3. Hierarchical Dirichlet Process Mixture Model

Let $(\mathcal{X}, \mathcal{A})$ be a measurable space. The hierarchical Dirichlet process (HDP) is a distribution over a set of random probability measures G_j over $(\mathcal{X}, \mathcal{A})$. Each G_j is conditionally independent given G_0, with distribution $G_j \sim \text{DP}(\alpha, G_0)$, where G_0 is a global probability measure distributed as $\text{DP}(\gamma, H_0)$ (Teh et al., 2006). Similar to the DP, the HDP can be used as the prior distribution in a mixture model. Specifically, it splits the data into a number of groups and inside each group the data are subdivided into subgroups; however, clusters characteristics (parameters) are shared among groups. By defining G_0 as a drawn from a DP, the HDP encourages groups to have atoms in common. The model is an immediate extension of the DPMM known as Hierarchical Dirichlet Process Mixture Model (HDPMM).

The HDPMM can be represented by the conditional distributions:

$$G_0|\gamma, H_0 \sim \text{DP}(\gamma, H_0) \qquad\qquad G_j|\alpha, G_0 \sim \text{DP}(\alpha, G_0) \tag{9a}$$

$$\theta_{ji}|G_j \sim G_j \qquad\qquad y_{ji}|\theta_{ji} \sim F(\theta_{ji}), \tag{9b}$$

where j indexes the group and i indexes the observations within a group; θ_{ji} specifies the mixture component associated with the observation y_{ji}. The HDPMM assumes that the observations are exchangeable within each group and also at the group level. Given G_j, the θ_{ji} are conditionally independent. The baseline H_0 provides the prior distribution for the θ_{ji}, and it can be continuous or discrete.

Given that G_0 is itself a draw from a DP, it has a *stick-breaking* representation as in (6). Then, the HDP in equation (9a) can be written as

$$G_0 = \sum_{k=1}^{\infty} \lambda_k \delta_{\theta_k} \qquad\qquad G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta_k}, \tag{10}$$

where θ_k is the parameter of the kth mixture component; $\boldsymbol{\pi}_j = \{\pi_{jk}\}_{k=1}^{\infty}$ and $\boldsymbol{\lambda} = \{\lambda_k\}_{k=1}^{\infty}$ are interpreted as probability measures on the positive integers. Since G_0 is the baseline distribution for each G_j, and G_0 places non-zero mass on the atoms $\boldsymbol{\theta} = \{\theta_k\}_{k=1}^{\infty}$, the atoms of G_j also come from $\boldsymbol{\theta}$. Then, the submodel j shares the same set of mixture components, but has different mixing proportions $\boldsymbol{\pi}_j$. Given that G_j is independent given G_0, the weights $\boldsymbol{\pi}_j$ are conditionally independent given $\boldsymbol{\lambda}$, and $\boldsymbol{\pi}_j \sim \text{DP}(\alpha, \boldsymbol{\lambda})$ (see Teh et al., 2006 for further details).

Let z_{ji} be *mode* variables, such that $\theta_{ji} = \theta_{z_{ji}}$, $z_{ji} \in \{1, 2, \ldots\}$, the HDPMM can be represented by the following conditional distributions:

$$
\begin{aligned}
G_0 &= \sum_{k=1}^{\infty} \lambda_k \delta_{\theta_k} & \boldsymbol{\lambda}|\gamma &\sim \text{GEM}(\gamma) \\
G_j &= \sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta_k} & \boldsymbol{\pi}_j|\alpha, \boldsymbol{\lambda} &\sim \text{DP}(\alpha, \boldsymbol{\lambda}) \\
&& \theta_k|H_0 &\sim H_0 \qquad (11)\\
&& z_{ji}|\boldsymbol{\pi}_j &\sim \boldsymbol{\pi}_j \\
&& y_{ji}|z_{ji}, \{\theta_k\}_{k=1}^{\infty} &\sim F(\theta_{z_{ji}}).
\end{aligned}
$$

Placing a DP prior on G_0 creates a shared and unbounded support for each of the group-specific distributions G_j. Posterior inference in the HDPMM is based in the Pólya urn representation of the DP; Gibbs sampling algorithms are proposed in Teh et al. (2006) and Wang & Wang (2013).

2.4. Sticky Hierarchical Dirichlet Process Mixture Model

One limitation of the HDP prior is that does not differentiate self-transitions from moves between different modes; as a result, the sampling algorithms often create redundant modes and rapidly switch among them. To address this issue, Fox et al. (2011*b*) propose to augment the HDP prior to include a parameter for self-transition bias, and place a separate prior on this parameter. Specifically, they propose to sample transition distributions $\boldsymbol{\pi}_j$ as follows:

$$
\begin{aligned}
\boldsymbol{\lambda}|\gamma &\sim \text{GEM}(\gamma) \\
\boldsymbol{\pi}_j|\alpha, \boldsymbol{\gamma}, \kappa &\sim \text{DP}\left(\alpha + \kappa, \frac{\alpha\boldsymbol{\lambda} + \kappa\delta_j}{\alpha + \kappa}\right),
\end{aligned} \tag{12}
$$

where δ_j is a measure concentrated at j. In equation (12), an amount $\kappa > 0$, named *sticky* parameter, is added to the jth component of $\alpha\boldsymbol{\lambda}$, which increases the expected probability of self-transition by an amount proportional to κ. That is, the expected set of weights for the transition distribution $\boldsymbol{\pi}_j$ is given by:

$$
\text{E}[\pi_{jk}|\boldsymbol{\lambda}, \kappa] = \frac{\alpha}{\alpha + \kappa}\lambda_k + \frac{\kappa}{\alpha + \kappa}\delta(j, k), \tag{13}
$$

where $\delta(j, k)$ is the Kronecker delta, such that $\text{E}[\pi_{jj}|\boldsymbol{\lambda}, \kappa] = \frac{\alpha}{\alpha+\kappa}\lambda_k + \frac{\kappa}{\alpha+\kappa}$.

The Sticky Hierarchical Dirichlet Process Mixture Model (*sticky* HDPMM) is represented by model (11), replacing the prior on $\boldsymbol{\pi}_j$ by (12).

Two inference algorithms for the *sticky* HDP are presented in Fox et al. (2011*b*): (1) *sampling via direct assignments*, which marginalizes over the infinite set of mode-specific transition distributions $\boldsymbol{\pi}_k$ and parameters θ_k, and sequentially sampling z_t given all other mode assignments $z_{-t} = \{z_1, \ldots, z_{t-1}, z_{t+1}, \ldots, z_T\}$, the observations $y_{1:T}$, and the global transition distribution $\boldsymbol{\lambda}$. Each z_t is sampled as if the associated y_t was the last observation. That is, the sampler initializes with z_1 given y_1, $\boldsymbol{\lambda}$, and the hyperparameters; it then samples z_2 given z_1, $y_{1:2}$, $\boldsymbol{\lambda}$, and the hyperparameters, and so on. When a sample $z_{1:T}$ is completed, $\boldsymbol{\lambda}$ and the hyperparameters are updated. This method is a Pólya urn Gibbs sampler that uses a one-coordinate-at-a-time updates for the parameters, similar to the update proposed by Escobar (1988), MacEachern (1994) and Escobar & West (1995) for DPMM, which suffers from slow mixing (see Ishwaran & James, 2001 to review several limitations of Pólya urn Gibbs sampling); (2) *Blocked sampling of mode sequence*, which jointly samples the mode sequence $z_{1:T}$ given the observations $y_{1:T}$, the transition probabilities $\boldsymbol{\pi}_k$, and the parameters θ_k. The sampler is based on a weak limit approximation to the DP prior on $\boldsymbol{\lambda}$ (Ishwaran & Zarepour, 2000; 2002*a*; 2002*b*), which induces a K-finite Dirichlet prior on $\boldsymbol{\pi}_k$, where K is the chosen truncation level. The sampling of $z_{1:T}$ is carried out by a *forward-backward* procedure that first sample z_1 from $p(z_1|y_{1:T}, \boldsymbol{\pi}, \boldsymbol{\theta})$, then conditioned on this value, sample z_2 from $p(z_2|z_1, y_{1:T}, \boldsymbol{\pi}, \boldsymbol{\theta})$, and so on (see Fox et al., 2011*b* for more details).

3. The Model

We start with a hierarchical model of the class of LDSs, which relates an observation y_t[2] to a latent state vector $\boldsymbol{\beta}_t$ through the same dynamic structure and assumptions of model (1):

$$\boldsymbol{\beta}_t = A\boldsymbol{\beta}_{t-1} + \mathbf{e}_t \tag{14a}$$

$$y_t = X_t\boldsymbol{\beta}_t + w_t, \tag{14b}$$

where the evolution error $\mathbf{e}_t \sim N(0, \Sigma_t)$ and the measurement error $w_t \sim N(0, r_t)$ satisfy the same independence assumptions of model (4). The evolution matrix A_t and the design matrix X_t are assumed known.

The LDS is useful in describing data with dynamic parameters; however, numerous of such data are structurally more complex, such that they can not be represented by a single model, but can be well-modeled as switching between a set of dynamic models. That is to say, the dynamic system is divided into segments, each one modeled by a potentially different LDS; separated segments could be described by a single model, but adjacent segments are modeled by different LDSs. These *switches* are specified by a discrete latent variable that identifies the state or mode of the system in each segment. When the latent mode variable is a discrete-time Markov process, the model is referred to as Switching Linear Dynamical System (SLDS).

[2] The general normal LDS is defined for a vector observation.

In a SLDS, each *mode* is associated with a linear dynamic process hierarchically as follows:

$$\begin{align} z_t|z_{t-1} &\sim \boldsymbol{\pi}_{z_{t-1}} \tag{15a}\\ \boldsymbol{\beta}_t &= A^{(z_t)}\boldsymbol{\beta}_{t-1} + \mathbf{e}_t^{(z_t)} \tag{15b}\\ y_t &= X_t\boldsymbol{\beta}_t + w_t^{(z_t)}, \tag{15c}\end{align}$$

where z_t is a Markovian *switching* variable that specifies which LDS is used, i.e., the mode. Given z_t, the evolution error $\mathbf{e}_t^{(z_t)}|z_t \sim N(0, \Sigma^{(z_t)})$ and the observational error $w_t^{(z_t)}|z_t \sim N(0, r^{(z_t)})$ exhibit the same independence assumptions as in the LDS. The LDS and the switching process are related by the dependence of $\{A^{(z_t)}, \Sigma^{(z_t)}, r^{(z_t)}\}$ on z_t. The output at each time, y_t, is determined by stochastically choosing the continuous latent variable, $\boldsymbol{\beta}_t$, using the mode of the discrete latent variable as a *switch*, and then emitting an observation from the corresponding conditional output distribution.

Similar to the inference problems in the LDS, the inference in the SLDS involves computing the posterior distribution of $\boldsymbol{\beta}_t$ and z_t, given the data $y_{1:T}$. Nevertheless, both exact filtered and smoothed inference in the SLDS are numerically intractable, scaling exponentially with time (see Barber, 2012). Fox et al. (2011*a*) propose an extension of the HDPMM of Teh et al. (2006) (described in Section 2.3) for the SLDS, assuming the latent states are mode-specific, while the measurement mechanism is not. That is, the $\boldsymbol{\beta}_t$ depends on z_t through both the evolution matrix $A^{(z_t)}$ and the evolution error $\mathbf{e}^{(z_t)}$, while the y_t is independent on z_t, such that $w_t \sim N(0, r)$. Additionally, the measurement matrix X_t is a shared matrix, such that for all t, $X_t = C$, where C contains fixed intercepts. To carry out inference, the SLDS of Fox et al. (2011*a*) is divided into two main components: the HDPMM and the LDS. The former is approximated by:

$$\begin{aligned} G_0 &= \sum_{k=1}^{K} \lambda_k \delta_{\theta_k} & \boldsymbol{\lambda}|\gamma &\sim \mathrm{Dir}(\gamma/K, \dots, \gamma/K)\\ G_j &= \sum_{k=1}^{K} \pi_{jk} \delta_{\theta_k} & \boldsymbol{\pi}_j|\alpha, \boldsymbol{\lambda}, \kappa &\sim \mathrm{Dir}\left(\alpha\lambda_1, \dots, \alpha\lambda_j + \kappa, \dots, \alpha\lambda_K\right)\end{aligned}$$

$$\begin{aligned} \theta_k|H_0 &\sim H_0 \\ z_t|z_{t-1} &\sim \boldsymbol{\pi}_{z_{t-1}} \\ \boldsymbol{\beta}_t &= A^{(z_t)}\boldsymbol{\beta}_{t-1} + \mathbf{e}_t^{(z_t)}, \end{aligned} \tag{16}$$

where $\theta_k = \{A^{(k)}, \Sigma^{(k)}\}$ and K is the truncation level. As $K \to \infty$, this model converges in distribution to the HDPMM (when $\kappa = 0$) described in Section 2.3, and to its *sticky* extension of Section 2.4 (see Ishwaran & Zarepour, 2002*b*; Teh et al., 2006). The LDS component uses the state equation as a latent process instead of being observable. Specifically,

$$\begin{aligned} \boldsymbol{\beta}_t &= A^{(z_t)}\boldsymbol{\beta}_{t-1} + \mathbf{e}_t^{(z_t)} \\ y_t &= C\boldsymbol{\beta}_t + w_t. \end{aligned} \tag{17}$$

Briefly, the algorithm iterates as follows:

1. Sample $(\boldsymbol{\beta}_{1:T}|z_{1:T}, \theta_k, r)$.
2. Sample $(z_{1:T}|\boldsymbol{\beta}_{1:T}, \boldsymbol{\pi}_k, \theta_k)$.
3. Sample $(\boldsymbol{\pi}_k|z_{1:T}, \boldsymbol{\beta}_{1:T}, y_{1:T})$.
4. Sample $(\theta_k, r|z_{1:T}, \boldsymbol{\beta}_{1:T}, y_{1:T})$.

The SDLS model is a way of modeling discontinuous changes over time in an evolving time series. Those changes are expressed by different values of the state vectors $\boldsymbol{\beta}_t$. The *discontinuity* is due to the DP prior that induces clustering at the θ_k values. In this paper, we propose an extension of the SLDS of Fox et al. (2011*a*) to allow for dependence of y_t on z_t through both the design matrix X_t, by introducing a indicator variable, and the observational error w_t. The model is based on the hypothesis that an explanatory variable can be relevant to a time segment or mode, but may not be significant to others, due to, for example, unanticipated event or an event that is not present through the complete time series. And by making the measurement noise be mode-specific, we allow the variance to change across modes.

The proposed hierarchical model is summarized in equations (18a)-(18d). The model has a level for the indicator variable γ_j, which includes or omits the jth predictor in the observation equation as described below.

$$z_t|z_{t-1} \sim \boldsymbol{\pi}_{z_{t-1}} \tag{18a}$$

$$\gamma_j^{(z_t)} \sim Be(1, p_j) \tag{18b}$$

$$\boldsymbol{\beta}_t = A^{(z_t)}\boldsymbol{\beta}_{t-1} + \mathbf{e}_t^{(z_t)} \tag{18c}$$

$$y_t = X_t^{(z_t)'}\boldsymbol{\beta}_t + w_t^{(z_t)}, \tag{18d}$$

where $X_t^{(z_t)'} = \{1\gamma_1^{(z_t)}x_{t1} \cdots \gamma_p^{(z_t)}x_{tp}\}$ and $\gamma_j^{(k)} = \{0, 1\}$, such that if $\gamma_j^{(k)} = 1$, then the jth predictor x_{tj} is included in the model, and it is excluded if $\gamma_j^{(k)} = 0$. All of the other parameters are defined in equation (15a)-(15c). This model, which we named VS-SLDS, allows us to describe changing relationships between variables as time evolves, an useful issue in practical applications due to the dynamic nature of the time series.

To carry out inference, we proceed by dividing the model in two main components, similar to Fox et al. (2011*a*). We sample the probability measures $\{\boldsymbol{\lambda}, \boldsymbol{\pi}_k\}$ and the hyperparameters $\{\alpha, \kappa, \gamma\}$ involved in model (16) as in Fox et al. (2011*b*). As proposed by Fox et al. (2011*a*), we consider the *Automatic Relevance Determination* (ARD) prior on the dynamic matrix $A^{(k)}$, an inverse-Wishart (IW) on $\Sigma^{(k)}$, and an inverse-Gamma (IG) on the measurement noise $r^{(k)}$. We block-sample the mode and state sequences by using a variant of the forward-backward algorithm. Blocked samplers (Ishwaran & James, 2001) are based on truncated approximations of the DP. They are straightforward to code and tend to have better mixing rate than marginal samplers. However, we block-sample the mode sequence by working on a marginal model. Specifically, we use the predictive distribution of the observations, which involves integration over the state sequence. By using standard results about the multivariate Gaussian distribution (see Section 2.1), we know the predictive distribution is Gaussian, so that it suffices to compute the moments by directly applying laws of the iterated expectation:

$$\begin{aligned}
\mathrm{E}(y_{t+1}|y_{1:t},\boldsymbol{\theta}) &= \mathrm{E}(\mathrm{E}(y_{t+1}|\boldsymbol{\beta}_{t+1},\boldsymbol{\theta})|y_{1:t}) = X_{t+1}^{(z_{t+1})\prime}\mathrm{E}(\boldsymbol{\beta}_{t+1}|y_{1:t},\boldsymbol{\theta}) \\
&= X_{t+1}^{(z_{t+1})\prime} f_{t,t+1} \qquad (19a) \\
\mathrm{Var}(y_{t+1}|y_{1:t},\boldsymbol{\theta}) &= \mathrm{Var}(\mathrm{E}(y_{t+1}|\boldsymbol{\beta}_{t+1},\boldsymbol{\theta})|y_{1:t}) + \mathrm{E}(\mathrm{Var}(y_{t+1}|\boldsymbol{\beta}_{t+1},\boldsymbol{\theta})|y_{1:t}) \\
&= X_{t+1}^{(z_{t+1})\prime}\mathrm{Var}(\boldsymbol{\beta}_{t+1}|y_{1:t},\boldsymbol{\theta})X_{t+1}^{(z_{t+1})} + \mathrm{E}(r^{(z_{t+1})}|y_{1:t}) \\
&= X_{t+1}^{(z_{t+1})\prime} F_{t,t+1} X_{t+1}^{(z_{t+1})} + r^{(z_{t+1})} \qquad (19b)
\end{aligned}$$

where

$$\begin{aligned}
f_{t,t+1} &= A^{(z_{t+1})} f_t^f \\
F_{t,t+1} &= \Sigma^{(z_{t+1})} + A^{(z_{t+1})} F_t^f A^{(z_{t+1})\prime}.
\end{aligned}$$

Note that (19a)-(19b) are analogous to the moments of the one-step-ahead predictive distribution of the Section 2.1.1. The terms f_t^f, F_t^f correspond to the mean vector and the variance matrix, respectively, of the filtering distribution $p(\boldsymbol{\beta}_t|y_{1:t},\boldsymbol{\theta})$ (see Appendix Appendix A for details).

The Kalman filter allows to compute the predictive and filtering distributions recursively, for $t = 1, 2, \dots, T$:

1. Compute the one-step-head predictive distribution for $\boldsymbol{\beta}_t$ given $\{y_{1:t-1}, \boldsymbol{\theta}\}$, based on the filtering distribution $p(\boldsymbol{\beta}_{t-1}|y_{1:t-1},\boldsymbol{\theta})$.
2. Compute the one-step-head predictive distribution $p(y_t|y_{1:t-1},\boldsymbol{\theta})$.
3. Compute the filtering distribution $p(\boldsymbol{\beta}_t|y_{1:t},\boldsymbol{\theta})$ with $p(\boldsymbol{\beta}_t|y_{1:t-1},\boldsymbol{\theta})$ as the prior distribution and the likelihood $p(y_t|\boldsymbol{\beta}_t,)$.

We propose using the predictive distribution $p(y_t|y_{1:t-1},\boldsymbol{\theta})$ directly in inference about the mode sequence. As described in Section 2.4, we use a truncated approximation of the sticky HDP, and then sample forwards in time each z_t from

$$p(z_t|z_{t-1}, y_{1:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) \propto p(z_t|\pi_{z_{t-1}})p(y_t|y_{1:t-1}, z_t, \boldsymbol{\theta})m_{t+1,t}(z_t),$$

where,

$$m_{t,t-1}(z_{t-1}) \propto \begin{cases} \sum_{z_t} p(z_t|\pi_{z_{t-1}})p(y_t|y_{1:t-1}, z_t, \boldsymbol{\theta})m_{t+1,t}(z_t) & t \leq T \\ 1 & t = T+1. \end{cases}$$

Conditioned on the mode sequence $z_{1:T}$, the observations $y_{1:T}$, and the dynamic parameters $\boldsymbol{\theta}$, we recursively sample each $\boldsymbol{\beta}_t$ as in Fox et al. (2011*a*) from

$$p(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1}, z_{1:T}, y_{1:T}, \boldsymbol{\theta}) \propto p(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1}, z_t, \boldsymbol{\theta})p(y_t|\boldsymbol{\beta}_t, z_t, \boldsymbol{\theta})p(y_{t+1:T}|\boldsymbol{\beta}_t, z_{t+1:T}, \boldsymbol{\theta}).$$

The z_t and $\boldsymbol{\beta}_t$ conditional distributions are based on the conditional independence assumption of the model. Explicit details on deriving these conditional distributions are given in appendices Appendix B.1 and Appendix B.2. Another sample scheme based on marginalizing the state sequence is proposed by Fox et al. (2011*a*). This sampler is implemented by sequentially sampling $z_{1:T}$. Although this sampler is computationally intensive, periodically interleaving it with the blocked sampler improve mixing.

Finally, we sample $\gamma_j^{(k)}$ using the variable selection method of Kuo & Mallick (1998). Before describing the sampling scheme, we have to make the following precisions:

1. $\gamma_j^{(k)}$ is determined by the measurements whose mode is k.

2. The index k specifies that the inclusion of the jth predictor is mode-specific.

3. $\gamma_j^{(k)}$, $j = 1, \ldots, p$, are chosen independently, each with Bernoulli prior distribution, $Be(p_j)$, where $p+1$ is the size of state vector, and p_j is the probability to include the jth predictor.

4. We let the intercept term be always included, such that there are 2^p possible submodels, and $\gamma_0^{(k)} = 1$ for all k.

Additionally, we use the following notation:

$$\vartheta_{tj}^{(z_t)} = \beta_{tj}\gamma_j^{(z_t)}, \quad j = 0, 1, \ldots, p; \quad t = 1, \ldots, T,$$
$$\boldsymbol{\vartheta}_t^{(z_t)} = (\vartheta_{t0}^{(z_t)}, \vartheta_{t1}^{(z_t)}, \ldots, \vartheta_{tp}^{(z_t)})' = (\beta_{t0}, \beta_{t1}\gamma_1^{(z_t)}, \ldots, \beta_{tp}\gamma_p^{(z_t)})'.$$

The equation (18d) can then be written as:

$$y_t = X_t'\boldsymbol{\vartheta}_t^{(z_t)} + w_t^{(z_t)}.$$

Based on Kuo & Mallick (1998), we sample variates $\gamma_j^{(k)}$, $j = 1, \ldots, p$, in random order from $Be(\tilde{p}_j^{(k)})$, with $\tilde{p}_j^{(k)} = c_j^{(k)}/(c_j^{(k)} + d_j^{(k)})$, where

$$c_j^{(k)} = p_j^{(k)} \exp\Big\{ -\frac{1}{2r^{(k)}} \sum_{t:\ z_t = k} (y_t - X_t'\boldsymbol{\vartheta}_t^{(z_t)*})^2 \Big\}, \tag{20a}$$

$$d_j^{(k)} = (1 - p_j^{(k)}) \exp\Big\{ -\frac{1}{2r^{(k)}} \sum_{t:\ z_t = k} (y_t - X_t'\boldsymbol{\vartheta}_t^{(z_t)**})^2 \Big\}, \tag{20b}$$

where $\boldsymbol{\vartheta}_t^{(z_t)*}$ is $\boldsymbol{\vartheta}_t^{(z_t)}$ with $\gamma_j^{(z_t)} = 1$, and $\boldsymbol{\vartheta}_t^{(z_t)**}$ is $\boldsymbol{\vartheta}_t^{(z_t)}$ with $\gamma_j^{(z_t)} = 0$. When $\{t : z_t = k\} = \{\emptyset\}$, the update is given by the prior distribution. The sampling algorithm uses the vector X_t to update $\gamma_j^{(k)}$, and it uses $X_t^{(z_t)}$ to update $\boldsymbol{\beta}_t$. Given an initial set $\{\boldsymbol{\pi}, \boldsymbol{\lambda}, \boldsymbol{\theta}, z_{1:T}\}$ and after initialization of the parameters, our Algorithm alternates through the following steps:

1. For each $t = \{T, \dots, 1\}$, sequentially sample z_t as in Fox et al. (2011*a*).

2. Starting with $F_T^b = \frac{1}{r^{(z_T)}} X_T^{(z_T)\prime} X_T^{(z_T)}$ and $f_T^b = \frac{1}{r^{(z_T)}} X_T^{(z_T)\prime} y_T$, for each $t = \{T-1, \dots, 1\}$, compute F_t^b, f_t^b as follows:

$$F_t^b = \frac{1}{r^{(z_t)}} X_t^{(z_t)\prime} X_t^{(z_t)} + F_{t+1,t}^{-1}$$
$$f_t^b = \frac{1}{r^{(z_t)}} X_t^{(z_t)\prime} y_t + F_{t+1,t}^{-1} f_{t+1,t},$$

   where

$$F_{t,t-1}^{-1} = A^{(z_t)\prime} \Sigma^{(z_t)^{-1}} A^{(z_t)} - A^{(z_t)\prime} \Sigma^{(z_t)^{-1}} (\Sigma^{(z_t)^{-1}} + F_t^b)^{-1} \Sigma^{(z_t)^{-1}} A^{(z_t)}$$
$$f_{t,t-1} = F_{t,t-1} (\Sigma^{(z_t)^{-1}} + F_t^b)^{-1} A^{(z_t)\prime} \Sigma^{(z_t)^{-1}} f_t^b.$$

3. For each $t = \{1, \dots, T\}$, sample $\boldsymbol{\beta}_t$:

$$\boldsymbol{\beta}_t \sim N\Big(\mu_{\boldsymbol{\beta}_t}, \Sigma_{\boldsymbol{\beta}_t}\Big)$$
$$\Sigma_{\boldsymbol{\beta}_t} = \big(\Sigma^{(z_t)^{-1}} + F_t^b\big)^{-1}$$
$$\mu_{\boldsymbol{\beta}_t} = \Sigma_{\boldsymbol{\beta}_t} \Big(\Sigma^{(z_t)^{-1}} A^{(z_t)} \boldsymbol{\beta}_{t-1} + f_t^b\Big).$$

4. Starting with with $f_0^f = 0$ and $F_0^f = \mathbf{I}$, for each $t = 1, \dots, T$ compute:

$$f_{t-1,t} = A^{(z_t)} f_{t-1}^f$$
$$F_{t-1,t} = \Sigma^{(z_t)} + A^{(z_t)} F_{t-1}^f A^{(z_t)\prime}$$
$$f_t^f = F_t^f \Big(\frac{1}{r^{z_t}} X_t^{(z_t)\prime} y_t + F_{t-1,t}^{-1} f_{t-1,t}\Big)$$
$$F_t^f = \Big(\frac{1}{r^{(z_t)}} X_t^{(z_t)\prime} X_t^{(z_t)} + F_{t-1,t}^{-1}\Big)^{-1}.$$

5. *Backward.* For each t in $\{T, \dots, 1\}$ and each k in $\{1, \dots, K\}$, compute $m_{t,t-1}(k)$ starting with $m_{T+1,T}(k) = 1$ for all k:

$$m_{t,t-1}(k) = \sum_{j=1}^{K} \pi_{kj} N(\tilde{a}_t^{(j)}, \tilde{Q}_t^{(j)}) m_{t+1,t}(j),$$

   where

$$\tilde{a}_t^{(j)} = X_t^{(j)\prime} f_{t-1,t}$$
$$\tilde{Q}_t^{(j)} = X_t^{(j)\prime} F_{t-1,t} X_t^{(j)} + r^{(j)}.$$

6. *Forward.* For each t in $\{1, \dots, T\}$:

   a) Compute for each k in $\{1, \dots, K\}$:

$$f_k(y_t) = N(\tilde{a}_t^{(k)}, \tilde{Q}_t^{(k)}) m_{t+1,t}(k).$$

b) Sample z_t from:

$$z_t \sim \sum_{k=1}^{K} \pi_{z_{t-1}}(k) f_k(y_t) \delta(z_t, k).$$

7. Update $\{\boldsymbol{\lambda}, \boldsymbol{\pi}\}$ as in Fox et al. (2011*b*), and $\boldsymbol{\theta}$ as in Fox et al. (2011*a*).
8. For each k in $\{1, \dots, K\}$ and $j = 1, \dots, p$, if $\{t\colon z_t = k\} = \{\emptyset\}$, then sample $\gamma_j^{(k)}$ from $B(p_j)$, with p_j the prior belief about the jth predictor. Else, sample $\gamma_j^{(k)}$ from $B(\tilde{p}_j^{(k)})$, with $p_j^{(k)} = c_j^{(k)}/(c_j^{(k)} + d_j^{(k)})$, where $c_j^{(k)}$ and $d_j^{(k)}$ are computed as in equation (20a) and (20b), respectively.

4. Simulation

In this section we analyze three synthetic datasets to highlight the versatility of the model proposed in the previous section. Each dataset is used to evaluate one of three scenarios. For the first and second scenarios, $T = 500$ data were generated from a two-mode SLDS with the following parameters:

1. The evolution matrices $A^{(k)}$, $k = 1, 2$, were set as

$$\mathbf{A}^{(1)} = \begin{bmatrix} 0.85 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}; \quad \mathbf{A}^{(2)} = \begin{bmatrix} 0.85 & 0 & 0 \\ 0 & 0.65 & 0 \\ 0 & 0 & 0.65 \end{bmatrix}$$

2. The covariance matrices $\Sigma^{(k)}$, $k = 1, 2$, were drawn from the prior distribution $\mathrm{IW}(n_0, S_0^{(k)})$, where $n_0 = p + 2$, $S_0^{(1)} = \mathbf{I}_p$, and $S_0^{(2)} = 2\mathbf{I}_p$.
3. The measurement noise precision $1/r^{(k)}$ was drawn from a $\mathrm{Gam}(a_r, b_r)$ prior, with $a_r = 1$ (shape parameter) and $b_r = 2$ (scale parameter).
4. The first scenario has the same number of observations per mode; each mode is observed in 2 non-adjacent subintervals of the same size. The mode sequence $z_{1:T}$ was set as

$$z_t = \begin{cases} 1 & \text{for} \quad t \in [1, 125] \text{ and } t \in [251, 375] \\ 2 & \text{for} \quad t \in [126, 250] \text{ and } t \in [376, 500]. \end{cases}$$

   The second scenario has different number of observations per mode. The mode sequence $z_{1:T}$ was drawn randomly from the prior distribution with large probability of self-transition. That is to say, $z_t \sim \boldsymbol{\pi}_{z_{t-1}}$, $t = 1, \dots, T$, where $\boldsymbol{\pi}_j = [\pi_{j1} \;\; \pi_{j2}]$, $j = 1, 2$, with $\pi_{jj} = 0.95$.

5. The vectors of indicator variables were set as

$$\boldsymbol{\gamma}^{(1)} = (1, 0, 1)$$
$$\boldsymbol{\gamma}^{(2)} = (1, 1, 0),$$

where the first element of each vector is associated with the intercept, and the other elements are associated with explanatory variables.

For the third scenario, $T = 600$ data were generated from a three-mode SLDS with the following parameters:

1. The evolution matrices $A^{(k)}$, $k = 1, 2, 3$, were set as

$$\mathbf{A}^{(1)} = \begin{bmatrix} 0.85 & \mathbf{0} \\ \mathbf{0} & 1\mathbf{I}_3 \end{bmatrix}; \ \mathbf{A}^{(2)} = \begin{bmatrix} 0.85 & \mathbf{0} \\ \mathbf{0} & 0.65\mathbf{I}_3 \end{bmatrix}; \ \mathbf{A}^{(3)} = \begin{bmatrix} 0.85 & \mathbf{0} \\ \mathbf{0} & 0.85\mathbf{I}_3 \end{bmatrix};$$

2. The covariance matrices $\Sigma^{(k)}$, $k = 1, 2$, were drawn from the prior distribution $\mathrm{IW}(n_0, S_0^{(k)})$, where $n_0 = p+2$, $S_0^{(1)} = S_0^{(3)} = 2\mathbf{I}_p$, and $S_0^{(2)} = \mathbf{I}_p$.

3. The measurement noise precision $1/r^{(k)}$ was drawn from a $\mathrm{Gam}(a_r, b_r)$ prior, with $a_r = 1$ (shape parameter) and $b_r = 2$ (scale parameter).

4. Each mode has the same number of observations, and is observed in one subinterval on the whole series.

5. Each $\gamma_j^{(k)}$, $k = 1, 2, 3$; $j = 1, 2, 3$, was drawn from a $Be(1, p)$, with $p \sim U(0, 1)$. The resulting vectors of indicator variables are

$$\begin{aligned} \boldsymbol{\gamma}^{(1)} &= (1, 0, 1, 1) \\ \boldsymbol{\gamma}^{(2)} &= (1, 0, 0, 1) \\ \boldsymbol{\gamma}^{(3)} &= (1, 1, 1, 0) \end{aligned}$$

The measurements for the three scenarios were simulated as follows:

1. The design matrix is build such that, for each $t = 1, \dots, T$ and $j = 1, \dots, p$:

$$\begin{aligned} &x_{tj} = \upsilon x_{t-1,j} + h_{tj}, \quad h_{tj} \sim \mathrm{N}(0, 1) \\ &x_{0j} = 0 \quad \forall j, \end{aligned}$$

where $\upsilon = 1$ for scenario 1, and $\upsilon = 0.9$ for scenarios 2 and 3. This choice of predictors as correlated adjacent points is due to the dynamic nature of the time series exhibiting dependence or time correlations between adjacent points in time.

2. For each $t = 1, \dots, T$, $\boldsymbol{\beta}_t$ and y_t satisfy the following set of equations:

$$\begin{aligned} \boldsymbol{\beta}_t &= A^{(z_t)}\boldsymbol{\beta}_{t-1} + e_t, & e_t &\sim \mathrm{N}(\mathbf{0}, \Sigma^{(z_t)}), \boldsymbol{\beta}_0 = \mathbf{0} \\ y_t &= \sum_{j=0}^{p} \beta_{tj}\gamma_j^{(z_t)} x_{tj} + w_t, & w_t &\sim \mathrm{N}(0, r^{(z_t)}), x_{t0} = 1. \end{aligned}$$

We employ the algorithm described in the previous section to investigate the accuracy of the model (18a)-(18d) in its ability to infer the simulated segmentation and model fitting. We use a truncation level of $K = 30$, an ARD prior distribution on $A^{(k)}$, IW prior on $\Sigma^{(k)}$, and IG prior on $r^{(k)}$, with the following initial values for the hyperparameters for each $k = 1, \dots, K$:

1. We place independent $N(\mathbf{0}, 1/\alpha_j^{(k)}\mathbf{I}_p)$ priors on the columns of the matrix $A^{(k)}$, where $\alpha_j^{(k)} \sim \text{Gam}(1, 0.01)$, such that $\text{E}(\alpha_j^{(k)}) = 100$.

2. We place a $\text{IW}(p + 2, 0.1\mathbf{I}_p)$ prior on $\Sigma^{(k)}$.

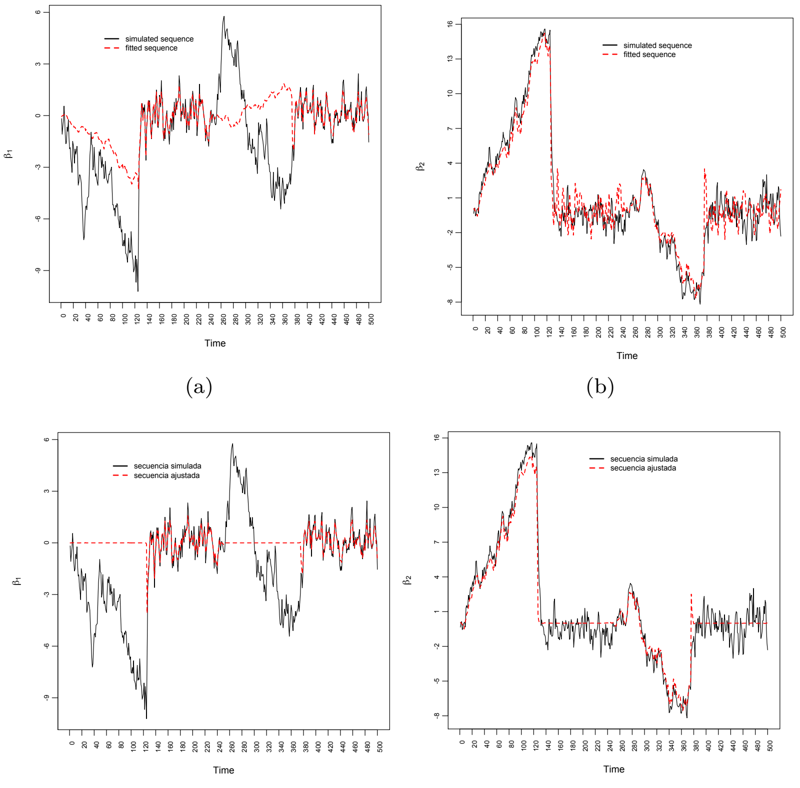
3. We place a $\text{Gam}(1, 0.1)$ prior on precision $1/r^{(k)}$, such that $\text{E}(1/r^{(k)}) = 10$.

The vectors of indicator variables are initialized with the full model $\gamma_j^{(k)} = 1$ for all j, and an equal probability of transition, $1/K$, to each mode. The hyperparameters $\{\alpha, \kappa, \gamma\}$ were set as in Fox et al. (2011*b*). All results are based on 15,000 iterations, obtained after a burn-in period of 5,000 samples.

4.1. Scenario 1

In Figure 1(a)-1(b) we show the slope components of the state vector fitted by our model (dash lines) along with the simulated points (continuous lines). As expected, the estimated sequences closely approximate the simulated ones in the intervals in which the explanatory variable is significant. This fact is clearer in Figure 1(c)-1(d), in which we show the simulated slope components (continuous lines) along with the corresponding posterior components of $\boldsymbol{\vartheta}_t^{(z_t)}$ (dash lines). We see that the VS-SLDS is able to correctly identify the significant variables.

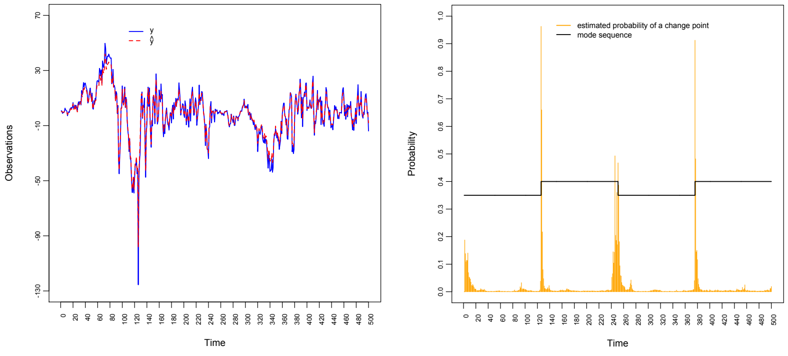
We display in Figure 2(a) the simulated data sequence (continuous line) and the fitted curve (dash line). Although it is well-known that a model that overfits the data typically leads to poor prediction performance, currently we are mainly interested in inferring about the number of dynamical modes and verifying if it correctly identifies the membership of the observations in each mode. However, we have to face the issue of summarizing the results from the posterior distribution over modes. The so-called *label switching* problem, used by Redner & Walker (1984) to describe the invariance of the likelihood under relabelling of the mixture components, makes it difficult to estimate quantities of interest by their posterior mean (see Stephens, 2000 for more details about this issue). In order to show the capability of the model to identify the correct grouping of observations, we display the posterior probability of a mode switching for each t in Figure 2(b). We can see that the higher probabilities match with sequence changes. Additionally, we display the number of modes to which at least 30% of observations are assigned (Figure 3); we can see that the observations are mainly clustered into 2 groups.



(a) (b)

(c) (d)

FIGURE 1: Scenario 1. (a) β_{t1}, $\hat{\beta}_{t1}$; (b) β_{t2}, $\hat{\beta}_{t2}$; (c) β_{t1}, $\hat{\vartheta}_{t1}^{(z_t)}$; (d) β_{t2}, $\hat{\vartheta}_{t2}^{(z_t)}$.



(a) (b)

FIGURE 2: Scenario 1. (a) Simulated and fitted sequences ($\hat{y}_t = X_t'\hat{\boldsymbol{\vartheta}}_t$, $t = 1, \ldots, 500$); (b) mode sequence and estimated probability of a change point for each t.

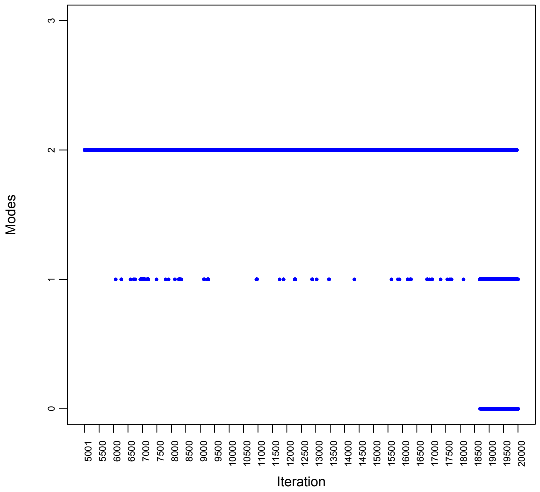


Figure 3: Scenario 1. Significant modes.

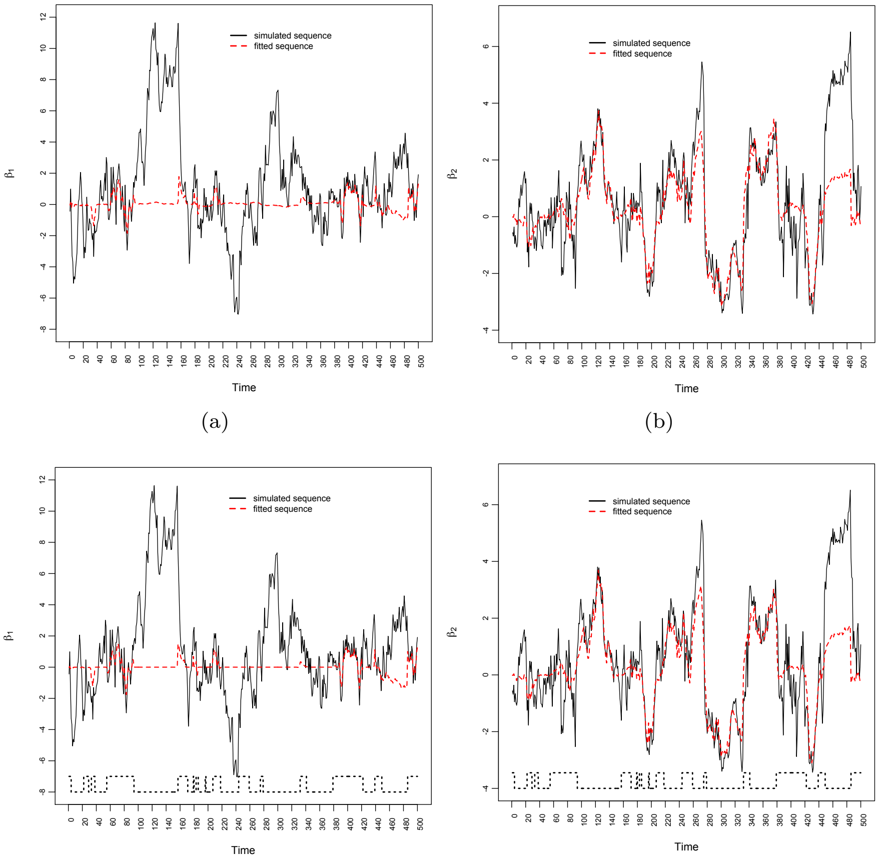
4.2. Scenario 2

The plots of Figure 4 show that the model is able to identify the relevant variables on each mode even if the mode has few data. We include the simulated mode sequence (dotted black line) in Figures 4(c) and 4(d) in order to provide a better understanding. We can see that the model has a good performance to identify the correct explanatory variables for each mode even with lower temporal mode persistence (Figures 4(c)-4(d)). This result is due to that the higher probabilities of a change point match with the true changes in the mode sequence (Figure 5(a)). Consequently, the data are mainly clustered into two modes (Figure 6), and the fitted values are accurate (Figure 5(b)).

4.3. Scenario 3

The results depicted in Figure 7 confirm good performance of the VS-SLDS to identify the significant variables on data generated from a model with three modes. The simulated slope components of the state vector are compared with the fitted ones (left panel), and with the corresponding fitted components $\hat{\vartheta}_{tj}^{(z_t)}$ (right panel), which is zero in the interval where the predictor is simulated not significant.

We display in Figure 8 the number of modes to which at least 25% of observations are assigned. Around 82% of iterations the data are clustered into three groups. We show in Figure 9(b) the estimated probability of a change point for each t. We notice that two change points with high probability divide the sequence into three sections. In real data without a priori knowledge of the number of modes, the probability of a change point does not give complete information by itself. By looking at the number of modes to which a significant number of observations are assigned, the estimated sequence of $z_{1:T}$, and the change point probabilities, we are able to infer the number of modes and where the mode changes originate. In Figure 9(b) we include the mode sequence of the iteration 16100[3] to have an idea about the estimated sequence. We display in Figure 9(a) the simulated and fitted data sequences.



(a)

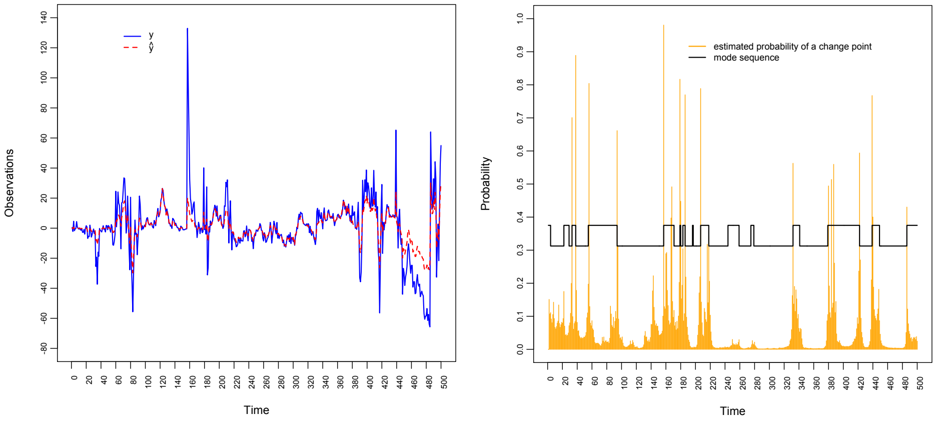
(b)

(c)

(d)

FIGURE 4: Scenario 2. (a) β_{t1}, $\hat{\beta}_{t1}$; (b) β_{t2}, $\hat{\beta}_{t2}$; (c) β_{t1}, $\hat{\vartheta}_{t1}^{(z_t)}$; (d) β_{t2}, $\hat{\vartheta}_{t2}^{(z_t)}$.

[3] We randomly chosen a mode sequence after of the iteration 5000.



(a) (b)

FIGURE 5: Scenario 2. (a) Simulated and fitted sequences ($\hat{y}_t = X_t'\hat{\boldsymbol{\vartheta}}_t$, $t = 1, \ldots, 500$); (b) mode sequences and estimated probability of a change point for each t.

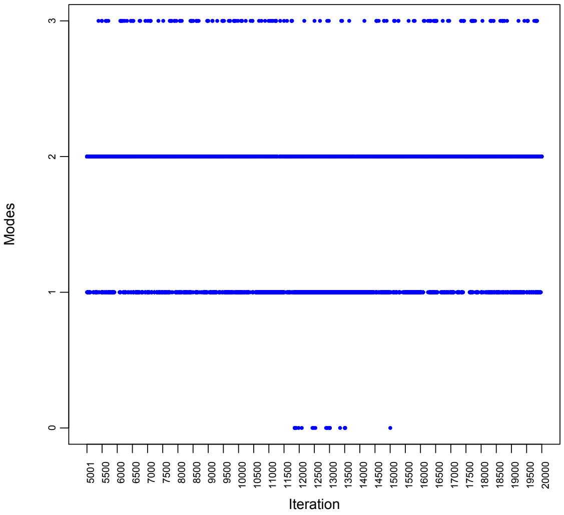
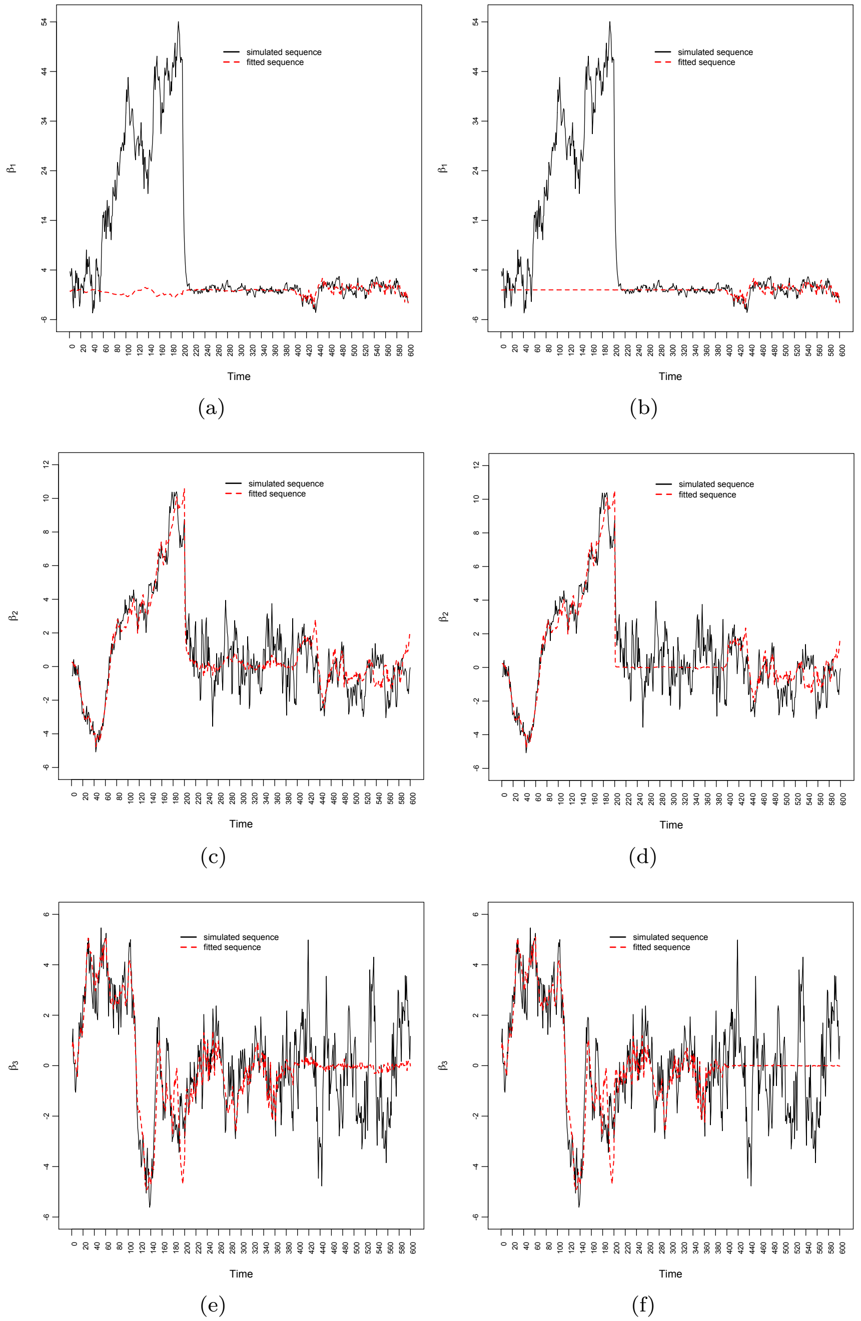


FIGURE 6: Scenario 2. Significant modes.



(a) (b)

(c) (d)

(e) (f)

Figure 7: Scenario 3. Simulated slope components of the state vector vs. fitted sequence (left panel) and components $\hat{\vartheta}_{tj}^{(z_t)}$ (right panel), $j = 1, 2, 3$.

5. Applications

The proposed method is on the real dataset of meteorological series of Barranquilla, Colombia, reported by the weather station Ernesto Cortissoz. The dataset can be downloaded from the National Centers for Environmental Information (2021). We have selected five series in all, including relative humidity, atmospheric pressure, visibility, wind speed and temperature. All the data are monthly average value ranging from January 1995 to December 2020, resulting in 295 months totally (there are 17 missing values).

In the first instance, we consider the temperature as observations y_t and all of the rest of variables in the design matrix X_t. In other example we consider the visibility as observations y_t and all of the rest of variables in the design matrix X_t. We assumed the hierarchical model in equations (18a)-(18d) to describe the dynamical behavior of both observations series by employing the algorithm described in Section 3. We use a truncation level of $K = 30$ and the following settings for each $k = 1, \dots, K$: independent $N(\mathbf{0}, 1/\alpha^{(k)}\mathbf{I}_p)$ priors on the columns of the matrix $A^{(k)}$, where $\alpha^{(k)} \sim \text{Gam}(1, 0.01)$; a $\text{IW}(p + 2, 0.1\mathbf{I}_p)$ prior on $\Sigma^{(k)}$; a $\text{Gam}(1, 0.1)$ prior on precision $1/r^{(k)}$. The vectors of indicator variables are initialized with the full model $\gamma_j^{(k)} = 1$ for all j, and an equal probability of transition, $1/K$, to each mode. The hyperparameters $\{\alpha, \kappa, \gamma\}$ were set as in Fox et al. (2011*b*). All results are based on 30,000 iterations, obtained after a burn-in period of 5,000 samples.

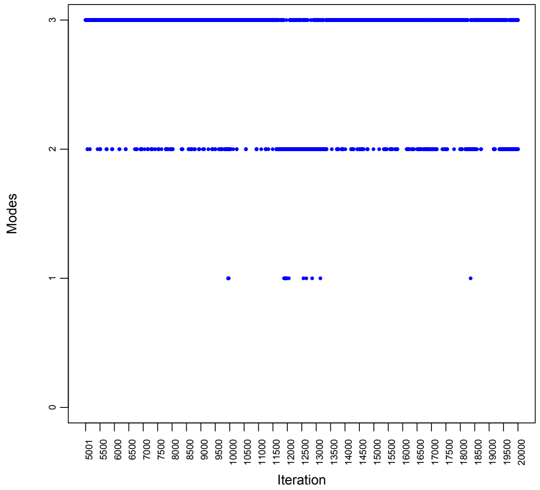
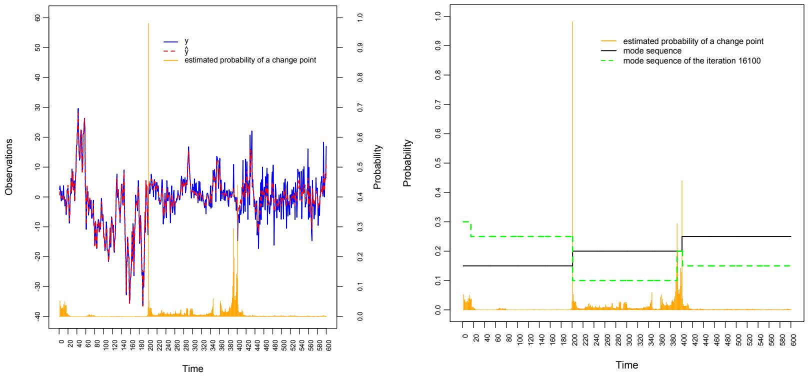


Figure 8: Scenario 3. Significant modes.

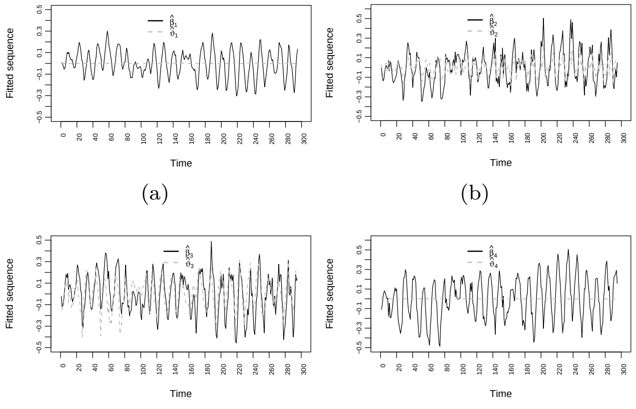


(a) (b)

FIGURE 9: Scenario 3. (a) Simulated and fitted sequences ($\hat{y}_t = X_t' \hat{\boldsymbol{\vartheta}}_t$, $t = 1, \ldots, 500$); (b) mode sequences and the estimated probability of a change point for each t.

5.1. Temperature Data

Temperature data are frequently related to meteorological parameters such as relative humidity, precipitation, wind speed, insolation, atmospheric pressure and so on (Han et al., 2015). In this instance, the design matrix is build such that $\boldsymbol{x}_1 =$ humidity , $\boldsymbol{x}_2 =$ atmospheric pressure, $\boldsymbol{x}_3 =$ visibility and $\boldsymbol{x}_4 =$ wind speed. The Figures 10(b) and 10(c) show that the model identifies two relevant variables for the temperature data: atmospheric pressure and visibility, which is described by the posterior of ϑ_{tj}, $t = 1, \ldots, 295$, $j = 2, 3$. On the other hand, in Figures 10(a) and 10(d) the fitted sequence of ϑ_{tj}, $t = 1, \ldots, 295$, $j = 1, 4$, is almost zero, which means that the variable j is absent.



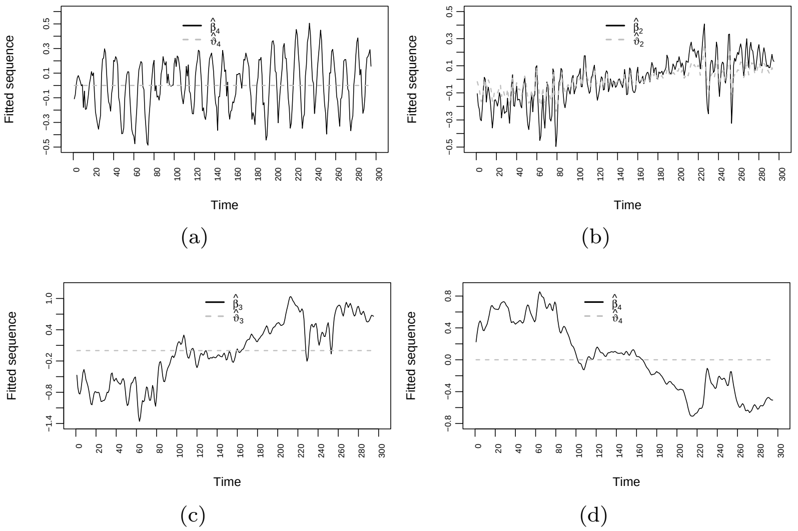
(a) (b)

(c) (d)

FIGURE 10: Fitted sequences of the state vectors $\boldsymbol{\beta}_j$ and $\boldsymbol{\vartheta}_j = \boldsymbol{\beta}_j \gamma_j$ for the temperature data.

5.2. Visibility Data

The study of visibility (the horizontal distance an object can be seen and identified) is of interest in a variety of fields including aeronautics (Watson et al., 2009), environment (Du et al., 2013; Tsai et al., 2007) and human health ((Huang et al., 2009; Thach et al., 2010; Majewski et al., 2015). Visibility is typically related to meteorological parameters and air pollution. In this instance, the design matrix is build such that $\boldsymbol{x}_1$ = humidity, $\boldsymbol{x}_2$ = atmospheric pressure, $\boldsymbol{x}_3$ = temperature and $\boldsymbol{x}_4$ = wind speed. The Figure 11 show that the model identifies the atmospheric pressure as the only relevant variable for the visibility data, which is described by the posterior of ϑ_{t2}, $t = 1, \ldots, 295$. In Figures 11(a), 11(c) and 11(d) the fitted sequence of ϑ_{tj}, $t = 1, \ldots, 295$, $j = 1, 3, 4$, shows that the variable j is absent. Influence of the atmospheric pressure on the visibility has been shown in the literature (see for example Tsai et al. 2007 and Majewski et al. 2011).



(a)

(b)

(c)

(d)

FIGURE 11: Fitted sequences of the state vectors $\boldsymbol{\beta}_j$ and $\boldsymbol{\vartheta}_j = \boldsymbol{\beta}_j \gamma_j$ for the visibility data.

6. Discussion

In this paper, we propose an extension of the SLDS. We make the measurement noise and mean mode-specific. The extension uses the predictive distribution of the observations, which involves integration over the latent states, and a method to select relevant explanatory variables among modes. By allowing the measurement noise be mode-specific we obtained a flexible model for complex time series where the stationarity assumption is not valid. Additionally, by selecting a mode-specific set of explanatory variables allows us to describe changing relationships between variables as time evolves. Our proposed model is a flexible dynamic regression model for learning about the number of modes, that is to say, it accurately helps finding out the appropriate number of distributions in a mixture of distributions. Since the modes divide the time series into non-overlapping groups, the proposed method helps finding the location of the modes, and within each mode, it identifies the significant explanatory variables and accurately estimate the regression coefficients. We judge the capability of the proposed model by examining three synthetic datasets. We note that the model performs well even with lower temporal mode persistence. We present two applications using a real dataset of meteorological data from Barranquilla, Colombia. The results are consistent with the literature. We conclude by addressing two avenues for further research: (1) to extend the model to include a penalty term to prevent overfitting, as proposed by West & Harrison (1997) for LDS, and as a consequence of it, (2) study the model's ability for predictions.

[Received: February 2020 — Accepted: December 2021]

References

Antoniak, C. (1974), 'Mixtures of dirichlet processes with applications to bayesian nonparametric problems', *The Annals of Statistics* **2**(6), 1152–1174.

Barber, D. (2012), *Bayesian Reasoning and Machine Learning*, Cambridge University Press.

Bishop, C. (2006), *Pattern Recognition and Machine Learning*, Springer.

Blackwell, D. & MacQueen, J. (1973), 'Ferguson distributions via Polya urn schemes', *The Annals of Statistics* **1**(2), 353–355.

Bregler, C. (1997, June), Learning and recognizing human dynamics in video sequences, *in* 'Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition', pp. 568–574.

Carvalho, C. & Lopes, H. (2007), 'Simulation based sequential analysis of markov switching stochastic volatility models', *Computational Statistics and Data Analysis* **51**, 4526–4542.

Du, K., Mu, C., Deng, J. & Yuan, F. (2013), 'Study on atmospheric visibility variations and the impacts of meteorological parameters using high temporal resolution data: an application of environmental internet of things in china', *International Journal of Sustainable Development and World Ecology* **20**(3), 238–247.

Escobar, M. (1988), Estimating the Means of Several Normal Populations by Nonparametric Estimation of the Distribution of the Means, PhD thesis, Yale University.

Escobar, M. & West, M. (1995), 'Bayesian density estimation and inference using mixtures', *Journal of the American Statistical Association* **90**(430), 577–588.

Ferguson, T. (1973), 'A bayesian analysis of some nonparametric problems', *The Annals of Statistics* **1**(2), 209–230.

Fox, E., Sudderth, E., Jordan, M. & Willsky, A. (2011*a*), 'Bayesian nonparametric inference of switching dynamic linear models', *IEEE Transactions on signal processing* **59**(4), 1569–1585.

Fox, E., Sudderth, E., Jordan, M. & Willsky, A. (2011*b*), 'A sticky hdp-hmm with application to speaker diarization', *The Annals of Applied Statistics* **5**(2A), 1020–1056.

Han, M., Ren, W. & Liu, X. (2015), 'Joint mutual information-based input variable selection for multivariate time series modeling', *Engineering Applications of Artificial Intelligence* **37**, 250–257.

Huang, W., Tan, J., Kan, H., Zhao, N., Song, W., Song, G., Chen, G., Jiang, L., Jiang, C., Chen, R. & Chen, B. (2009), 'Visibility, air quality and daily mortality in shanghai, china', *Science of The Total Environment* **407**(10), 3295–3300.

Huerta, G., Sansó, B. & Stroud, J. R. (2004), 'A spatiotemporal model for mexico city ozone levels', *Journal of the Royal Statistical Society* **53**(2), 231–248.

Ishwaran, H. & James, L. (2001), 'Gibbs sampling methods for stick-breaking priors', *Journal of the American Statistical Association* **96**(453), 161–173.

Ishwaran, H. & James, L. (2002), 'Approximate dirichlet process computing in finite normal mixtures: Smoothing and prior information', *Journal of Computational and Graphical Statistics* **11**(3), 1–26.

Ishwaran, H. & Zarepour, M. (2000), 'Markov chain monte carlo in approximate dirichlet and beta two-parameter process hierarchical models', *Biometrika* **87**(2), 371–390.

Ishwaran, H. & Zarepour, M. (2002*a*), 'Dirichlet prior sieves in finite normal mixtures', *Statistica Sinica* **12**(3), 941–963.

Ishwaran, H. & Zarepour, M. (2002*b*), 'Exact and approximate sum representations for the dirichlet process', *The Canadian Journal of Statistics* **30**(2), 269–283.

Kalman, R. (1960), 'A new approach to linear filtering and prediction problems', *Journal of Basic Engineering* **82**, 35–45.

Kalman, R. (1963), 'Mathematical description of linear dynamical systems', *Journal of the Society for Industrial and Applied Mathematics* **1**(2), 152–192.

Kim, C. (1994), 'Dynamic linear models with markov switching', *Journal of Econometrics* **60**(1-2), 1–22.

Kuo, L. & Mallick, B. (1998), 'Variable selection for regression models', *The Indian Journal of Statistics. Special Issue on Bayesian Analysis* **60**(1), 65–81.

Lamon III, E., Carpenter, S. & Stow, C. (1998), 'Forecasting PCB concentrations in Lake Michigan salmonids: a dynamic linear model approach', *Ecological Applications* **8**(3), 659–668.

MacEachern, S. N. (1994), 'Estimating normal means with a conjugate style dirichlet process prior', *Communications in Statistics-Simulation and Computation* **23**(3), 727–741.

Majewski, G., Kleniewska, M. & Brandyk, A. (2011), 'Seasonal variation of particulate matter mass concentration and content of metals', *Polish Journal of Environmental Studies* **20**(2), 417–427.

Majewski, G., Rogula-Kozłowska, W., Czechowski, P. O., Badyda, A. & Brandyk, A. (2015), 'he impact of selected parameters on visibility: First results from a long-term campaign in warsaw, poland', *Atmosphere* **6**, 1154–1174.

McAlinn, K. & West, M. (2016), Dynamic bayesian predictive synthesis in time series forecasting, Technical report, Duke University.

Meinhold, R. & Singpurwalla, N. (1983), 'Understanding the kalman filter', *The American Statistician* **37**(2), 123–127.

National Centers for Environmental Information (2021), 'Local climatological data'. https://www.ncei.noaa.gov/data/local-climatological-data/

Pavlović, V., Rehg, J. & MacCormick, J. (2001), Learning switching linear models of human motion., *in* 'Advances in Neural Information Processing Systems', Vol. 13, Neural Information Processing Systems (NIPS) 2000.

Petris, G., Petrone, S. & Campagnoli, P. (2009), *Dynamic Linear Models with R*, Springer-Verlag.

Rauch, H., Striebel, C. & Tung, F. (1965), 'Maximum likelihood estimates of linear dynamic systems', *AIAA Journal* **3**(8), 1445–1450.

Redner, R. & Walker, H. (1984), 'Mixture densities, maximum likelihood and the em algorithm', *SIAM Review* **26**(2), 195–239.

Rodríguez, A. (2007), Some Advances in Bayesian Nonparametric Modeling, PhD thesis, Duke University.

Sethuraman, J. (1994), 'A constructive definition of dirichlet priors', *Statistica Sinica* **4**, 639–650.

Stephens, M. (2000), 'Dealing with label switching in mixture models', *Journal of the Royal Statistical Society* **62**(4), 795–809.

Teh, Y. W., Jordan, M. I., Beal, M. J. & Blei, D. M. (2006), 'Hierarchical dirichlet processes', *Journal of the American Statistical Association* **101**, 1566–1581.

Thach, T.-Q., Wong, C.-M., Chan, K.-P., Chau, Y.-K., Chung, Y.-N., Ou, C.-Q., Yang, L. & Hedley, A. J. (2010), 'Daily visibility and mortality: Assessment of health benefits from improved visibility in hong kong', *Environmental Research* **110**(6), 617–623.

Tsai, Y., Kuo, S.-C., Lee, W.-J., Chen, C.-L. & Chen, P.-T. (2007), 'Long-term visibility trends in one highly urbanized, one highly industrialized, and two rural areas of taiwan', *Science of The Total Environment* **382**(2-3), 324–341.

Velasco-Cruz, C., Leman, S. C., Hudy, M. & Smith, E. P. (2012), 'Assessing the risk of rising temperature on brook trout: a spatial dynamic linear risk model', *Journal of Agricultural, Biological, and Environmental Statistics* **17**(2), 246–264.

Wang, L. & Wang, X. (2013), 'Hierarchical dirichlet process model for gene expression clustering', *EURASIP Journal on Bioinformatics and Systems Biology* **1**(5).

Watson, A., Ramirez, C. & Salud, E. (2009), 'Predicting visibility of aircraft', *PLOS ONE* **5**(7), 1–16.

West, M. (2013), *Bayesian Dynamic Modelling*, Oxford University Press, chapter 8.

West, M. & Harrison, J. (1997), *Bayesian Forecasting and Dynamic Models*, 2 edn, Springer.

Zeng, Y. & Wu, S., eds (2013), *State-space models. Applications in Economics and Finance*, Springer.

Appendix A. Filtering Distribution $p(\boldsymbol{\beta}_t|y_{1:T},\boldsymbol{\theta})$

$$\begin{aligned}
p(\boldsymbol{\beta}_t|y_{1:t},\boldsymbol{\theta}) &\propto p(y_t|\boldsymbol{\beta}_t,\boldsymbol{\theta})p(\boldsymbol{\beta}_t|y_{1:t-1},\boldsymbol{\theta})\\
&\propto \exp\Big\{-\frac{1}{2}\boldsymbol{\beta}_t'\Big(\frac{1}{r^{(z_t)}}X_t^{(z_t)\prime}X_t^{(z_t)}+F_{t-1,t}^{-1}\Big)\boldsymbol{\beta}_t+\\
&\qquad\qquad \boldsymbol{\beta}_t'\Big(\frac{1}{r^{(z_t)}}X_t^{(z_t)\prime}y_t+F_{t-1,t}^{-1}f_{t-1,t}\Big)\Big\}\\
&\propto N(f_t^f,F_t^f),
\end{aligned}$$

where

$$\begin{aligned}
f_t^f &= F_t^f\Big(\frac{1}{r^{z_t}}X_t^{(z_t)\prime}y_t+F_{t-1,t}^{-1}f_{t-1,t}\Big)\\
F_t^f &= \Big(\frac{1}{r^{(z_t)}}X_t^{(z_t)\prime}X_t^{(z_t)}+F_{t-1,t}^{-1}\Big)^{-1}.
\end{aligned}$$

Appendix B. Forward-Backward Procedure for Sampling the Mode and State Sequences

We provide the derivations of the expressions to jointly sample the dynamic *modes* $z_{1:T}$ as well as the *state* sequence $\boldsymbol{\beta}_{1:T}$. The expressions apply for both the SLDS (15a)-(15c) and VS-SLDS (18a)-(18d), we simply replace X_t with $X_t^{(z_t)}$. For compactness, we omit the dependency of dynamic parameters $\boldsymbol{\theta}$ and design matrix X_t on z_t.

Appendix B.1. Sampling $z_{1:T}$

The joint distribution of $z_{1:T}$, given the observation sequence, can be decomposed as follows:

$$\begin{aligned} p(z_{1:T}|y_{1:T},\boldsymbol{\pi},\boldsymbol{\theta}) =& p(z_T|z_{T-1},y_{1:T},\boldsymbol{\pi},\boldsymbol{\theta})\ p(z_{T-1}|z_{T-2},y_{1:T},\boldsymbol{\pi},\boldsymbol{\theta}) \\ & \cdots p(z_2|z_1,y_{1:T},\boldsymbol{\pi},\boldsymbol{\theta})\ p(z_1|y_{1:T},\boldsymbol{\pi},\boldsymbol{\theta}). \end{aligned}$$

We can use a forward-backward procedure for jointly sample the mode sequence, as in Section 2.1. We first sample z_1 from $p(z_1|y_{1:T},\boldsymbol{\pi},\boldsymbol{\theta})$, then we sample z_2 conditioning on z_1 from $p(z_2|z_1,y_{1:T},\boldsymbol{\pi},\boldsymbol{\theta})$, and so on. For each t, the conditional distribution is given by:

$$\begin{aligned} & p(z_t|z_{t-1},y_{1:T},\boldsymbol{\pi},\boldsymbol{\theta}) \\ & \quad \propto p(z_t|\pi_{z_{t-1}})p(y_1|z_t,\boldsymbol{\pi},\boldsymbol{\theta})p(y_2|y_1,z_t,\boldsymbol{\pi},\boldsymbol{\theta})p(y_3|y_{1:2},z_t,\boldsymbol{\pi},\boldsymbol{\theta}) \\ & \qquad \cdots p(y_t|y_{1:t-1},z_t,\boldsymbol{\pi},\boldsymbol{\theta})p(y_{t+1}|y_{1:t},z_t,\boldsymbol{\pi},\boldsymbol{\theta})\cdots p(y_T|y_{1:T-1},z_t,\boldsymbol{\pi},\boldsymbol{\theta}) \\ & \quad \propto p(z_t|\pi_{z_{t-1}})\prod_{i=1}^{t} p(y_i|y_{1:i-1},z_t,z_i,\boldsymbol{\pi},\boldsymbol{\theta}) \cdot \prod_{j=t+1}^{T} p(y_j|y_{1:j-1},z_t,z_j,\boldsymbol{\pi},\boldsymbol{\theta}) \\ & \quad \propto p(y_1,\dots,y_t,z_t|z_{1:t-1},\boldsymbol{\pi},\boldsymbol{\theta}) \cdot p(y_{t+1},\dots,y_T|z_t,z_{t+1:T},\boldsymbol{\pi},\boldsymbol{\theta}). \end{aligned}$$

If $t=1$, then:

$$\begin{aligned} p(z_1|y_{1:T},\boldsymbol{\pi},\boldsymbol{\theta}) & \propto p(y_1,z_1|\boldsymbol{\pi},\boldsymbol{\theta}) \cdot p(y_2,\dots,y_T|z_1,z_{2:T},\boldsymbol{\pi},\boldsymbol{\theta}) \\ & \propto p(z_1)p(y_1|z_1,\boldsymbol{\pi},\boldsymbol{\theta}) \cdot \sum_{z_2} p(z_2|\pi_{z_1})p(y_2|y_1,z_2,\boldsymbol{\pi},\boldsymbol{\theta})p(y_{3:T}|z_{3:T},\boldsymbol{\pi},\boldsymbol{\theta}) \\ & \propto p(z_1)p(y_1|z_1,\boldsymbol{\theta})m_{2,1}(z_1). \end{aligned}$$

When $t=2$:

$$\begin{aligned} & p(z_2|z_1,y_{1:T},\boldsymbol{\pi},\boldsymbol{\theta}) \\ & \quad \propto p(y_1,y_2,z_2|z_1,\boldsymbol{\pi},\boldsymbol{\theta}) \cdot p(y_3,\dots,y_T|z_2,z_{3:T},\boldsymbol{\pi},\boldsymbol{\theta}) \\ & \quad \propto p(z_2|\pi_{z_1})p(y_1|z_2,\boldsymbol{\pi},\boldsymbol{\theta})p(y_2|y_1,z_2,\boldsymbol{\pi},\boldsymbol{\theta}) \\ & \qquad\qquad \cdot \sum_{z_3} p(z_3|\pi_{z_2})p(y_3|y_{1:2},z_3,\boldsymbol{\pi},\boldsymbol{\theta})p(y_{4:T}|z_{4:T},\boldsymbol{\pi},\boldsymbol{\theta}) \\ & \quad \propto p(z_2|\pi_{z_1})p(y_2|y_1,z_2,\boldsymbol{\theta})m_{3,2}(z_2). \end{aligned}$$

In general, the conditional distribution of z_t, for all t, can be decomposed as follows:

$$p(z_t|z_{t-1}, y_{1:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) \propto p(z_t|\pi_{z_{t-1}})p(y_t|y_{1:t-1}, z_t, \boldsymbol{\theta})m_{t+1,t}(z_t),$$

where,

$$m_{t,t-1}(z_{t-1}) \propto \begin{cases} \sum_{z_t} p(z_t|\pi_{z_{t-1}})p(y_t|y_{1:t-1}, z_t, \boldsymbol{\theta})m_{t+1,t}(z_t) & t \leq T \\ 1 & t = T+1. \end{cases}$$

The term $m_{t,t-1}(z_{t-1})$ is known as backward message passed from z_t to z_{t-1} in the *machine learning* literature (see e.g. Bishop, 2006).

Appendix B.2. Sampling $\boldsymbol{\beta}_t$

A similar forward-backward procedure can be used to sample the latent state sequence $\boldsymbol{\beta}_{1:T}$. We begin by considering

$$\begin{aligned} m_{t,t-1}(\boldsymbol{\beta}_{t-1}) &\propto \int p(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1}, z_t, \boldsymbol{\theta})p(y_{t:T}|\boldsymbol{\beta}_t, z_{t:T}, \boldsymbol{\theta})d\boldsymbol{\beta}_t \\ &\propto \int p(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1}, z_t, \boldsymbol{\theta})p(y_t|\boldsymbol{\beta}_t, z_t, \boldsymbol{\theta})p(y_{t+1:T}|\boldsymbol{\beta}_t, z_{t+1:T}, \boldsymbol{\theta})d\boldsymbol{\beta}_t \\ &\propto \int p(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1}, z_t, \boldsymbol{\theta})p(y_t|\boldsymbol{\beta}_t, z_t, \boldsymbol{\theta})m_{t+1,t}(\boldsymbol{\beta}_t)d\boldsymbol{\beta}_t, \end{aligned}$$

where it is assumed that $m_{t+1,t}(\boldsymbol{\beta}_t) \propto N(f_{t+1,t}, F_{t+1,t})$. Then,

$$\begin{aligned} m_{t,t-1}(\boldsymbol{\beta}_{t-1}) &\propto \int p(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1}, z_t, \boldsymbol{\theta})p(y_t|\boldsymbol{\beta}_t, z_t, \boldsymbol{\theta})m_{t+1,t}(\boldsymbol{\beta}_t)d\boldsymbol{\beta}_t \\ &\propto \int \exp\left\{ -\frac{1}{2} \begin{bmatrix} \boldsymbol{\beta}_{t-1} \\ \boldsymbol{\beta}_t \end{bmatrix}' \begin{bmatrix} A'\Sigma^{-1}A & -A'\Sigma^{-1} \\ -\Sigma^{-1}A & \Sigma^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_{t-1} \\ \boldsymbol{\beta}_t \end{bmatrix} \right\} \\ &\quad \cdot \exp\left\{ -\frac{1}{2} \begin{bmatrix} \boldsymbol{\beta}_{t-1} \\ \boldsymbol{\beta}_t \end{bmatrix}' \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{1}{r}X_t'X_t \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_{t-1} \\ \boldsymbol{\beta}_t \end{bmatrix} + \begin{bmatrix} \boldsymbol{\beta}_{t-1} \\ \boldsymbol{\beta}_t \end{bmatrix}' \begin{bmatrix} \mathbf{0} \\ \frac{1}{r}X_t'y_t \end{bmatrix} \right\} \\ &\quad \cdot \exp\left\{ -\frac{1}{2} \begin{bmatrix} \boldsymbol{\beta}_{t-1} \\ \boldsymbol{\beta}_t \end{bmatrix}' \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & F_{t+1,t}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_{t-1} \\ \boldsymbol{\beta}_t \end{bmatrix} + \begin{bmatrix} \boldsymbol{\beta}_{t-1} \\ \boldsymbol{\beta}_t \end{bmatrix}' \begin{bmatrix} \mathbf{0} \\ F_{t+1,t}^{-1}f_{t+1,t} \end{bmatrix} \right\} d\boldsymbol{\beta}_t \\ &\propto \int \exp\left\{ -\frac{1}{2} \begin{bmatrix} \boldsymbol{\beta}_{t-1} \\ \boldsymbol{\beta}_t \end{bmatrix}' \begin{bmatrix} A'\Sigma^{-1}A & -A'\Sigma^{-1} \\ -\Sigma^{-1}A & \Sigma^{-1} + F_t^b \end{bmatrix} \left(\begin{bmatrix} \boldsymbol{\beta}_{t-1} \\ \boldsymbol{\beta}_t \end{bmatrix} \right.\right. \\ &\qquad \left.\left. - 2 \begin{bmatrix} A'\Sigma^{-1}A & -A'\Sigma^{-1} \\ -\Sigma^{-1}A & \Sigma^{-1} + F_t^b \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{0} \\ f_t^b \end{bmatrix} \right) \right\} d\boldsymbol{\beta}_t, \end{aligned}$$

where,

$$\begin{aligned} F_t^b &= \frac{1}{r}X_t'X_t + F_{t+1,t}^{-1} \\ f_t^b &= \frac{1}{r}X_t'y_t + F_{t+1,t}^{-1}f_{t+1,t}. \end{aligned}$$

It is easy to see that $m_{t,t-1}(\boldsymbol{\beta}_{t-1}) \propto N(f_{t,t-1}, F_{t,t-1})$, where

$$\begin{aligned} f_{t,t-1} &= F_{t,t-1}(\Sigma^{-1} + F_t^b)^{-1} A' \Sigma^{-1} f_t^b \\ F_{t,t-1} &= [A'\Sigma^{-1}A - A'\Sigma^{-1}(\Sigma^{-1} + F_t^b)^{-1}\Sigma^{-1}A]^{-1}. \end{aligned}$$

The conditional distribution of $\boldsymbol{\beta}_t$ is then computed as:

$$\begin{aligned} p(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1}, z_{1:T}, y_{1:T}, \boldsymbol{\theta}) &\propto p(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1}, z_t, \boldsymbol{\theta}) p(y_t|\boldsymbol{\beta}_t, z_t, \boldsymbol{\theta}) m_{t+1,t}(\boldsymbol{\beta}_t) \\ &\propto \exp\Big\{ -\frac{1}{2}(\boldsymbol{\beta}_t - A\boldsymbol{\beta}_{t-1})'\Sigma^{-1}(\boldsymbol{\beta}_t - A\boldsymbol{\beta}_{t-1})\Big\} \\ &\quad \cdot \exp\Big\{ -\frac{1}{2r}(y_t - X_t\boldsymbol{\beta}_t)'(y_t - X_t\boldsymbol{\beta}_t)\Big\} \\ &\quad \cdot \exp\Big\{ -\frac{1}{2}(\boldsymbol{\beta}_t - f_{t+1,t})' F_{t+1,t}^{-1}(\boldsymbol{\beta}_t - f_{t+1,t})\Big\} \\ &\propto \exp\Big\{ -\frac{1}{2}\Big[\boldsymbol{\beta}_t'(F_t^b + \Sigma^{-1})\boldsymbol{\beta}_t - 2\boldsymbol{\beta}_t'\Big(\Sigma^{-1}A\boldsymbol{\beta}_{t-1} + f_t^b\Big)\Big]\Big\} \\ &\propto N(\mu_{\boldsymbol{\beta}_t}, \Sigma_{\boldsymbol{\beta}_t}), \end{aligned}$$

where,

$$\begin{aligned} \Sigma_{\boldsymbol{\beta}_t} &= (F_t^b + \Sigma^{-1})^{-1} \\ \mu_{\boldsymbol{\beta}_t} &= \Sigma_{\boldsymbol{\beta}_t}\Big(\Sigma^{-1}A\boldsymbol{\beta}_{t-1} + f_t^b\Big). \end{aligned}$$

We first compute the messages $m_{t+1,t}(\boldsymbol{\beta}_t)$ backward in time, by initializing it with $F_T^b = \frac{1}{r}X_T'X_T, f_T^b = \frac{1}{r}X_T'y_T$, and we then sample the state sequence $\boldsymbol{\beta}_{1:T}$ forwards in time.