



# COLEGIO DE POSTGRADUADOS

---

---

INSTITUCIÓN DE ENSEÑANZA E INVESTIGACIÓN EN CIENCIAS AGRÍCOLAS

## CAMPUS MONTECILLO

POSTGRADO DE SOCIOECONOMÍA, ESTADÍSTICA E INFORMÁTICA  
ESTADÍSTICA

### MODELACIÓN DE FENÓMENOS DINÁMICOS COMPLEJOS MEDIANTE MODELOS DINÁMICOS LINEALES DE CAMBIO DE RÉGIMEN

DAYNA PRISCILA SALDAÑA ZEPEDA

T E S I S

PRESENTADA COMO REQUISITO PARCIAL PARA  
OBTENER EL GRADO DE:

DOCTORA EN CIENCIAS

MONTECILLO, TEXCOCO, EDO. DE MÉXICO  
2017

---

---

La presente tesis titulada: **Modelación de fenómenos dinámicos complejos mediante modelos dinámicos lineales de cambio de régimen**, realizada por la alumna: **Dayna Priscila Saldaña Zepeda**, bajo la dirección del Consejo Particular indicado, ha sido aprobada por el mismo y aceptada como requisito parcial para obtener el grado de:

**DOCTORA EN CIENCIAS**

**SOCIOECONOMÍA, ESTADÍSTICA E INFORMÁTICA  
ESTADÍSTICA**

**CONSEJO PARTICULAR**

CONSEJERO

  
\_\_\_\_\_  
Dr. Ciro Velasco Cruz

ASESOR

  
\_\_\_\_\_  
Dra. Elizabeth González Estrada

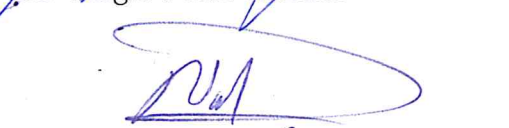
ASESOR

  
\_\_\_\_\_  
Dr. Paulino Pérez Rodríguez

ASESOR

  
\_\_\_\_\_  
Dr. Sergio Pérez Elizalde

ASESOR

  
\_\_\_\_\_  
Dr. Víctor Hugo Torres Preciado

Montecillo, Texcoco, Estado de México, junio de 2017


**CARTA DE CONSENTIMIENTO DE USO DE LOS DERECHOS DE AUTOR Y DE LAS REGALIAS COMERCIALES DE PRODUCTOS DE INVESTIGACION**

En adición al beneficio ético, moral y académico que he obtenido durante mis estudios en el Colegio de Postgraduados, el que suscribe Dayna Priscila Saldaña Zepeda, Alumno (a) de esta Institución, estoy de acuerdo en ser participe de las regalías económicas y/o académicas, de procedencia nacional e internacional, que se deriven del trabajo de investigación que realicé en esta institución, bajo la dirección del Profesor Dr. Ciro Velasco Cruz, por lo que otorgo los derechos de autor de mi tesis

Modelación de fenómenos dinámicos complejos mediante modelos dinámicos lineales de cambio de régimen

y de los producto de dicha investigación al Colegio de Postgraduados. Las patentes y secretos industriales que se puedan derivar serán registrados a nombre el colegio de Postgraduados y las regalías económicas que se deriven serán distribuidas entre la Institución, El Consejero o Director de Tesis y el que suscribe, de acuerdo a las negociaciones entre las tres partes, por ello me comprometo a no realizar ninguna acción que dañe el proceso de explotación comercial de dichos productos a favor de esta Institución.

Montecillo, Mpio. de Texcoco, Edo. de México, a 29 de junio de 2017

  
\_\_\_\_\_  
Firma del  
Alumno (a)

  
\_\_\_\_\_  
Dr. Ciro Velasco Cruz  
Vo. Bo. del Consejero o Director de Tesis

# MODELACIÓN DE FENÓMENOS DINÁMICOS COMPLEJOS MEDIANTE MODELOS DINÁMICOS LINEALES DE CAMBIO DE RÉGIMEN

Dayna Priscila Saldaña Zepeda, Dra.

Colegio de Postgraduados, 2017

## RESUMEN

En la práctica, el supuesto de que las observaciones en una muestra aleatoria son idénticamente distribuidas no es adecuado para muchos tipos de fenómenos. Para relajar este supuesto, se han propuesto mecanismos que permiten la agrupación de observaciones similares. Cuando la agrupación es por medio de la distribución que se genera las observaciones, el problema se enmarca en un modelo de mezclas. Los procesos Dirichlet para modelos de mezclas permiten determinar, simultáneamente, el número de distribuciones en la mezcla y los parámetros que las definen.

En series de tiempo, la violación del supuesto de estacionariedad es frecuente y natural, debido a su inherente dinámica que les permite evolucionar con el tiempo. En algunos casos la evolución que exhiben es simple, y se puede representar satisfactoriamente por un modelo dinámico lineal. Fenómenos más complejos, en los que la dinámica se relaciona con eventos que originan cambios estructurales en el tiempo, se aproximan mediante sistemas dinámicos lineales de cambio de régimen (SLDS, por sus siglas en inglés).

En esta tesis se propone un modelo de regresión dinámica que permite saber cuántas distribuciones diferentes están presentes, dónde se encuentran, y estimar los parámetros que las definen. Adicionalmente, como el modelo de regresión dinámica incluye covariables, es de interés incorporar selección de variables como un elemento para distinguir entre las distribuciones de las que se generan los datos. Las propuestas son extensiones de los SLDS. El desempeño de los modelos se examina mediante simulación, y su utilidad se respalda con problemas prácticos.

**Palabras clave:** procesos Dirichlet jerárquicos, modelos de espacio-estado, selección de variables.

# MODELING COMPLEX DYNAMICAL DATA WITH SWITCHING DYNAMIC LINEAR MODELS (SDLM)

Dayna Priscila Saldaña Zepeda, Dra.

Colegio de Postgraduados, 2017

## ABSTRACT

In practical applications, the assumption that observations in a random sample are identically distributed is not suitable for many phenomena. In order to relax this assumption, mechanisms have been proposed for clustering similar observations, such that observations in the same group are similar but different from those in other groups. When clustering is based on distributions, a mixture of distributions is more appropriate to model the uncertainty of the generating data process. The Dirichlet Process Mixture Models (DPMM) are able to simultaneously infer the number of distributions in the data and learn about the distributions' parameters.

In time series data, the stationarity assumption is violated mostly because their evolution in time. In some cases, the behavior exhibited by the data is simple and can be satisfactorily explained by a linear dynamical model. However, more complex phenomena in which dynamics is related to events that cause structural changes over time, are well described by the switching linear dynamical systems (SLDS).

In this thesis, we propose a flexible dynamic regression model to learn about the number of components in the mixture of distributions, and associate some events that might be responsible for the changes in distributions. Additionally, we are interested in incorporating variable selection as an element to distinguish between the distributions in the mixture. The proposal is an extension of the SLDS, its performance is evaluated by simulation, and its usefulness is illustrated by practical applications.

**Keywords:** Hierarchical Dirichlet processes, state-space models, variable selection.

*A los hombres que me dieron la vida,  
mi padre, Pedro Saldaña†  
y mi hijo, Aarón Julián.*

## AGRADECIMIENTOS

La vida no es la que uno vivió,  
sino la que uno recuerda y  
cómo la recuerda para contarla.  
*Gabriel García Márquez*

Esta tesis es fruto del soporte económico fundamental de dos instituciones: el Consejo Nacional de Ciencia y Tecnología (CONACYT) y la Universidad de Colima (UDC). No debo menos que expresar mi profundo agradecimiento por la invaluable oportunidad de continuar mi formación académica.

La investigación comenzó con el enorme entusiasmo contagiado por mi Consejero, el Dr. Ciro Velasco Cruz, en temas de teoría de la medida y procesos Dirichlet. Su constante ánimo e interés de nuevo aprendizaje, su apertura a nuevas ideas, y su pericia para proveer de una crítica perspicaz, han hecho que trabajar con él sea una gran aventura. Confesar un inmenso agradecimiento por su apoyo, dirección, infinita paciencia, el aliento de elegir libremente mis intereses académicos, y sobretodo por ser mi fuente de estimulación intelectual durante toda mi estancia en el Colpos, en especial para la culminación del presente trabajo, es muy poco. Seguro que no pude tener mejor consejero y amigo.

Tengo también una gran deuda de agradecimiento con los miembros de mi consejo particular, mis asesores: Dr. Paulino Pérez, Dr. Víctor Hugo Torres, Dr. Sergio Pérez y Dra. Elizabeth González, por sus oportunos comentarios, sus observaciones y su tiempo dedicado a la revisión, que han sido fundamentales en la conclusión de este trabajo.

No menos agradecida estoy con todo el personal administrativo del departamento de Estadística por el gran apoyo en todo momento. El ambiente de buenas vibras que se respira, y contar con alguien que está siempre pendiente de mis necesidades escolares, es algo de lo mucho que voy a extrañar. Quiero hacer una mención particular para Isa y Jacque, que siempre fueron tan atentas conmigo.

De manera más personal, necesito manifestar con lo más grande y sincero de mi corazón algunos *gracias*, primero a mi familia, mi mamá, Catalina, y mis hermanos, Ely y Pedro Aarón, por su comprensión y apoyo que desde siempre me han brindado. Y en especial a mi chiquitín, Aarón Julián, que es mi manantial de fuerza e inspiración, y quién más sufrió el proceso de esta osadía que él no eligió. Gracias por ayudarme a entender que mi vida no queda consumada por completo en la investigación. Espero pronto devolver ese favor.

A mis grandes amigos de Colima, Dora y Ricardo, que han sido un gran soporte emocional, y que, sin importar las circunstancias y sus propias preocupaciones, invariablemente tienen un corazón sincero, lleno de optimismo y de sentido del humor conmigo. Los momentos con ustedes siempre son una fiesta. Soy realmente afortunada de tenerlos tan cerca.

Finalmente, quiero agradecer al Universo que ha conspirado tantas experiencias de vida a mi favor. Simple principio de la Teoría del Caos, con un pequeño cambio no estaría donde estoy. Y reza una canción: *es una barca con dos remos en el mar, un remo aprietan mis manos, el otro lo mueve el azar*. No soy sólo yo. Al final, el Universo me ha hecho quien soy.

Gracias,

Dayna Saldaña  
Verano 2017



# CONTENIDO

<b>LISTA DE TABLAS</b>	<b>xiii</b>
<b>LISTA DE FIGURAS</b>	<b>xiv</b>
<b>Capítulo 1: Introducción</b>	<b>1</b>
1.1 Organización y resumen capitular de la tesis . . . . .	4
1.1.1 Resumen: Métodos Bayesianos No-paramétricos . . . . .	5
1.1.2 Resumen: Modelos lineales dinámicos . . . . .	5
1.1.3 Resumen: SLDS para problemas de regresión . . . . .	6
1.1.4 Resumen: Selección de variables en SLDS . . . . .	7
1.1.5 Resumen: Conclusiones y trabajo futuro . . . . .	7
<b>Capítulo 2: Métodos Bayesianos No-paramétricos</b>	<b>8</b>
2.1 Introducción . . . . .	8
2.2 Caracterización de la distribución Dirichlet . . . . .	10
2.2.1 Propiedades . . . . .	13
2.3 El proceso Dirichlet . . . . .	15
2.3.1 Definición teórica . . . . .	15
2.3.2 Propiedades . . . . .	17

2.3.3	Representaciones del DP . . . . .	20
2.4	Proceso Dirichlet para modelos de mezclas . . . . .	23
2.4.1	Ejemplo 1: datos simulados . . . . .	26
2.4.2	Ejemplo 2: geyser Old Faithful . . . . .	36
2.4.3	Ejemplo 3: muestreo Gibbs por bloques . . . . .	38
2.5	Modelo de mezclas proceso Dirichlet jerárquico . . . . .	44
2.5.1	Proceso Dirichlet jerárquico . . . . .	46
2.6	Conclusiones . . . . .	47
<b>Capítulo 3: Modelos lineales dinámicos</b>		<b>48</b>
3.1	Introducción . . . . .	48
3.2	Cadenas de Markov . . . . .	51
3.3	Modelos de espacio-estado . . . . .	52
3.3.1	Modelos de Markov ocultos . . . . .	53
3.3.2	Sistemas dinámicos lineales . . . . .	62
3.3.3	Sistemas dinámicos lineales con matrices de covarianzas desconocidas . . . . .	74
3.4	Conclusiones . . . . .	78
<b>Capítulo 4: SLDS para problemas de regresión</b>		<b>80</b>
4.1	Introducción . . . . .	80
4.2	Modelos de Markov ocultos y sistemas dinámicos lineales de cambio de régimen . . . . .	83
4.3	Modelos de Markov ocultos y procesos Dirichlet jerárquicos . . . . .	85

4.3.1	<i>Sticky</i> HDP-HMM . . . . .	88
4.3.2	HDP-SLDS . . . . .	89
4.4	Regresión con SLDS . . . . .	91
4.4.1	Muestreo Gibbs para el modelo de regresión HDP-SLDS . . . . .	93
4.4.2	Estudio de simulación . . . . .	107
4.4.3	Caso 1: Dinámica del tipo de cambio en México, 1970-2016 . . . . .	116
4.4.4	Caso 2: Niveles de ozono de la ciudad de México . . . . .	126
4.5	Regresión espuria y cointegración . . . . .	129
4.5.1	Modelo de corrección de error . . . . .	135
4.5.2	Caso 3: Crecimiento, crédito bancario e inflación en México: un modelo de regresión lineal dinámica . . . . .	137
4.6	Conclusiones . . . . .	145
<b>Capítulo 5: Selección de variables en SLDS</b>		<b>147</b>
5.1	Introducción . . . . .	147
5.2	El modelo . . . . .	148
5.3	Estudio de simulación . . . . .	151
5.4	Conclusiones . . . . .	164
<b>Capítulo 6: Conclusiones y trabajo futuro</b>		<b>166</b>
6.1	Trabajo futuro . . . . .	168
6.1.1	Sobreajuste . . . . .	168
6.1.2	Convergencia . . . . .	169
6.1.3	Predicción . . . . .	169

<b>Referencias</b>	<b>171</b>
<b>Anexos</b>	<b>182</b>
A    Procesos Dirichlet para modelos de mezclas . . . . .	183
B    Distribuciones a priori y posteriori de los parámetros dinámicos de un SLDS . . . . .	186
<b>Glosario</b>	<b>194</b>

# LISTA DE TABLAS

2.1	Datos simulados. Algoritmo 1: Escobar & West (1995) . . . . .	28
2.2	Valores ajustados. Algoritmo 1: Escobar & West (1995) . . . . .	30
2.3	Otras estimaciones para la media. Algoritmo 1: Escobar & West (1995) . . . . .	31
2.4	Datos simulados. Algoritmo 2: Neal (2000) . . . . .	33
2.5	Valores ajustados. Algoritmo 2: Neal (2000) . . . . .	34
2.6	Otras estimaciones para la media. Algoritmo 2: Neal (2000) . . . . .	35
2.7	Tiempos de espera entre erupciones: geyser Old Faithful . . . . .	37
2.8	Media del tiempo de espera entre erupciones: geyser Old Faithful . . . . .	37
2.9	Parámetros de datos simulados para el Algoritmo BGS . . . . .	41
2.10	Parámetros estimados con el Algoritmo BGS . . . . .	43
3.1	Distribuciones para el proceso de actualización: $\{1, 1, R_t, \Sigma_t\}$ . . . . .	70
4.1	Eventos asociados a los cambios de régimen: México, 1970-2016 . . . . .	120

# LISTA DE FIGURAS

1.1	Serie de tiempo con 3 modos . . . . .	4
2.1	Tiempos de espera entre erupciones sucesivas del géiser Old Faithful	9
2.2	Distribución Dirichlet, $k = 3$ y distintos valores de $\alpha$ . . . . .	12
2.3	Proceso Dirichlet . . . . .	18
2.4	Esquema de urnas . . . . .	19
2.5	Muestras DP generadas por un CRP con distintos valores de $\alpha$ . . . . .	22
2.6	<i>Stick-breaking</i> con distintos valores de $\alpha$ . . . . .	23
2.7	Realizaciones del proceso Dirichlet generadas con Stick Breaking . . . . .	24
2.8	Datos simulados. Algoritmo 1: Escobar & West (1995) . . . . .	29
2.9	Agrupación con el Algoritmo 1: Escobar & West (1995) . . . . .	30
2.10	Densidad estimada. Algoritmo 1: Escobar & West (1995) . . . . .	31
2.11	Datos simulados. Algoritmo 2: Neal (2000) . . . . .	34
2.12	Agrupación con el Algoritmo 2: Neal (2000) . . . . .	35
2.13	Densidad estimada con el Algoritmo 2: Neal (2000) . . . . .	36
2.14	Geyser Old Faithful (a) 272 observaciones; (b) Clusters: Algoritmo 2-Neal (2000) . . . . .	38

2.15	Datos simulados para el Algoritmo BGS . . . . .	42
2.16	Agrupación con el BGS . . . . .	43
2.17	Representación del HDPMM . . . . .	45
3.1	Series macroeconómicas en México. . . . .	49
3.2	HMM de T pasos en el tiempo . . . . .	54
3.3	Valores pronosticados y bandas de confianza de 95 % para el crecimiento del producto interno bruto en México. . . . .	71
3.4	Medias <i>forward</i> y <i>backward</i> para el crecimiento del producto interno bruto en México. . . . .	74
3.5	Valores <i>filtering</i> y bandas de confianza de 90 % para el crecimiento del producto interno bruto en México con distintos valores del <i>factor de descuento</i> . . . . .	78
4.1	Mediciones en el tiempo del órgano de una planta. . . . .	81
4.2	SLDS de T pasos en el tiempo . . . . .	85
4.3	Representación gráfica de un HDP-HMM . . . . .	86
4.4	Representación gráfica de un HDP-SLDS . . . . .	90
4.5	Representación gráfica de un HDP-SLDS con observaciones dependientes de los modos . . . . .	97
4.6	Datos simulados para el modelo de regresión HDP-SLDS. . . . .	110
4.7	Ajuste para el modelo de regresión HDP-SLDS. . . . .	112
4.8	Número de modos en cada iteración para los datos simulados. . . . .	113
4.9	Secuencia de modos en la iteración 7400. . . . .	114
4.10	Probabilidad estimada en cada $t$ de un cambio de modo para los datos simulados. . . . .	115

4.11	Estimación de las pendientes del vector de estados para el modelo de regresión HDP-SLDS. . . . .	116
4.12	Tipo de cambio peso-dólar, 1970-1983. . . . .	117
4.13	Tipo de cambio peso-dólar, 1970-2016. . . . .	118
4.14	Ajuste de la serie: tipo de cambio peso-dólar, 1970-2016 . . . . .	123
4.15	Número de modos en cada iteración para los datos de tipo de cambio	124
4.16	Probabilidad estimada diaria de un cambio de modo . . . . .	125
4.17	Parámetros estimados y ajuste para la estación Plateros de monitoreo atmosférico. . . . .	128
4.18	Parámetros estimados y ajuste para la estación Tlalnepantla de monitoreo atmosférico. . . . .	129
4.19	Curvas senoidales con distinta <i>fase</i> . . . . .	130
4.20	Variación de la correlación entre dos elementos infinitesimales de curvas armónicas . . . . .	131
4.21	Datos simulados de dos procesos $I(1)$ independientes . . . . .	134
4.22	$\log(PIB, GOB, INV, XP), INF, CP$ : México, 1961-2011. . . . .	140
4.23	Estimaciones dinámicas de los coeficientes de pendiente $\beta_{jt}$ . . . . .	141
4.24	Ajuste para los datos del logaritmo del producto interno bruto (PIB), 1961-2011. . . . .	145
5.1	Escenario 1. Observaciones, secuencia de modos y covariables simuladas. . . . .	154
5.2	Escenario 1. Secuencia simulada y estimada del vector de estados. . . . .	155
5.3	Escenario 1. Número de modos detectados. . . . .	156
5.4	Escenario 1. Ajuste de observaciones, secuencia de modos y probabilidad de punto de cambio. . . . .	156



5.5	Escenario 2. Observaciones, secuencia de modos y covariables simuladas. . . . .	157
5.6	Escenario 2. Secuencia simulada y estimada del vector de estados. . . . .	158
5.7	Escenario 2. Número de modos detectados. . . . .	159
5.8	Escenario 2. Ajuste de observaciones, secuencia de modos y probabilidad de punto de cambio. . . . .	159
5.9	Escenario 3. Observaciones, secuencia de modos y covariables simuladas. . . . .	161
5.10	Escenario 3. Secuencia simulada y estimada del vector de estados. . . . .	162
5.11	Escenario 3. Número de modos detectados. . . . .	164
5.12	Escenario 3. Ajuste de observaciones, secuencia de modos y probabilidad de punto de cambio. . . . .	164

# Capítulo 1

## Introducción

El carácter dinámico de muchos fenómenos en diversos campos de la ciencia, como en economía y finanzas (Kim, 1994; Carvalho & Lopes, 2007; West, 2013; Zeng & Wu, 2013; McAlinn & West, 2016), en estudios del movimiento del cuerpo humano (Bregler, 1997; Pavlović et al., 2001), y en estudios ambientales (Conrad Lamon III et al., 1998; Huerta et al., 2004; Velasco Cruz et al., 2012), ha motivado el desarrollo de métodos flexibles para representarlos. Estos métodos se adaptan a las condiciones del sistema que los origina, y permiten modelar fenómenos en los que el supuesto de estacionariedad no se sostiene, ya que son capaces de capturar cambios repentinos experimentados por el sistema como respuesta a modificaciones de algunas características que directa o indirectamente pudieran afectarlo. Por ejemplo, la proliferación de una enfermedad transmitida por un virus puede tener cambios debido a las condiciones climáticas, al desarrollo de una nueva vacuna, o a la implementación de otros mecanismos para su control; un país puede sufrir cambios en su economía debido a una recesión, a un evento de impacto nacional (como para Cuba, la muerte de su presidente Fidel Castro en 2016), o a un evento global externo (como el que implicó la llegada del presidente Trump en 2017 a Estados Unidos). Los eventos asociados a estos ejemplos representan los *estados* (o *modos*) del sistema que origina el fenómeno, y la serie de tiempo retrata la evolución del fenómeno en cada *estado* del sistema, que se modelan adecuadamente mediante un tipo de modelos dinámicos llamado *sistemas dinámicos lineales de cambio de régimen* (SLDS) (Ghahramani & Hinton, 1998; Fox et al., 2011a; Barber, 2012).

Los SLDS son una extensión de los modelos de Markov ocultos (HMM) (E. & Petrie, 1966; Rabiner, 1989; Bishop, 2006), en los que en cada *modo* se asocia un sistema dinámico lineal (LDS). Aunque la dinámica de muchas series de tiempo encontradas en la práctica es compleja y generalmente no lineal, se puede aproximar

## 1. Introducción

---

eficientemente mediante cambios entre un conjunto de *modos* condicionalmente lineales. La idea básica de los **SLDS** es dividir la serie en segmentos (modos), y modelar cada uno con un modelo dinámico lineal. Explícitamente, un **SLDS** se puede representar, siguiendo la definición de [Fox et al. \(2011a\)](#), por el siguiente modelo jerárquico

$$z_t | z_{t-1} \sim \boldsymbol{\pi}_{z_{t-1}} \quad (1.1)$$

$$\boldsymbol{\beta}_t = A^{(z_t)} \boldsymbol{\beta}_{t-1} + \mathbf{e}_t^{(z_t)} \quad (1.2)$$

$$\mathbf{y}_t = \mathbf{C} \boldsymbol{\beta}_t + \mathbf{w}_t, \quad (1.3)$$

donde  $z_t$  es la variable oculta del **HMM** al tiempo  $t^1$  que representa los modos. Las ecuaciones (1.2)-(1.3) definen el **LDS**;  $\boldsymbol{\beta}_t$  es una variable latente (oculta) continua, comúnmente Gaussiana,  $\mathbf{y}_t^2$  es el vector de respuestas,  $\mathbf{C}$  es una matriz constante de interceptos,  $\mathbf{e}_t^{(z_t)} \sim N(\mathbf{0}, \Sigma^{(z_t)})$ , y  $\mathbf{w}_t \sim N(\mathbf{0}, \mathbf{R})$ ; los términos de error se asumen interna y mutuamente independientes. El **LDS** y el **HMM** se relacionan mediante la dependencia de  $\{A^{(z_t)}, \Sigma_t^{(z_t)}\}$  en  $z_t$ .

El **SLDS** definido por el sistema (1.1)-(1.3) permite describir la complejidad de la serie de tiempo mediante un proceso latente **HMM** que induce agrupamiento de los parámetros que definen la distribución de  $\boldsymbol{\beta}_t$ ; las observaciones se modelan entonces por medio de la dependencia en  $\boldsymbol{\beta}_t$ , y no dependen directamente de  $z_t$ . Para hacer inferencia con este modelo, [Fox et al. \(2011a\)](#) proponen un algoritmo *forward-backward* que itera entre el muestreo por bloques ([Ishwaran & James, 2001](#)) de la secuencia de modos (variables ocultas del **HMM**) y la secuencia de estados (variables ocultas del **LDS**). El muestreo de la secuencia de modos usa una aproximación truncada del proceso Dirichlet ([Ishwaran & James, 2002](#); [Ishwaran & Zarepour, 2002](#)), condicionando sobre la secuencia de estados. El muestreo por bloques de la secuencia de estados se realiza condicionado a la secuencia de modos, tal como fue propuesto originalmente por [Carter & Kohn \(1994, 1996\)](#). El algoritmo de [Fox et al. \(2011a\)](#) es un procedimiento Bayesiano no-paramétrico que permite modelar series de tiempo complejas mediante un conjunto de posibles modelos dinámicos más simples. El número de modelos en el conjunto está determinado por el truncamiento del **DP**; al mismo tiempo, el **DP** permite relajar el supuesto de otras metodologías que asumen fijo y conocido el número de modelos presentes en la serie (ver por ejemplo [Ghahramani & Hinton, 2000](#); [Kotsalis et al., 2006](#); [Chen et al., 2011](#)).

En esta tesis se presenta una extensión del **SLDS** de [Fox et al. \(2011a\)](#) en el que la variable de interés se relaciona directamente con los *modos* que originan los

---

<sup>1</sup>En esta tesis se estudian sistemas de tiempo discreto.

<sup>2</sup>En general, un sistema dinámico lineal se define para un vector de observaciones; sin embargo, en esta tesis se trabaja con una observación para cada  $t$ .

## 1. Introducción

---

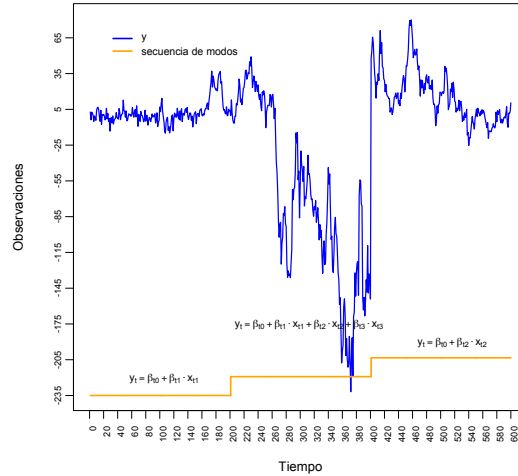
cambios dinámicos a través del término de error  $w_t$ , similar a la dependencia de los modos sobre los estados. Debido a que se tiene mejor comprensión del sistema en estudio cuando se consideran variables relevantes con influencia en la variable de interés, se establece un modelo de regresión dinámico que incluye variables explicativas, o covariables, y no únicamente intercepto, como en el SLDS de Fox et al. (2011a). La interpretación de los coeficientes asociados a este modelo es similar a la de los modelos de regresión clásicos, es decir, cuantifican el efecto que tienen sobre la variable de interés; pero en un modelo de regresión dinámico el efecto de una variable sobre otra evoluciona con el tiempo. Esto agrega flexibilidad al modelo; permite describir relaciones cambiantes entre variables conforme el tiempo transcurre. El SLDS extendido se presenta entonces por el siguiente modelo jerárquico:

$$z_t | z_{t-1} \sim \pi_{z_{t-1}} \quad (1.4)$$

$$\beta_t = A^{(z_t)} \beta_{t-1} + \mathbf{e}_t^{(z_t)} \quad (1.5)$$

$$y_t = X_t' \beta_t + w_t^{(z_t)}, \quad (1.6)$$

donde se asume que  $w_t^{(z_t)} \sim N(0, R^{(z_t)})$ ,  $y_t$  es escalar, y  $X_t$  es un vector que incluye intercepto y una o más covariables al tiempo  $t$ . Una segunda extensión del modelo SLDS propuesta en esta tesis consiste en permitir que la variable de interés dependa de los modos a través del vector de covariables  $X_t$ . Esta dependencia se expresa mediante una variable indicadora que señala qué covariables ejercen un efecto significativo en cada modo. La propuesta se basa en la hipótesis de que una variable explicativa puede ser relevante para la variable respuesta en un segmento del tiempo, pero puede no ser significativa en otros segmentos del tiempo debido, por ejemplo, a eventos no anticipados o no presentes en toda la serie. Para ilustrar esto, considere la Figura 1.1. La gráfica corresponde a una serie de tiempo de 600 observaciones (línea azul),  $t = 1, \dots, 600$ , en la que la dinámica está caracterizada por tres modos presentes en intervalos de tiempo de igual longitud (línea amarilla). En cada modo, las observaciones están en función de un subconjunto de un conjunto de tres covariables más el intercepto. En el modo 1, la variable observada depende de la covariable 1; en el modo 2 las tres covariables ejercen un efecto significativo sobre las observaciones, y la varianza en este modo es más grande que en los otros dos; y en el modo 3, las observaciones están definidas por el efecto de la segunda covariable. En los tres modos el intercepto está presente. La implicación de las dos extensiones al modelo SLDS es que representan elementos adicionales para distinguir entre modos; es decir, un modo  $k$  está definido por el conjunto  $\{A^{(k)}, \Sigma_t^{(k)}, R^{(k)}\}$  y un subconjunto de covariables significativas.



**Figura 1.1:** Serie de tiempo con 3 modos

Para hacer inferencia sobre las extensiones propuestas, se muestrean por bloques las secuencias de modos y estados usando una variante del algoritmo *forward-backward* como en [Fox et al. \(2011a\)](#), pero trabajando sobre un modelo marginal de los estados para mostrar la secuencia de modos.

## 1.1. Organización y resumen capitular de la tesis

La tesis está organizada por capítulos; los capítulos 2 y 3 describen los fundamentos teóricos sobre los que se basan las contribuciones, desarrolladas en los capítulos 4 y 5. En cada uno se presentan ejemplos con datos simulados y datos reales. El objetivo es clarificar las metodologías expuestas y emplearlas en aplicaciones prácticas. Los ejemplos con datos reales tienen justificación teórica o empírica, y las contribuciones son resultado del interés por hacer aportaciones a la estadística aplicada. Los lectores con bases suficientemente sólidas en procesos Dirichlet y modelos dinámicos pueden obviar la lectura de los capítulos 2 y 3, y concentrarse directamente en las aportaciones presentadas a partir del capítulo 4. Las conclusiones generales se exponen en el último capítulo. Debido a la limitante de tiempo para cubrir un objetivo más ambicioso en la investigación, y a que todo trabajo puede ser mejorable o extendible, se incluye también en ese capítulo una discusión sobre el trabajo futuro derivado de los resultados de la investigación.

En esta sección se presenta una breve descripción del contenido de cada capítulo,

## 1.1. Organización y resumen capitular de la tesis

---

con el fin de dar una apreciación preliminar de la aportación de cada uno hacia los objetivos y principales contribuciones de la tesis.

### 1.1.1. Resumen: Métodos Bayesianos No-paramétricos

En la práctica es común encontrar datos para los que no es adecuado suponer que una única distribución o modelo sea suficiente para describirlos o caracterizarlos. Para este problema, las mezclas de distribuciones paramétricas son de más utilidad. Típicamente, la estimación de los parámetros de las distribuciones que definen los componentes de la mezcla y las probabilidades de pertenencia a cada componente, se realiza siempre y cuando el número de componentes de mezcla sea conocido. En un contexto de modelos Bayesianos no paramétricos, los modelos de mezclas basados en el proceso Dirichlet como distribución a priori sobre los parámetros, se conocen como proceso Dirichlet para modelos de mezclas (DPMM). El DPMM permite encontrar, simultáneamente, el número de los componentes en la mezcla y los parámetros que los definen.

El capítulo comienza motivando un problema de mezclas de distribuciones. La sección 2.2 describe a la distribución Dirichlet como introducción al proceso Dirichlet, que se presenta en la sección 2.3. Las propiedades de agrupación inducidas por los procesos Dirichlet se han explotado en muchas aplicaciones de modelos de mezclas. Los DPMM se ilustran en la sección 2.4 mediante tres algoritmos conocidos para estimar estos modelos. Una extensión del proceso Dirichlet, conocido como proceso Dirichlet jerárquico (HDP) se trata en la sección 2.5. El HDP involucra datos agrupados, en donde cada observación dentro de un grupo se toma de un modelo de mezclas, y se permite que los grupos compartan componentes de mezclas. En el Capítulo 4, el HDP se usa como a priori para un modelo de Markov oculto de manera similar a la construcción del DPMM.

### 1.1.2. Resumen: Modelos lineales dinámicos

Los modelos dinámicos describen el comportamiento de un sistema cambiante en el tiempo. La representación de estos modelos en la forma *espacio-estado* es de gran utilidad para modelar las observaciones. De manera particular, los sistemas dinámicos lineales (LDS) en la forma *espacio-estado* son consistentes con los modelos de regresión dinámica, es decir, con los modelos de regresión en los que los coeficientes asociados a las variables independientes varían con el tiempo. Los HMM y los LDS son los dos tipos más importantes de los modelos de

## 1.1. Organización y resumen capitular de la tesis

---

*espacio-estado*; en el primero, la dinámica se describe en términos de transiciones de una variable aleatoria discreta, mientras que en el último la variable aleatoria es continua tipo Gaussiana.

La sección 3.2 presenta la definición de cadenas de Markov, que son los modelos de Markov más simples. En la sección 3.3 se discuten los modelos de *espacio-estado HMM* y los **LDS**, incluyendo ilustraciones del algoritmo iterativo conocido como *forward-backward*, utilizada para inferencia acerca de los *estados* del sistema; además, se presenta una discusión sobre la habilidad de los **LDS** para incorporar parámetros desconocidos, y la implicación que tienen en la estimación de los *estados*.

### 1.1.3. Resumen: **SLDS** para problemas de regresión

Muchos fenómenos complejos no se pueden describir adecuadamente por un solo **LDS**; sin embargo, se pueden aproximar mediante *cambios* de modelos que pueden ocurrir como respuesta a modificaciones de las condiciones del sistema. Cuando esos *cambios* se modelan con base en una variable latente que sigue un **HMM** de tiempo discreto, el modelo que resulta es un **SLDS**. La sección 4.2 introduce los **SLDS**; la sección 4.3 define al **HDP** como a priori para el **HMM**, y una extensión del modelo, conocida como *sticky HDP-HMM*, que introduce un parámetro para incrementar la probabilidad de permanecer en un mismo *estado*. La sección 4.4 generaliza el **SLDS** para problemas de regresión; el objetivo es modelar series de tiempo con dinámicas complejas, relacionándolas a variables explicativas. Para evaluar el desempeño en ajuste del modelo, se realiza un estudio de simulación. Adicionalmente, se examinan dos conjuntos de datos reales: (1) el modelo detecta cambios de *modo* en la serie diaria del tipo de cambio en México, en el periodo 01/01/1970-11/05/2016, coincidentes con una serie de eventos asociados a periodos de crisis; (2) se ajusta un modelo de regresión de los niveles de ozono en la Cd. de México con la temperatura como covariable.

Cuando se hace análisis de regresión, se parte del supuesto de que las covariables están causalmente relacionadas con la variable respuesta; de otro modo, si la dependencia entre variables no tiene sentido o interpretación, pero se obtienen relaciones significativas, entonces se dice que se ha producido una *relación espuria*. La sección 4.5 está dedicada al tema de regresión espuria y cointegración. La sección concluye con un estudio de caso que explora los efectos de la inflación y el crédito privado en el crecimiento económico de México. Los resultados se comparan con los obtenidos en un estudio basado en un modelo de corrección de error.

### 1.1.4. Resumen: Selección de variables en **SLDS**

Cuando se especifica un modelo de regresión, es común incorporar variables explicativas que se presupone contribuyen en la descripción de la variable de interés o respuesta. Cuando el conjunto de posibles variables a incluir como covariables es muy grande, es deseable que el modelo recurra al principio de parsimonia, que sugiere que la explicación completa más simple es preferible. Los métodos de selección de variables ofrecen alternativas cuando no se tiene una justificación contundente a priori sobre qué variables deben ser incluidas.

En los **SLDS**, la selección de variables es incorporada a los diferentes *modos* dinámicos, de manera que representen más adecuadamente los cambios que el fenómeno exhibe. La idea central es que una variable es relevante en el modelo dada su interacción con algún evento presente en un periodo, pero no necesariamente es significativa cuando tal evento no está presente.

En la sección 5.2 se adapta el método de selección de variables de [Kuo & Mallick \(1998\)](#) en cada *modo* para el modelo de regresión **SLDS**. El método de selección de variables en forma general hace cero los componentes del vector de estados asociados a las covariables que no sean significativas por modo. En la sección 5.3 se evalúa el desempeño en ajuste de la propuesta mediante un estudio de simulación.

### 1.1.5. Resumen: Conclusiones y trabajo futuro

El capítulo resume los contenidos y principales resultados de la tesis. Cabe señalar que hasta ahora, la investigación está enfocada al ajuste del modelo, por lo que en este capítulo se incluye una breve dirección sobre dos posibles tópicos pendientes para futura investigación: predicción y un estudio de cómo controlar sobreajuste si estuviera presente.



# Capítulo 2

## Métodos Bayesianos No-paramétricos

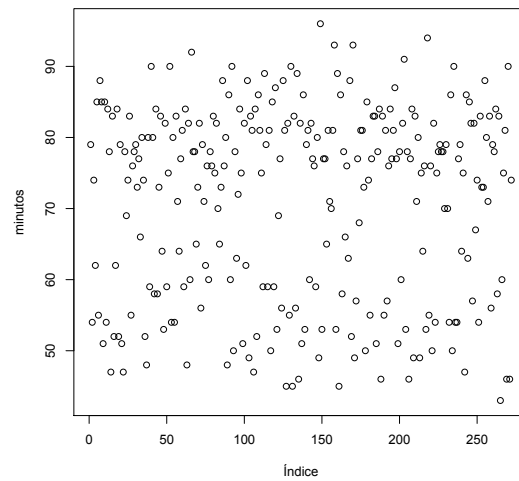
### 2.1. Introducción

Considere los datos presentados en la Figura 2.1. Las observaciones en (a) corresponden a los tiempos de espera (en minutos) entre erupciones sucesivas del géiser *Old Faithful* en el Parque Nacional de Yellowstone en Wyoming, USA. La densidad en (b) se estimó usando estimación no paramétrica de densidades (Silverman, 1986) implementada en la función *density* del paquete estadístico R (R Core-Team, 2016). Como se puede observar, las gráficas dan fuerte evidencia de que la distribución de estos datos es bimodal, y sugieren que las observaciones se pueden dividir en grupos (clusters), según la distribución de la cual fueron generadas.

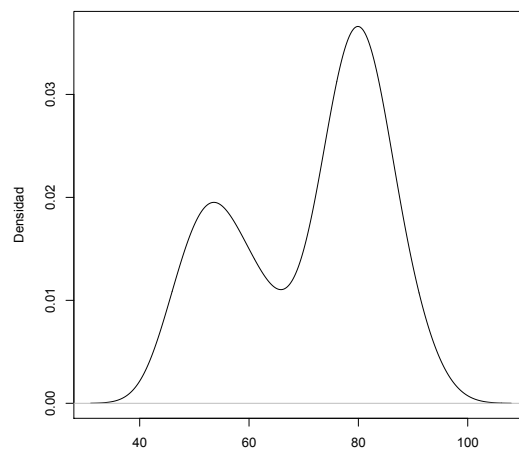
En la práctica, es común encontrar estructuras de datos similares a las de la Figura 2.1, en las que no es adecuado suponer que todas las observaciones provienen de la misma distribución de probabilidad, es decir, que son idénticamente distribuidas. Por el contrario, se identifican grupos de observaciones homogéneas, tal que pertenecen a un mismo grupo si son tomadas de la misma distribución. En algunos casos, el conjunto de datos podría reflejar una agrupación fácil de encontrar mediante examen visual. En otros, la estructura de los datos es más compleja, y la agrupación no es tan evidente o fácil, o se requiere discriminar entre todas las características presentes en los datos, aquellas que son importantes para definir los clusters. Resulta entonces de gran interés contar con procedimientos que permitan encontrar de manera eficiente la agrupación a la que obedecen.

## 2.1. Introducción

---



(a)



(b)

**Figura 2.1:** Tiempos de espera entre erupciones sucesivas del géiser Old Faithful

Motivados por el ejemplo anterior, se parte de un conjunto de observaciones que no están totalmente caracterizadas por una única distribución o modelo, si no que varias distribuciones explican mejor el mecanismo aleatorio que las genera. Sin embargo, no se conoce de antemano cuántas de tales distribuciones, ni los parámetros que las definen, están presentes en los datos. La pregunta es entonces: Qué hacer?

## 2.2. Caracterización de la distribución Dirichlet

---

Modelos Bayesianos no paramétricos, tales como procesos Dirichlet para modelos de mezclas (DPMM), permiten encontrar, simultáneamente, el número de *componentes en la mezcla*, es decir, el número de distribuciones presentes en los datos, y la estimación de sus parámetros. Esta característica es la que los distingue de otros métodos de clasificación en los que se conoce a priori el número de componentes o se usan etiquetas de clase a priori para identificar a las observaciones.

El objetivo de este capítulo es hacer una breve revisión del proceso Dirichlet como un método Bayesiano no paramétrico para modelos de mezclas. En la sección 2.2 se define y caracteriza a la distribución Dirichlet; la sección 2.3 presenta una descripción del proceso Dirichlet, que puede ser entendido como una generalización de la distribución Dirichlet, y se incluyen además dos métodos de representación: proceso de restaurant Chino y *stick-breaking*; la sección 2.4 usa al proceso Dirichlet como a priori para los parámetros de los componentes de un modelo de mezclas, el resultado se denomina modelo de mezclas proceso Dirichlet. Tres algoritmos para estimar estos modelos se describen también en esta sección; una extensión del proceso Dirichlet a dos niveles, llamado proceso Dirichlet jerárquico, se describe brevemente en la sección 2.5. La sección 2.6 presenta conclusiones del capítulo.

## 2.2. Caracterización de la distribución Dirichlet

La distribución Dirichlet es una familia de distribuciones de probabilidad continuas multivariada, parametrizada en Ferguson (1973) por  $\alpha$ , un vector de valores de reales no negativos. Esta distribución es la generalización multivariada de la distribución beta y, por sus características, es frecuentemente utilizada en el contexto Bayesiano como a priori conjugada para los parámetros de la distribución multinomial.

Sea  $(Y_1, \dots, Y_k)$  un vector aleatorio tal que

$$Y_j = Z_j / \sum_{i=1}^k Z_i, \quad j = 1, 2, \dots, k$$

donde  $Z_1, Z_2, \dots, Z_k$  son v.a. independientes con distribución Gamma,  $Z_j \sim \text{Ga}(\alpha_j, 1)$ , con parámetro de forma  $\alpha_j \geq 0$  para todo  $j$ , y  $\alpha_j > 0$  para algún  $j$ ,  $j = 1, 2, \dots, k$ . Entonces, el vector  $(Y_1, \dots, Y_k)$  tiene distribución Dirichlet (Ferguson, 1973). Esta definición permite a algunas de las variables ser degeneradas en cero, es decir, si algún  $\alpha_j = 0$ , entonces  $Z_j$  y  $Y_j$  correspondientes son degeneradas en cero.

## 2.2. Caracterización de la distribución Dirichlet

De la definición es fácil ver que  $Y_1 + \dots + Y_k = 1$ , y que cada  $Y_j$  es Beta:  $Y_j \sim \text{Be}(\alpha_j, (\sum_1^k \alpha_i) - \alpha_j)$ . Por ejemplo, para  $k = 2$ :  $Y_1 \sim \text{Be}(\alpha_1, \alpha_2)$  y  $Y_2 \sim \text{Be}(\alpha_2, \alpha_1)$ , o bien,  $Y_2 = 1 - Y_1 \sim \text{Be}(\alpha_2, \alpha_1)$  por la propiedad de imagen espejo de la distribución Beta. Se sigue que  $(Y_2 = 1 - Y_1) + Y_1 = 1$ .

La densidad Dirichlet está dada por:

$$f(y_1, \dots, y_{k-1} | \alpha_1, \dots, \alpha_k) = \frac{\Gamma(\alpha_1 + \dots + \alpha_k)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_k)} \left( \prod_{j=1}^{k-1} y_j^{\alpha_j - 1} \right) \left( 1 - \sum_{j=1}^{k-1} y_j \right)^{\alpha_k - 1} I_S(y_1, \dots, y_{k-1}) \quad (2.1)$$

donde:

$$S = \{(y_1, \dots, y_{k-1}) : y_j \geq 0, \sum_{j=1}^{k-1} y_j \leq 1\}$$

El soporte de la distribución es el conjunto  $k$ -dimensional, cuyos valores para cada  $y_i$  son números reales en el intervalo  $(0, 1)$ ; estos pueden ser vistos como las probabilidades de un evento categórico (con  $k$  categorías). Dicho de otro modo, el dominio de la distribución Dirichlet es por sí mismo un conjunto de distribuciones de probabilidad, por lo que comúnmente es llamada una *distribución de distribuciones*.

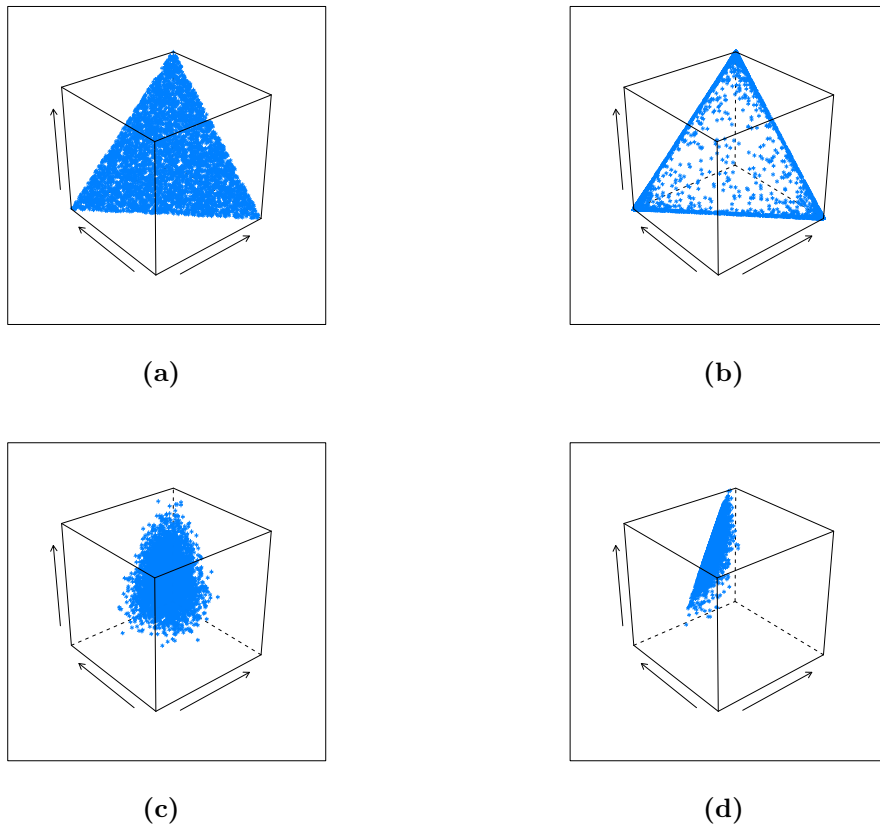
Ejemplo. Considere un conjunto de  $N$  dados comunes. Cada dado puede ser visto como una pmf; para muestrear la pmf se lanza el dado y se obtiene un número del uno al seis. Suponga ahora que los  $N$  dados se meten en una bolsa. La bolsa con los dados es un ejemplo de una pmf con distribución Dirichlet, esto es, un conjunto de pmfs. Para muestrear de esta pmf, se saca de la bolsa un dado, es decir, se toma una pmf. El dado extraído es una realización de la distribución Dirichlet.

En adelante, se denotará a la distribución Dirichlet de parámetro  $\alpha$  como  $\text{Dir}(\alpha)$ .

La Figura 2.2 ilustra a la distribución Dirichlet, para  $k = 3$ , con varios valores del parámetro de concentración  $\alpha$ . De las gráficas se puede notar que cuando  $\alpha = (c, c, c)$  para algún  $c > 0$ , la densidad es simétrica sobre la pmf uniforme. El caso especial  $\alpha = (1, 1, 1)$  en (a), es la distribución uniforme sobre el simplex. Cuando  $0 < c < 1$  como en (b), la densidad se concentra casi en los vertices del simplex, es decir, pequeños valores del parámetro de concentración favorece valores extremos de la distribución. Si  $c > 1$  como en (c), la densidad se concentra en el centro del simplex. Finalmente, si  $\alpha$  no es un vector constante como ocurre en (d), la densidad no es simétrica.

## 2.2. Caracterización de la distribución Dirichlet

---



**Figura 2.2:** Distribución Dirichlet,  $k = 3$  y distintos valores de  $\alpha$ , (a):  $\text{Dir}(1, 1, 1)$ , (b):  $\text{Dir}(0.1, 0.1, 0.1)$ , (c):  $\text{Dir}(10, 10, 10)$ , (d):  $\text{Dir}(1, 10, 30)$

## 2.2. Caracterización de la distribución Dirichlet

---

### 2.2.1. Propiedades

#### Propiedad de agregación

La distribución Dirichlet tiene una propiedad útil del tipo fractal, tal que si se particiona el espacio muestral, la distribución conjunta de sumas de los elementos de la partición es también una distribución Dirichlet sobre el nuevo conjunto de eventos (Frigyik et al., 2010).

Ejemplo. Considere una distribución Dirichlet sobre dados de seis caras con  $\alpha \in \mathbb{R}_+^6$ . Suponga que se desea conocer la probabilidad de obtener un número impar versus la probabilidad de obtener un número par. Por la propiedad de agregación, la distribución Dirichlet sobre las seis caras de dados produce una distribución Dirichlet sobre el espacio muestral formado por dos eventos, pares e impares, con parámetro  $(\alpha_1 + \alpha_3 + \alpha_5, \alpha_2 + \alpha_4 + \alpha_6)$ .

De manera formal, la propiedad se enuncia de la siguiente manera:

Si  $(Y_1, \dots, Y_k) \sim \text{Dir}(\alpha_1, \dots, \alpha_k)$  y  $r_1, \dots, r_l$  son enteros tal que  $0 < r_1 < \dots < r_l = k$ , entonces

$$\left( \sum_1^{r_1} Y_i, \sum_{r_1+1}^{r_2} Y_i, \dots, \sum_{r_{l-1}+1}^{r_l} Y_i \right) \sim \text{Dir} \left( \sum_1^{r_1} \alpha_i, \sum_{r_1+1}^{r_2} \alpha_i, \dots, \sum_{r_{l-1}+1}^{r_l} \alpha_i \right).$$

Es fácil notar que la propiedad se sigue directamente de la definición de la distribución Dirichlet y de la propiedad aditiva de la distribución gamma, por ejemplo: si  $Z_1 \sim Ga(\alpha_1, 1)$ ,  $Z_2 \sim G(\alpha_2, 1)$ , y si  $Z_1, Z_2$  son independientes, entonces  $Z_1 + Z_2 \sim Ga(\alpha_1 + \alpha_2, 1)$ .

#### A priori conjugada

Otra característica de la distribución Dirichlet es que es una a priori conjugada para los parámetros de la distribución multinomial. Esto es, si la distribución a priori de  $(Y_1, \dots, Y_k) \sim \text{Dir}(\alpha_1, \dots, \alpha_k)$  y si

$$P\{X = j | Y_1, \dots, Y_k\} = Y_j \quad \text{para } j = 1, \dots, k$$

## 2.2. Caracterización de la distribución Dirichlet

entonces la distribución a posteriori  $[Y_1, \dots, Y_k | X = j]$  es  $\text{Dir}(\alpha_1^{(j)}, \dots, \alpha_k^{(j)})$ , donde

$$\begin{aligned} \alpha_i^{(j)} &= \alpha_i && \text{si } i \neq j \\ &= \alpha_j + 1 && \text{si } i = j \end{aligned}$$

Un caso particular se tiene cuando se asume una distribución beta como a priori para la probabilidad de éxito en la distribución binomial. Una propiedad similar caracteriza al proceso Dirichlet descrito en la siguiente sección.

Ejemplo. Considere  $k = 2$ :

$$\begin{aligned} P\{X = j | Y_1, Y_2\} &= Y_j \text{ para } j = 1, 2 \\ P\{Y_1, Y_2 | X = j\} &\propto P\{X = j | Y_1, Y_2\} \cdot P\{Y_1, Y_2\} \\ &\propto y_j y_j^{\alpha_1 - 1} (1 - y_j)^{\alpha_2 - 1} \end{aligned}$$

Si  $j = 1$ , la última expresión es  $y_1^{(\alpha_1 + 1) - 1} (1 - y_1)^{\alpha_2 - 1}$  que es el Kernel de  $\text{Dir}(\alpha_1 + 1, \alpha_2)$ , y si  $j = 2$ , entonces  $y_1^{(\alpha_1 - 1)} (1 - y_1)^{(\alpha_2 + 1) - 1}$ , esto es, el Kernel de  $\text{Dir}(\alpha_1, \alpha_2 + 1)$ .

Intuitivamente,  $\alpha$  representa el número de observaciones que ya han sido obtenidas en cada categoría.

### Momentos

La distribución se caracteriza por las siguientes expresiones que definen los momentos:

$$\begin{aligned} E[Y_i] &= \alpha_i / \alpha \\ E[Y_i^2] &= \alpha_i(\alpha_i + 1) / \alpha(\alpha + 1) \\ E[Y_i Y_j] &= \alpha_i \alpha_j / \alpha(\alpha + 1), \end{aligned}$$

donde  $\alpha = \sum \alpha_i$ . Estas expresiones son fácilmente verificables cuando  $k = 2$ , en donde la distribución Dirichlet se reduce a una distribución Beta,  $\text{Be}(\alpha_1, \alpha_2)$ .

Una revisión más extensa de la distribución Dirichlet se presenta en [Frigyik et al. \(2010\)](#), e incluye además algoritmos para muestrear de la distribución basados en urna de Pólya ([Blackwell & MacQueen, 1973](#)), *stick-breaking*, y en transformación de variables aleatorias Gamma. El paquete estadístico R ([R Core-Team, 2016](#)) incluye varias librerías de funciones para muestrear de la distribución, por ejemplo MCMCpack ([Martin et al., 2017](#)) y gtools ([Warnes et al., 2015](#)).

## 2.3. El proceso Dirichlet

La distribución Dirichlet es limitada en el sentido de que asume un conjunto finito de eventos. En el contexto del ejemplo de la sección anterior, esto significa que cada dado tiene un número finito de caras. En cambio, el proceso Dirichlet (DP, por sus siglas en inglés) permite trabajar con un conjunto infinito de eventos, y por tanto, modelar la distribución de probabilidad sobre un espacio muestral más amplio (infinito). El DP puede ser entendido como una generalización de la distribución Dirichlet. La distribución Dirichlet es una distribución de distribuciones, mientras que el DP es una distribución sobre medidas de probabilidad (Antoniak, 1974).

El DP fue formalmente descrito por Ferguson (1973) como una conveniente distribución a priori para problemas noparamétricos. La utilidad radica en que posee dos propiedades deseables:

- I El soporte de la distribución a priori es *grande*.
- II La distribución a posteriori, dada una muestra de observaciones de la verdadera distribución de probabilidad, es analíticamente tratable.

Sin embargo, el conjunto de todas las distribuciones de probabilidad sobre un espacio muestral infinito no es fácilmente manejable. Para tratar con esto, el proceso Dirichlet restringe la clase de distribuciones bajo consideración al conjunto de distribuciones de probabilidad discretas sobre el espacio muestral infinito, escritas como una suma infinita de funciones indicadoras ponderadas.

### 2.3.1. Definición teórica

Considere a  $(\mathcal{X}, \mathcal{A})$  un espacio medible, donde  $\mathcal{X}$  es un espacio y  $\mathcal{A}$  un  $\sigma$ -álgebra de conjuntos en  $\mathcal{X}$ . Sea  $\tilde{\alpha} = \alpha G_0$  una medida finita no nula en  $(\mathcal{X}, \mathcal{A})$ . Entonces, un proceso estocástico  $G$ , indexado por elementos  $A$  de  $\mathcal{A}$ , es un proceso Dirichlet en  $(\mathcal{X}, \mathcal{A})$  con parámetro  $\tilde{\alpha}$  si para cualquier partición medible  $(A_1, \dots, A_k)$  de  $\mathcal{X}$ , el vector aleatorio  $(G(A_1), \dots, G(A_k))$  tiene distribución Dirichlet con parámetro  $(\tilde{\alpha}(A_1), \dots, \tilde{\alpha}(A_k))$  (Ferguson, 1973), o equivalentemente  $(\alpha G_0(A_1), \dots, \alpha G_0(A_k))$  (Rodríguez, 2007).  $G$  es una medida de probabilidad aleatoria en  $(\mathcal{X}, \mathcal{A})$ ,  $G(\emptyset)$  es degenerada en cero,  $G(\mathcal{X})$  es degenerada en 1, y  $G(A)$  toma valores sólo en  $[0, 1]$  (Ferguson, 1973).



### 2.3. El proceso Dirichlet

---

El uso de  $\alpha$  y  $G_0$  en la definición es conveniente por el rol que estos parámetros tienen en describir el DP.  $G_0$  es el centro del DP, tal que  $E[G(A)] = G_0(A)$ . Comúnmente es llamada medida (o distribución) base.  $\alpha$  es un parámetro de concentración: para  $\alpha$  grande hay poca variabilidad en las realizaciones del DP, es decir, entre más grande sea  $\alpha$ , más cercano se espera  $G$  de  $G_0$ . Cuando el DP se usa como una a priori sobre distribuciones en un modelo Bayesiano no paramétrico,  $G$  y  $G_0$  son funciones de distribución en  $\mathcal{X}$ .

Una definición alternativa: el DP, con probabilidad uno, es una medida de probabilidad discreta en  $(\mathcal{X}, \mathcal{A})$ . La idea básica es que, así como la distribución Dirichlet es definida como la distribución conjunta de variables aleatorias gamma independientes divididas por la suma, el DP es definido como la suma de un proceso gamma con incrementos independientes divididos por la suma. Es decir, se define la medida de probabilidad aleatoria  $G$  en  $(\mathcal{X}, \mathcal{A})$  como

$$G(A) = \sum_{j=1}^{\infty} w_j \delta_{\theta_j}(A) \quad (2.2)$$

donde  $w_j \geq 0$ ,  $\sum_1^{\infty} w_j = 1$  con probabilidad uno, y  $\theta_1, \theta_2, \dots$  es una secuencia de v.a. iid con valores en  $\mathcal{X}$ . Las  $w_j$  son construidas tal que

$$w_j = J_j / Z_1$$

donde  $Z_1 = \sum_1^{\infty} J_j$ ,  $Z_1 \in Gam(\alpha, 1)$ . La distribución de las  $J_j$  define un proceso gamma con incrementos independientes (ver Ferguson, 1973 para más detalles sobre esta definición).

Una definición más explícita del DP, semejante a la de la Ec. (2.2), pero donde los pesos  $w_j$  son construidos por un proceso *stick breaking*, caracterizado más adelante, es dada por Sethuraman (1994). De manera breve, la definición se establece considerando dos variables aleatorias independientes,  $V_i \sim Beta(1, \alpha)$  y  $\nu_i^* \sim G_0$ , para  $i = \{1, 2, \dots\}$ , tales que:

$$\begin{aligned} \pi_i(\mathbf{v}) &= v_i \prod_{j=1}^{i-1} (1 - v_j) \\ G &= \sum_{i=1}^{\infty} \pi_i(\mathbf{v}) \delta_{\nu_i^*}. \end{aligned} \quad (2.3)$$

Esta representación del DP hace claro que  $G$  es discreta (con probabilidad uno); el soporte de  $G$  consiste de un conjunto infinito contable de átomos, tomados

## 2.3. El proceso Dirichlet

---

independiente de  $G_0$ . Las proporciones de mezcla  $\pi_i(\mathbf{v})$  son dadas *quebrando* sucesivamente un *stick* de medida uno, en un número infinito de piezas. El tamaño de cada pieza, proporcional al resto del stick, se toma de una distribución  $\text{Be}(1, \alpha)$ .

Otra caracterización del DP puede consultarse en [Ishwaran & Zarepour \(2002\)](#), quienes proponen una representación de suma finita Dirichlet que aproxima al DP, considerando la construcción mediante procesos gamma de [Ferguson \(1973\)](#) y la representación *stick-breaking* de [Sethuraman \(1994\)](#).

En lo sucesivo se usará  $DP(\alpha, G_0)$  para referir a un proceso Dirichlet con parámetros  $\alpha$  y  $G_0$  ya definidos.

### 2.3.2. Propiedades

#### Parámetros de la distribución

El DP es parametrizado por  $\alpha$  (el parámetro de concentración), y una medida base o distribución base,  $G_0$ , que es la media de la distribución. La Figura 2.3 muestra realizaciones del DP con diferentes valores de  $\alpha$ . La columna de la derecha es una realización del proceso, la columna de la izquierda presenta la cdf para 5 realizaciones del proceso. La línea negra es la cdf de la distribución base. En todos los casos se usó  $k = 3$  y distribución base uniforme. La primera fila corresponde a procesos Dirichlet con  $\alpha = 0.1$ , en la segunda  $\alpha = 10$  y la última a  $\alpha = 100$ . La figura ilustra que a medida que  $\alpha$  aumenta, la distribución del proceso se aproxima a la distribución base.

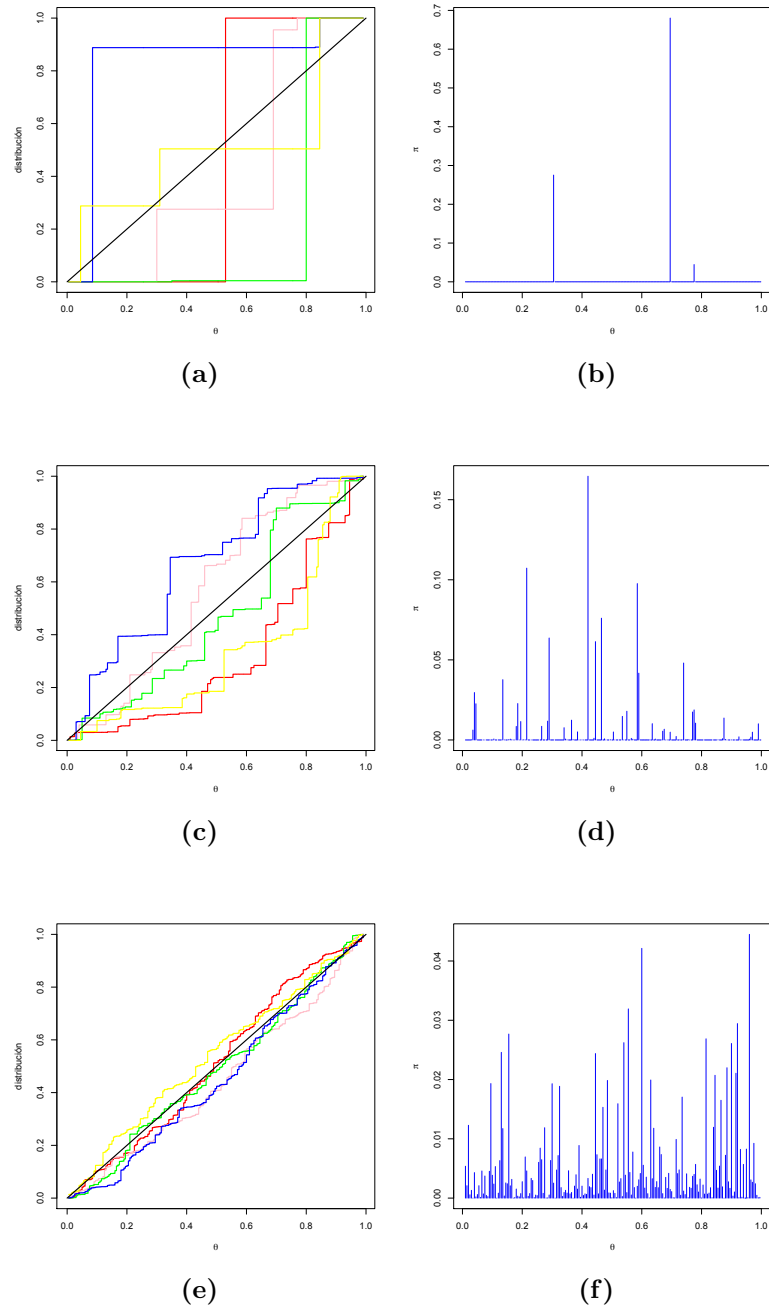
#### Distribución a posteriori

El DP es conjugado para sí mismo, tal que

$$P(G|\theta_1 \dots \theta_n) = DP \left( \alpha + n, \frac{\alpha}{\alpha + n} G_0 + \frac{1}{\alpha + n} \sum_{j=1}^n \delta_{\theta_j} \right)$$

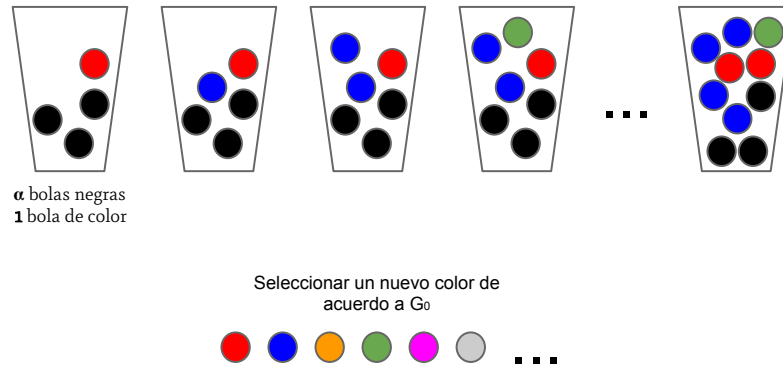
La media a posteriori puede ser interpretada como un promedio ponderado entre la distribución base  $G_0$  (media a priori) y la distribución empírica  $\frac{1}{n} \sum \delta_{\theta_i}$  de los datos observados. La distribución a posteriori converge relativamente rápido a la cdf empírica conforme el tamaño de muestra crece.

## 2.3. El proceso Dirichlet



**Figura 2.3:** Proceso Dirichlet: (a)-(b):  $\alpha = 0.1$ , (c)-(d):  $\alpha = 10$ , (e)-(f):  $\alpha = 100$ .  $\theta$  denota a la variable, y  $\pi$  son los pesos.

## 2.3. El proceso Dirichlet



**Figura 2.4:** Esquema de urnas

### Predictiva a posteriori

El proceso asociado con la distribución predictiva induce un esquema de cluster, la distribución está dada por la siguiente expresión

$$\int p(\theta_{n+1}|G)p(G|\theta_1, \dots, \theta_n)dG =$$

$$\theta_{n+1}|\theta_1, \dots, \theta_n \sim \frac{\alpha}{n + \alpha}G_0(\cdot) + \frac{1}{n + \alpha} \sum_{j=1}^n \delta_{\theta_j}(\cdot) \quad (2.4)$$

Ejemplo: considere una urna que inicialmente tiene  $\alpha$  pelotas negras y una pelota de color (el color fue seleccionado aleatoriamente de acuerdo a  $G_0$ ). Se eligen pelotas de la urna de manera secuencial ( $\theta_i$  representa el color de la  $i$ -ésima pelota seleccionada;  $\theta_i \sim G_0$ ); si se elige una pelota de color, se regresa esa pelota a la urna junto con otra del mismo color; si se elige una pelota negra, regresamos esa pelota a la urna junto con una pelota de un nuevo color aleatoriamente seleccionado de acuerdo a  $G_0$ . Note que, conforme se eligen más y más pelotas de un cierto color, se hace más y más probable elegir una pelota de ese color en las siguientes iteraciones.

El proceso del ejemplo es ilustrado en la Figura 2.4.

## 2.3. El proceso Dirichlet

---

### 2.3.3. Representaciones del DP

Se encuentran en la literatura varios procesos para generar muestras de un DP. Los más comunes se conocen como: urna de Pólya, proceso de restaurant Chino, y *stick-breaking* (ver por ejemplo: Frigyük et al., 2010; Gershman & Blei, 2012; Ishwaran & Zarepour, 2002). En esta sección se describen los dos últimos, ya que, en la práctica, urna de Pólya y el proceso de restaurant Chino son diferentes nombres del mismo proceso.

#### Proceso de restaurant Chino

Considere un restaurant con un número infinito de mesas. En cada mesa hay un número infinito de sillas. El restaurant abre, clientes comienzan a llegar uno a uno y eligen una mesa. Es más probable que un cliente elija una mesa si en esta ya hay muchas personas sentadas. Sin embargo, con probabilidad proporcional a  $\alpha$ , el cliente se sentará en una nueva mesa.

Se define la siguiente notación:

- $K$  = número de mesas ocupadas.
- $n_k$  = número de clientes sentados en la mesa  $k$ ,  $k = 1, \dots, K$ , donde  $\sum n_k = N$
- A cada mesa  $k$  se le asigna  $\theta_k \sim G_0$ .
- Los clientes sentados en una mesa dada, comparten los mismos atributos definidos por  $\theta_k$ .

El primer cliente entra al restaurant y se sienta en una mesa. Para esa mesa, se toma  $\theta_1$  de  $G_0$ . En este momento,  $N = 1$  (hay sólo un cliente en el restaurant),  $K = 1$  (hay sólo una mesa ocupada), y  $n_1 = 1$  (un cliente está sentado en la mesa 1). El segundo cliente entra y elige una mesa. Con probabilidad  $1/(1 + \alpha)$  se sienta en la misma mesa ocupada por el cliente 1, y con probabilidad  $\alpha/(1 + \alpha)$  se sienta en una nueva mesa. Si eligió una nueva mesa,  $\theta_2$  tomado de  $G_0$  es asignado al cliente, y  $n_2 = 1$ . De otro modo,  $\theta_1$  es asignado al cliente, y  $n_1 = 2$ . Ahora hay dos clientes en el restaurant,  $N = 2$ . El proceso se realiza sucesivamente. Después de  $N$  pasos, el resultado del CRP es una partición de  $N$  clientes en  $K$  mesas, o equivalentemente, una partición de los números naturales  $1, 2, \dots, N$  en  $K$  conjuntos.

### 2.3. El proceso Dirichlet

---

El algoritmo que describe el proceso es el siguiente:

El cliente 1 entra al restaurant y se sienta en la mesa 1.

$\phi_1 = \theta_1$  donde  $\theta_1 \sim G_0$ ,  $K = 1$ ,  $N = 1$ ,  $n_1 = 1$

**for**  $N = 2, \dots$

$$N \text{ se sienta en la mesa } \begin{cases} k & \text{con prob } \frac{n_k}{N-1+\alpha}, \quad k = 1, \dots, K \\ K + 1 & \text{con prob } \frac{\alpha}{N-1+\alpha} \quad (\text{nueva mesa}) \end{cases}$$

**if** fue elegida una nueva mesa **then**  $K = K + 1$ ,  $\theta_{K+1} \sim G_0$

**else** asignar  $\theta_k$  de la mesa  $k$  en la que el cliente  $N$  se sentó; hacer  $n_k = n_k + 1$

La correspondiente distribución inducida por el CRP es invariable ante permutaciones, es decir,  $\theta_1, \theta_2, \dots$  son intercambiables. La función es equivalente a la Ec. (2.4):

$$\theta_N | \theta_1, \dots, \theta_{N-1}, G_0, \alpha \sim \frac{\alpha}{N-1+\alpha} G_0 + \frac{\sum n_k \delta_{\theta_k}}{N-1+\alpha}$$

La Figura (2.5) presenta muestras generadas por un DP con distribución base  $N(0, 1)$  y distintos valores de  $\alpha$ . El CRP hace evidente la tendencia a formar clusters, y que esto puede ser controlado por  $\alpha$ . Conforme  $\alpha \rightarrow 0$ , se hace más y más probable que todos los  $\theta$  sean iguales; conforme  $\alpha \rightarrow \infty$ , el número de clusters se incrementa. Los dos casos limite serían: un cluster cuando  $\alpha \rightarrow 0$ , y  $N$  clusters (uno asociado con cada observación), cuando  $\alpha \rightarrow \infty$ . Además,  $\alpha$  determina la concentración.

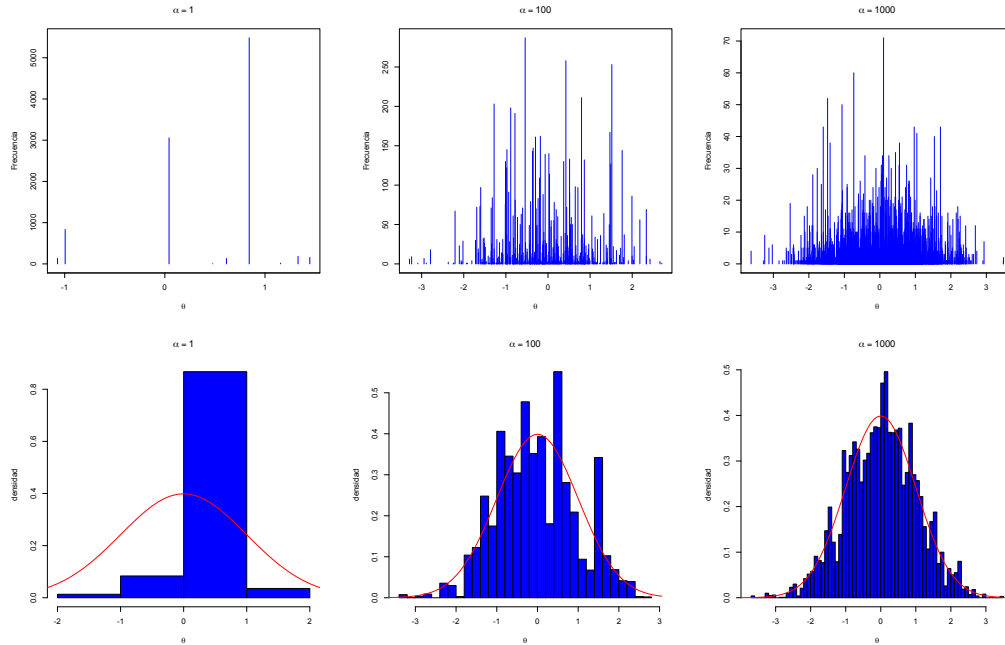
#### *Stick-breaking*

Este proceso trabaja con un *stick* de longitud 1. Suponga que se genera una secuencia de pesos  $\{\pi_k\}_{k=1}^{\infty}$  de acuerdo con lo siguiente:

$$\begin{aligned} \pi_k &= \beta_k \prod_{j=1}^{k-1} (1 - \beta_j) \\ G &= \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k} \end{aligned} \tag{2.5}$$

donde:  $\beta_k \sim \text{Beta}(1, \alpha)$ ,  $\theta_k \sim G_0$  y  $\sum_{k=1}^{\infty} \pi_k = 1$ . El proceso se resume como sigue:

### 2.3. El proceso Dirichlet



**Figura 2.5:** Muestras [DP](#) generadas por un CRP con distintos valores de  $\alpha$

- Tomar una porción aleatoria  $\beta_1$  del stick. La longitud de esta pieza dará el primer *peso*,  $\pi_1$ .
- Tomar  $\theta_1$  de  $G_0$ .
- Del *stick* sobrante, tomar una porción aleatoria  $\beta_2$ . Calcular la longitud del segundo *peso*,  $\pi_2$ .
- Tomar  $\theta_2$  de  $G_0$ .
- y así sucesivamente ...

Conforme  $k$  crece, la longitud del stick, o los pesos, se hacen más pequeños. El parámetro de concentración  $\alpha$  determina la distribución de la longitud del stick. Para  $\alpha$  pequeño, sólo los primeros *sticks* tendrán longitud significativa, el resto de *sticks* tendrán longitud muy pequeña. Para  $\alpha$  grande, la longitud de los *sticks* tenderá a ser más uniforme (ver Figura 2.6).

La Figura (2.7) muestra realizaciones del [DP](#) con diferente  $\alpha$  y distribución base  $\text{Gam}(2,2)$ . La distribución base  $G_0$  determina la localización del *stick*, y  $\alpha$  controla la distribución de la longitud. Con  $\alpha$  pequeño, la longitud se concentra en unos pocos *sticks*.

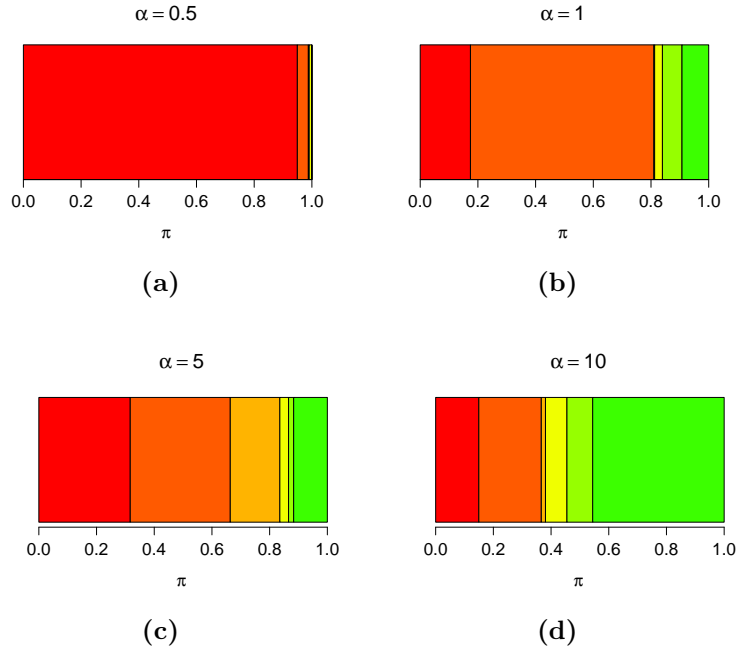


Figura 2.6: *Stick-breaking* con distintos valores de  $\alpha$

## 2.4. Proceso Dirichlet para modelos de mezclas

En el contexto de modelos de mezclas, se usa un DP como distribución a priori sobre los parámetros de los componentes de la mezcla. El modelo resultante se denomina proceso Dirichlet para modelos de mezclas (DPMM). En otras palabras, un DPMM es un modelo de mezclas donde la distribución de la mezcla es desconocida, y se asigna a ella un DP a priori. El objetivo es modelar un conjunto de datos que se presume no provienen todos de la misma distribución, pero que se desconoce cuántas distribuciones caracterizan a las observaciones y cómo están definidas. Por ejemplo, la estructura de una mezcla de  $k = 2$  densidades Gaussianas es dada por:

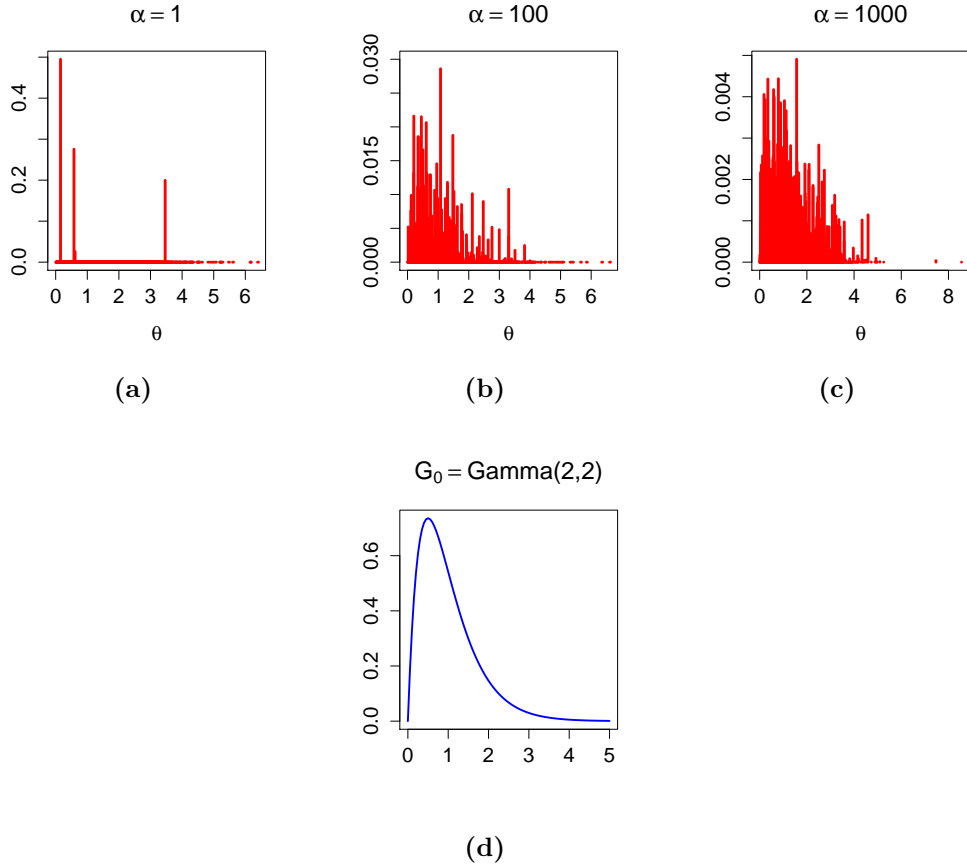
$$y_i | w, \{\mu_k\}_{k=1}^2, \{\sigma_k^2\}_{k=1}^2 \sim wN(y_i; \mu_1, \sigma_1^2) + (1 - w)N(y_i; \mu_2, \sigma_2^2) \quad (2.6)$$

es decir, para cada  $i = 1, \dots, n$  independientes, la observación  $y_i$  proviene de una distribución  $N(\mu_1, \sigma_1^2)$  con probabilidad  $w$ , o de una distribución  $N(\mu_2, \sigma_2^2)$  con probabilidad  $1 - w$ . La densidad en la Figura (2.1)-(a) ilustra la Ec. (2.6). Adicionalmente, en el escenario Bayesiano se establece una distribución a priori para los parámetros desconocidos de la mezcla, como

$$(w, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2) \sim p(w, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$$



## 2.4. Proceso Dirichlet para modelos de mezclas



**Figura 2.7:** Realizaciones del proceso Dirichlet generadas con Stick Breaking

Otra forma de establecer el modelo (2.6) es mediante variables aleatorias auxiliares  $c_1, \dots, c_n$  tal que  $c_i = 1$  si  $y_i$  proviene del componente  $N(\mu_1, \sigma_1^2)$ , y  $c_i = 2$  si  $y_i$  es toma de la  $N(\mu_2, \sigma_2^2)$ . Entonces,

$$\begin{aligned}
 y_i | w, \{\mu_k\}_{k=1}^2, \{\sigma_k^2\}_{k=1}^2 &\sim N(y_i; \mu_{c_i}, \sigma_{c_i}^2) \\
 P(c_i = 1 | w) &= w = 1 - P(c_i = 2 | w) \\
 (w, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2) &\sim p(w, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)
 \end{aligned} \tag{2.7}$$

Si se marginaliza sobre  $c_i$  se regresa a la formulación original. Otra manera es escribir

$$wN(y_i; \mu_1, \sigma_1^2) + (1 - w)N(y_i; \mu_2, \sigma_2^2) = \int N(y_i; \mu, \sigma^2) dG(\mu, \sigma^2), \tag{2.8}$$

donde

$$G(\cdot) = w\delta_{(\mu_1, \sigma_1^2)}(\cdot) + (1 - w)\delta_{(\mu_2, \sigma_2^2)}(\cdot)$$

## 2.4. Proceso Dirichlet para modelos de mezclas

---

y  $\delta_x(\cdot)$  denota la delta de Dirac. Expresiones similares a (2.6), (2.7), y (2.8) se pueden crear para el caso general de  $k$  componentes en la mezcla. En la Ec. (2.8)  $G$  es discreta (y aleatoria), por lo que una alternativa es usar un DP como a priori para  $G$ , resultando en un DPMM (Antoniak, 1974). Usar un DP como a priori permite a los datos decidir cuántos componentes son apropiados para describirlos.

Un DPMM es descrito por el modelo jerárquico

$$\begin{aligned} G &\sim DP(\alpha, G_0) \\ \theta_i | G &\sim G \\ y_i | \theta_i &\sim F(\theta_i) \end{aligned} \tag{2.9}$$

donde  $y_i$  son las variables que se desean modelar, tomadas independientemente de una distribución desconocida.  $\theta_i$  son los parámetros del componente de mezcla (puede ser uno, o un vector de múltiples parámetros) que corresponde a  $y_i$ ,  $F$  representa la distribución de los componentes de mezcla,  $G_0$  es la distribución base, y  $\alpha$  es un parámetro de precisión que determina la concentración de  $G$  sobre  $G_0$ . Con frecuencia  $F$  y  $G_0$  dependen de hiperparámetros adicionales, por lo que el modelo puede extenderse para incluir distribuciones a priori para estos parámetros y para  $\alpha$ .

Si los  $\theta$  fueran tomados, por ejemplo, de una Gaussiana, ningún valor sería igual, pero como se toman de un DP, algunos componentes del vector  $\theta$  son iguales, de tal manera que el total de valores diferentes en ese vector es igual al número de componentes en la mezcla. Por lo tanto, cuando dos observaciones  $y_i$  y  $y_j$  pertenecen al mismo componente, entonces  $\theta_i = \theta_j$ .

Introduciendo variables aleatorias auxiliares, el modelo (2.9) puede ser reescrito como

$$\begin{aligned} y_i | c_i, \phi &\sim F(\phi_{c_i}) \\ c_i | w &\sim Discrete(w_1, \dots, w_k) \\ \phi_c &\sim G_0 \\ w &\sim Dirichlet(\alpha/k, \dots, \alpha/k) \end{aligned} \tag{2.10}$$

donde  $c_i$  indica cuál clase está asociada con la observación  $y_i$ . Para cada clase  $c$ , los parámetros  $\phi_c$  determinan la distribución de las observaciones de esa clase.

## 2.4. Proceso Dirichlet para modelos de mezclas

La a priori para  $c_i$  es

$$\begin{aligned}
 P(c_i = c | c_1, \dots, c_{i-1}) &= P(c_1, \dots, c_{i-1}, c_i = c) / P(c_1, \dots, c_{i-1}) \\
 &= \frac{\int w_{c_1} \cdots w_{c_{i-1}} w_c \frac{\Gamma(\sum_1^k \frac{\alpha}{k})}{\prod_1^k \Gamma(\alpha/k)} \prod_{i=1}^k w_i^{\alpha/k-1} dw}{\int w_{c_1} \cdots w_{c_{i-1}} \frac{\Gamma(\sum_1^k \frac{\alpha}{k})}{\prod_1^k \Gamma(\alpha/k)} \prod_{i=1}^k w_i^{\alpha/k-1} dw} \\
 &= \frac{n_{i,c} + \alpha/k}{i - 1 + \alpha}
 \end{aligned}$$

donde  $n_{i,c}$  es el número de  $c_j$ , para  $j < i$ , que son iguales a  $c$ . Haciendo  $k \rightarrow \infty$ , las probabilidades condicionales tienen los siguientes límites (ver [Neal, 2000](#)):

$$\begin{aligned}
 P(c_i = c | c_1, \dots, c_{i-1}) &\rightarrow \frac{n_{i,c}}{i - 1 + \alpha} \\
 P(c_i \neq c_j \text{ for all } j < i | c_1, \dots, c_{i-1}) &\rightarrow \frac{\alpha}{i - 1 + \alpha}
 \end{aligned} \tag{2.11}$$

Diversos métodos MCMC para muestrear de la distribución a posteriori de un DPMM son revisados en [Neal \(2000\)](#). El paper incluye algoritmos que son de fácil implementación para modelos basados en distribuciones a priori conjugadas y no conjugadas. En esta sección se ilustran dos de los algoritmos para el caso conjugado; el primero corresponde también al usado por [Escobar \(1994\)](#) y [Escobar & West \(1995\)](#). Un tercer ejemplo en el contexto de regresión se usa para presentar el algoritmo propuesto por [Ishwaran & James \(2001\)](#). Otras propuestas de muestreo para el DPMM no revisadas aquí, pueden consultarse en [Blei & Jordan \(2006\)](#), [Jain & Neal \(2004\)](#), [Walker \(2007\)](#) y [Kalli et al. \(2011\)](#).

### 2.4.1. Ejemplo 1: datos simulados

Suponga que  $y_1, \dots, y_n$  son condicionalmente independientes y normalmente distribuidas,  $y_i | \theta_i \sim N(\mu_i, V_i)$ , con medias  $\mu_i$  y varianzas  $V_i$ . En el DPMM descrito por el modelo (2.9),  $\theta_i = (\mu_i, V_i)$ ,  $i = 1, \dots, n$ . Suponga además que las medias y varianzas vienen de alguna distribución a priori  $G_0(\cdot)$ . Si  $G(\cdot)$  es modelado como un proceso Dirichlet, entonces los datos vienen de una mezcla Dirichlet de normales ([Escobar & West, 1995](#)). Bajo esta configuración, el modelo jerárquico puede verse como en (2.12). El objetivo es conocer el número  $k$  de distintos valores de  $(\mu, V)$ , y qué distribución caracteriza a cada una de las observaciones.

$$\begin{aligned}
 G &\sim DP(\alpha, G_0) \\
 (\mu_i, V_i) | G &\sim G \\
 y_i | \mu_i, V_i &\sim N(\mu_i, V_i)
 \end{aligned} \tag{2.12}$$

## 2.4. Proceso Dirichlet para modelos de mezclas

---

Una forma conveniente para  $G_0$  cuando las observaciones son normalmente distribuidas, con media y varianza desconocidas, es la Normal-Inversa Gamma, o Normal-Gamma si se usa precisión en lugar de varianza. Entonces, bajo  $G_0$ :

$$\begin{aligned} V_i^{-1} &\sim \text{Ga}(a/2, 2/b) \\ \mu_i | V_i &\sim N(m_0, V_i/\rho) \end{aligned}$$

donde  $a/2$  y  $2/b$  son parámetros de forma y escala, respectivamente, y  $\rho > 0$  es un factor de escala. Esta elección de  $G_0$  es propuesta por [Ferguson \(1983\)](#). Una guía para los valores de los hiperparámetros puede consultarse en la misma referencia.

### Algoritmo 1: [Escobar & West \(1995\)](#)

Una aproximación para muestrear del modelo (2.12) es propuesta por [Escobar & West \(1995\)](#). El método consiste en tomar repetidamente valores para cada  $\theta_i = (\mu_i, V_i)$  de su distribución condicional dados los datos y  $\theta_{-i}$ . Esta distribución condicional se obtiene combinando la verosimilitud para  $y_i$ , y la a priori condicional a  $\theta_{-i}$ , dada por la Ec. (2.13) ([Blackwell & MacQueen, 1973](#)).

$$\theta_i | \theta_{-i} \sim \frac{\alpha}{\alpha + n - 1} G_0(\theta_i) + \frac{1}{\alpha + n - 1} \sum_{j=1, j \neq i}^n \delta_{\theta_j}(\theta_i) \quad (2.13)$$

Como las  $\theta_i$  son iid  $\sim G$ , su distribución conjunta es invariante ante permutaciones, por lo que  $\theta_1, \theta_2, \dots$  son intercambiables. Es fácil notar que, cuando se combina la verosimilitud con la Ec. (2.13), la distribución condicional para usar en el muestreo Gibbs está dada por

$$\theta_i | \theta_{-i}, y_i \sim q_0 G_i(\theta_i) + \sum_{j=1, j \neq i}^n q_j \delta_{\theta_j}(\theta_i) \quad (2.14)$$

donde  $\sum_{j \neq i} q_j + q_0 = 1$ ,  $G_i(\cdot)$  es la distribución a posteriori para  $\theta$  bajo la a priori  $G_0$  y la observación  $y_i$ .  $q_j$  y  $q_0$  están definidas por

$$q_0 \propto \alpha \int F(y_i, \theta) dG_0(\theta) \quad (2.15)$$

$$q_j \propto F(y_i, \theta_j) \quad (2.16)$$

En el caso de Mezclas Dirichlet de Normales,  $q_0$  es proporcional a  $\alpha$  veces la función de densidad de  $T_\alpha(m_0, M)$ , donde  $M = (1 + \rho)b/a\rho$  (ver Anexo A), y  $q_j$  es

## 2.4. Proceso Dirichlet para modelos de mezclas

proporcional a la verosimilitud de los datos  $y_i$ , esto es, una muestra de la función de densidad  $N(\mu_j, V_j)$  en  $y_i$ , con  $j = 1, \dots, n^*$ , donde  $n^*$  es el número de valores distintos de  $\theta = (\mu, V)$ . La distribución a posteriori  $G_i(\cdot)$  resulta Normal-Gamma; su derivación es fácil, y puede consultarse en el Anexo A.

Cuando  $G_0$  es a priori conjugada para la verosimilitud dada por  $F$ , como en el caso de la Normal-Gamma, la integral en (2.15) es factible, de otro modo, puede ser analíticamente intratable.

Para muestrear de la distribución dada en (2.14), Escobar (1994), Escobar & West (1995) y Neal (2000) usan el Algoritmo siguiente:

1. Elegir valores iniciales para los hiperparámetros y muestrear para cada  $i = 1, \dots, n$ , un valor inicial de la a posteriori  $G_i(\cdot)$  dada en la Ec. (2.14).
2. Asignar cada observación a un cluster  $j$ ,  $j \in \{1, \dots, n\}$ .
3. Para cada  $i = 1, \dots, n$ : actualizar los valores de  $\theta_i | \theta_{-i}, y_i$  de la Ec. (2.14).
4. Regresar al paso 3 y repetir hasta convergencia.

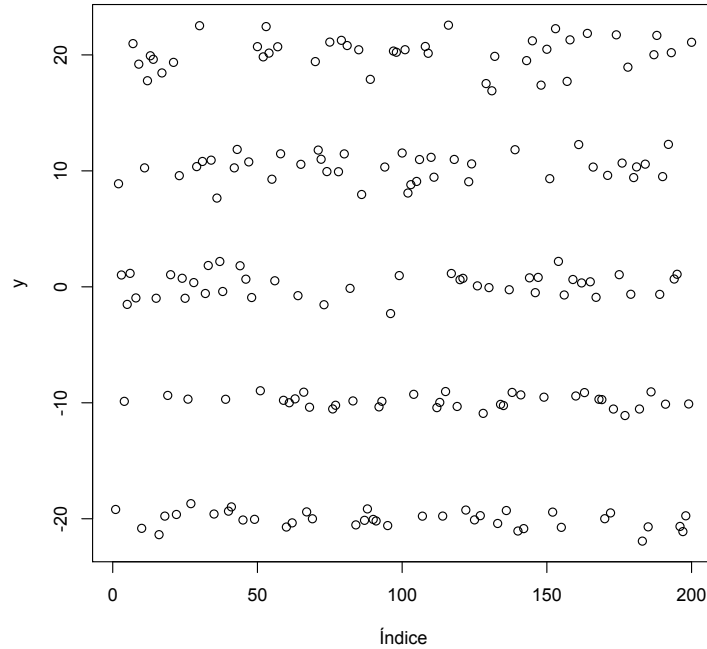
Para ilustrar el método, se simularon 200 observaciones provenientes de  $k = 5$  distribuciones normales, con probabilidad igual de pertenecer a cada uno de los  $k$  clusters. Las especificaciones se resumen en la Tabla 2.1. Las proporciones de mezcla  $w$  se refieren al porcentaje de observaciones en cada cluster. Los datos simulados se muestran en la Figura (2.8). Como se puede observar, los datos tienen una evidente agrupación, sin embargo, el ejemplo pretende ilustrar el desempeño del método en un caso simple.

**Tabla 2.1:** Datos simulados. Algoritmo 1: Escobar & West (1995)

Grupo	Media	Varianza	$w$
1	-20	0.500	0.195
2	-10	0.675	0.180
3	0	0.850	0.210
4	10	1.025	0.205
5	20	1.200	0.210
$k = 5$			1.000

## 2.4. Proceso Dirichlet para modelos de mezclas

---



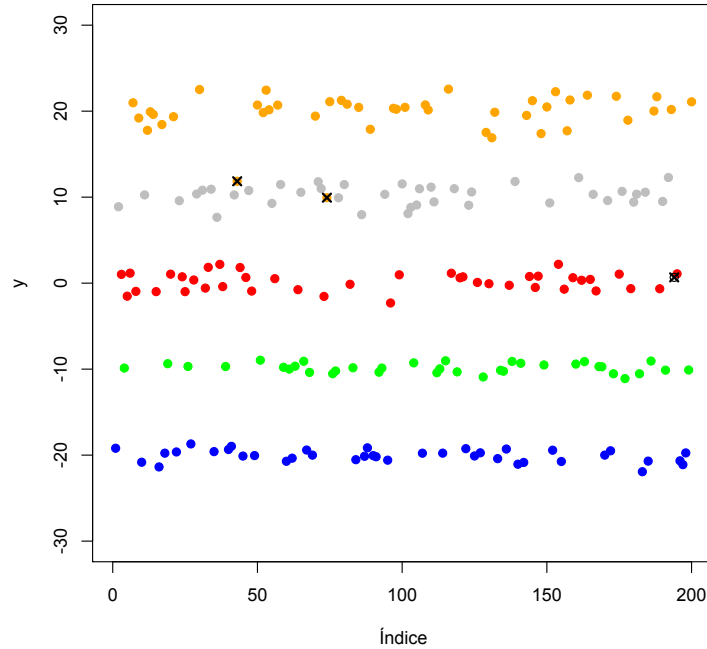
**Figura 2.8:** Datos simulados. Algoritmo 1: [Escobar & West \(1995\)](#)

La Figura (2.9) muestra el resultado de agrupación del algoritmo para los datos simulados. Se tomaron como valores de los parámetros:  $\alpha = 0.1$ ,  $a = 1$ ,  $b = 0.01$  y  $\rho = 0.1$ . El tiempo de ejecución fue de 11:07.19 para 10000 repeticiones. En la Figura, los grupos encontrados en la última iteración se muestran por colores. Los puntos marcados con cruz corresponden a observaciones mal agrupadas. El punto sin color es una observación no clasificada en alguno de los 5 grupos.

La Tabla 2.2 muestra las estimaciones. La media, varianza, y  $w$  corresponden a la media, varianza y proporción muestrales, calculadas con las observaciones clasificadas en cada grupo en la última iteración. La Figura 2.10 muestra la densidad estimada con los valores de la Tabla 2.2 (línea roja), y la densidad con los valores de la Tabla 2.1 (línea azul), usando un grid de 500 puntos.

Otras estimaciones para la media de los grupos se dan en la Tabla 2.3: la primera calculada con el promedio de la media a posteriori de las últimas 5000 iteraciones, y la segunda la media a posteriori de la última iteración.

## 2.4. Proceso Dirichlet para modelos de mezclas



**Figura 2.9:** Agrupación con el Algoritmo 1: Escobar & West (1995)

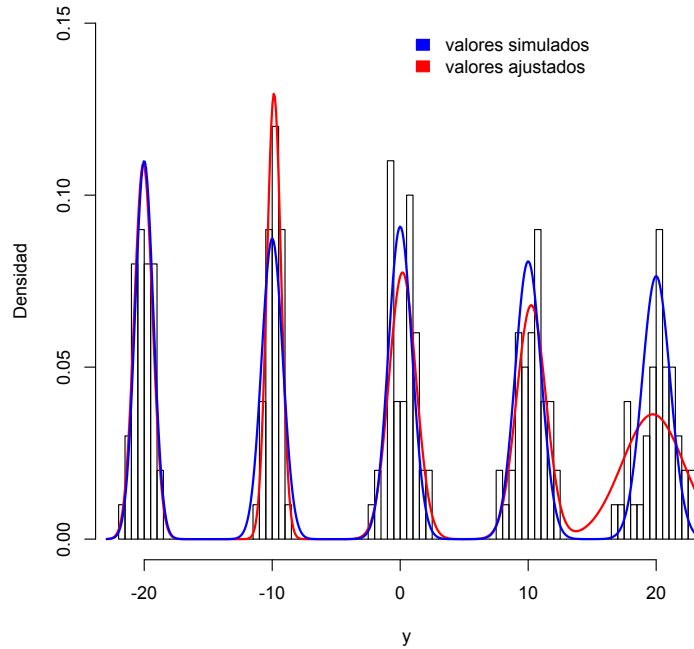
### Algoritmo 2: Neal (2000) - Algoritmo 2

El problema con el método de Escobar & West (1995) es que actualiza cada  $\theta$  uno a la vez, por lo que la convergencia a la distribución a posteriori puede ser lenta (Neal, 2000). Para evitar esto, Neal (2000) aplica Gibbs sampling al modelo formulado en (2.10). Así, en lugar de muestrear  $\theta_1, \dots, \theta_n$  directamente, se muestrea  $c_1, \dots, c_n$ , y  $\phi_c$  para toda  $c \in \{c_1, \dots, c_n\}$ . El muestreo Gibbs para  $c_i$

**Tabla 2.2:** Valores ajustados. Algoritmo 1: Escobar & West (1995)

Grupo	Media	Varianza	$w$
1	-20.07	0.507	0.195
2	-9.86	0.307	0.180
3	0.18	1.112	0.205
4	10.24	1.308	0.195
5	19.74	5.85	0.220
$k = 5$			0.995

## 2.4. Proceso Dirichlet para modelos de mezclas



**Figura 2.10:** Densidad estimada. Algoritmo 1: [Escobar & West \(1995\)](#)

se basa en las siguientes probabilidades condicionales ([Neal, 2000](#)):

$$\text{If } c = c_j \text{ for some } j \neq i : P(c_i = c | c_{-i}, y_i, \phi) \propto \frac{n_{-i,c}}{n-1+\alpha} F(y_i, \phi_c) \quad (2.17)$$

$$P(c_i \neq c_j \text{ for all } j \neq i | c_{-i}, y_i, \phi) \propto \frac{\alpha}{n-1+\alpha} \int F(y_i, \phi) dG_0(\phi) \quad (2.18)$$

donde  $\phi$  es el conjunto de  $\phi_c$  asociados con al menos una observación. Si el muestreo para  $c_i$  elige un nuevo valor (distinto de los demás), entonces  $\phi_{c_i}$  se toma de

**Tabla 2.3:** Otras estimaciones para la media. Algoritmo 1: [Escobar & West \(1995\)](#)

Grupo	5000 iteraciones	Última iteración
1	-19.96	-19.96
2	-9.13	-9.13
3	0.13	0.17
4	6.06	11.79
5	15.24	20.68



## 2.4. Proceso Dirichlet para modelos de mezclas

---

$G_i$ . También aquí  $G_0$  es conjugada, lo que permite calcular  $\int F(y_i, \phi) dG_0(\phi)$  y muestrear de  $G_i$ . El algoritmo es el siguiente

## 2.4. Proceso Dirichlet para modelos de mezclas

Asuma que el estado actual de la cadena de Markov consiste de  $\mathbf{c} = (c_1, \dots, c_n)$ , y considere los parámetros  $\phi_c$ ,  $\phi = (\phi_c : c \in \{c_1, \dots, c_n\})$ .  
Para  $i = 1, \dots, n$ :

1. Si el valor actual de  $c_i$  no está asociado con ninguna otra observación, remover  $\phi_{c_i}$  del estado.
2. Tomar un nuevo valor para  $c_i$  de la distribución condicional dada en las ecuaciones (2.17) y (2.18).
3. Si el nuevo  $c_i$  no está asociado con ninguna otra observación, tomar un valor para  $\phi_{c_i}$  de

$$G_i(\phi) = P(\phi|y_i) \propto F(y_i, \phi)G_0(\phi) \quad (2.19)$$

Para todo  $c \in \{c_1, \dots, c_n\}$ :

1. Tomar un nuevo valor de  $\phi_c$  de la distribución a posteriori, basada en la a priori  $G_0$  y todas las observaciones asociadas con el componente  $c$ , esto es, para las cuales  $c_i = c$ :

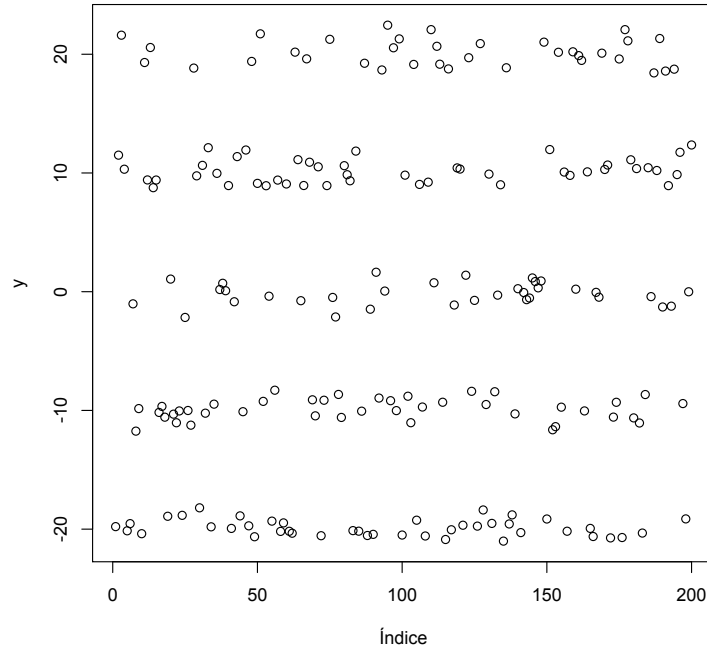
$$P(\phi|\mathbf{y}_c) \propto G_0(\phi) \prod_{i:c_i=c} F(y_i, \phi) \quad (2.20)$$

Para ilustrar el desempeño de este Algoritmo, se simularon 200 observaciones con las mismas características del ejemplo anterior. La Tabla 2.4 y la Figura 2.11 resumen los datos. El algoritmo se ejecutó 10000 veces usando  $\alpha = 0.1$ ,  $a = 1$ ,  $b = 0.01$  y  $\rho = 0.1$ , como en el caso anterior. El tiempo de ejecución fue 02:17.41. La Figura 2.12 muestra el resultado de la agrupación, diferenciando los grupos resultantes por colores. Todas las observaciones fueron clasificadas correctamente, y el tiempo de ejecución del algoritmo fue considerablemente menor al empleado por Escobar & West (1995).

**Tabla 2.4:** Datos simulados. Algoritmo 2: Neal (2000)

Grupo	Media	Varianza	$w$
1	-20	0.500	0.215
2	-10	0.675	0.210
3	0	0.850	0.170
4	10	1.025	0.230
5	20	1.200	0.175
$k = 5$			1.000

## 2.4. Proceso Dirichlet para modelos de mezclas



**Figura 2.11:** Datos simulados. Algoritmo 2: Neal (2000)

La Tabla 2.5 muestra las estimaciones muestrales para la media, varianza y proporción, calculadas con los grupos de la última iteración. La Tabla 2.6 muestra la estimación para la media calculada con el promedio de la media a posteriori de las últimas 5000 iteraciones, así como la media a posteriori de la última iteración. La línea roja de la Fig 2.11 es la densidad estimada con los valores de la Tabla 2.4, y la línea azul es la densidad estimada con los valores de la Tabla 2.5.

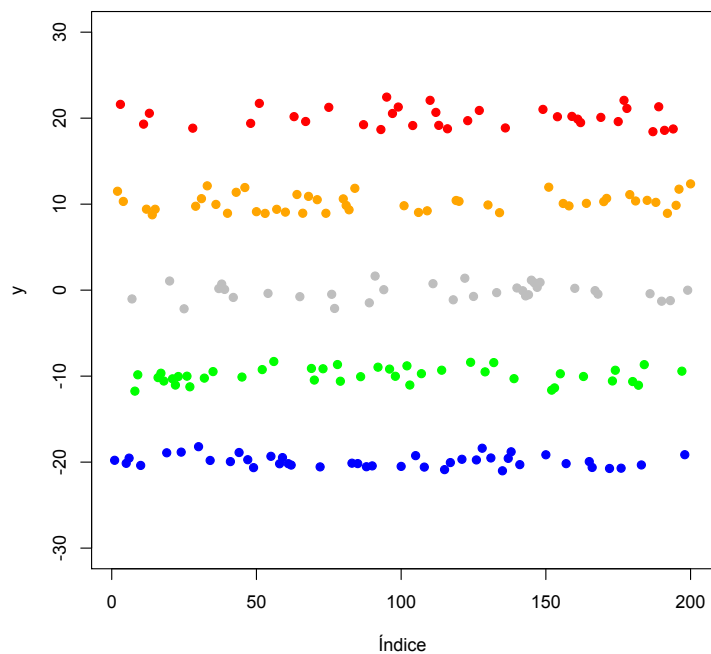
**Tabla 2.5:** Valores ajustados. Algoritmo 2: Neal (2000)

Grupo	Media	Varianza	$w$
1	-19.89	0.473	0.215
2	-9.91	0.801	0.210
3	-0.19	0.878	0.170
4	10.18	1.010	0.230
5	20.13	1.304	0.175
$k = 5$			1.000

Note que en los algoritmos presentados se usaron valores fijos para los hiperparámetros y para el parámetro de concentración  $\alpha$ . Sin embargo, se puede asignar

## 2.4. Proceso Dirichlet para modelos de mezclas

---



**Figura 2.12:** Agrupación con el Algoritmo 2: Neal (2000)

a estos una distribución a priori y actualizarlos de su distribución condicional. El procedimiento puede consultarse en Escobar (1994), Escobar & West (1995), Görür & Rasmussen (2010) y West (1992).

**Tabla 2.6:** Otras estimaciones para la media. Algoritmo 2: Neal (2000)

Grupo	5000 iteraciones	Última iteración
1	-19.28	-19.98
2	-8.81	-9.86
3	0.87	-0.44
4	10.54	10.02
5	19.30	19.90

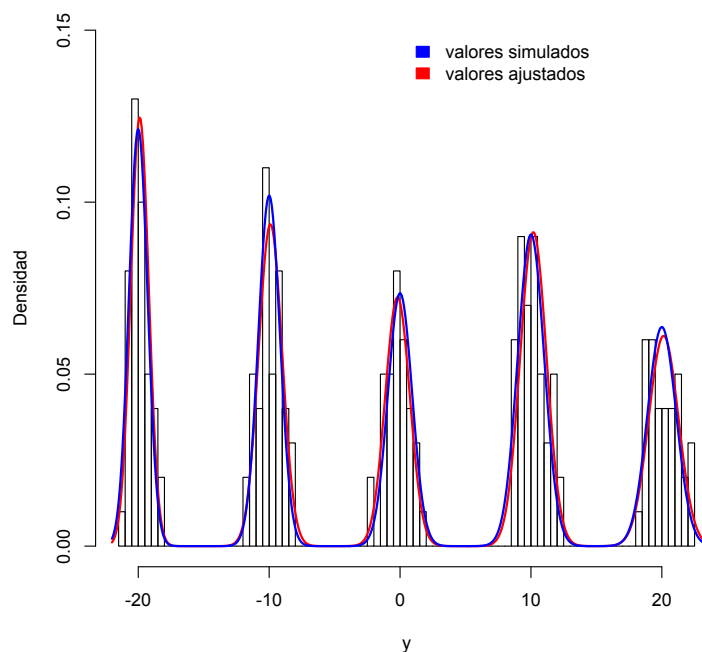


Figura 2.13: Agrupación con el Algoritmo 2: Neal (2000)

### 2.4.2. Ejemplo 2: geyser Old Faithful

Un Geyser es un tipo especial de fuente termal que se caracteriza por la descarga intermitente de agua expulsada de forma turbulenta y acompañada de vapor. Los geyser son usado para varias actividades económicas, como generación de electricidad, calefacción y turismo. En el mundo se encuentran muchas reservas geotermales que los albergan, una de las más famosas es el Yellowstone National Park en Wyoming, Estados Unidos, en donde se encuentra el geyser Old Faithful, uno de los geyser más predecibles.

Del 1 al 15 de Agosto de 1985, el geyser Old Faithful fue observado y se registraron los tiempos de espera entre erupciones sucesivas y la duración de cada erupción. Dos versiones de los datos se encuentran en R. [Everitt & Hothorn \(2010\)](#) usaron la versión con 272 observaciones para ajustar una estimación paramétrica de la densidad de los tiempos de espera entre erupciones, basada en un modelo de mezclas de dos componentes normales. Usando los mismos datos, se ajustó un [DPMM](#) con el Algoritmo 2 de [Neal \(2000\)](#), con parámetros:  $\alpha = 0.1$ ,  $a = 1$ ,  $b = 0.01$ ,  $\rho = 0.1$ , y 10000 réplicas. La [Tabla 2.7](#) compara los resultados de los dos ajustes. Para el Algoritmo de [Neal \(2000\)](#), los valores que se reportan son el

## 2.4. Proceso Dirichlet para modelos de mezclas

resultado de la última iteración. Otras estimaciones para la media se presentan en la Tabla 2.8, calculadas como el promedio de la media a posteriori después de convergencia, y el promedio de las observaciones clasificadas en cada grupo en la última iteración. La Figura 2.14 muestra las observaciones y la agrupación resultante.

**Tabla 2.7:** Tiempos de espera entre erupciones: geyser Old Faithful

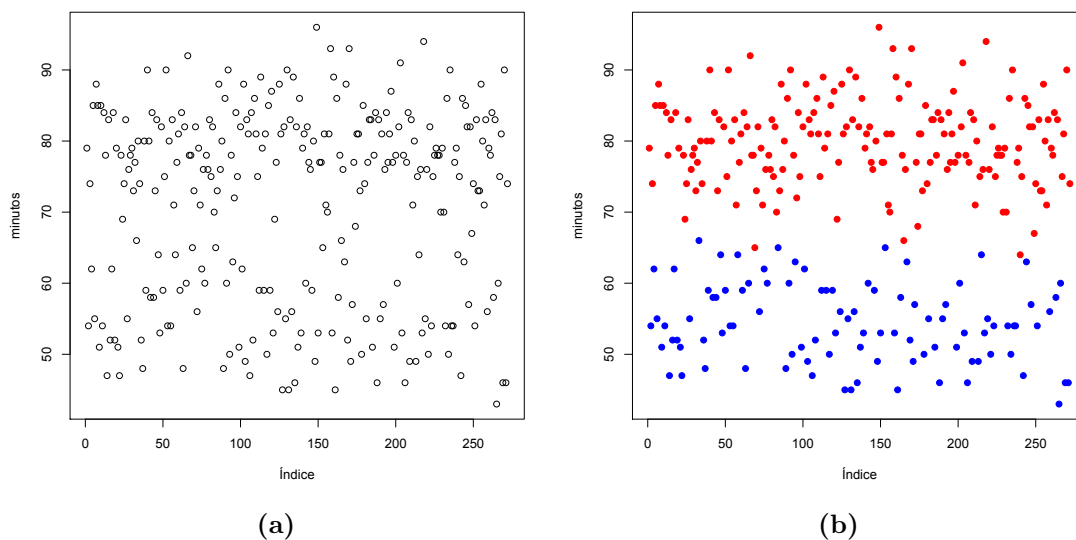
	Everitt & Hothorn (2010)	Neal (2000)
No. Grupos	2(fijos)	2
Media	54.63 80.1	54.56 79.88
Varianza	5.90 5.87	5.78 5.93
$w$	0.36 0.64	0.35 0.65

**Tabla 2.8:** Media del tiempo de espera entre erupciones: geyser Old Faithful

	Últimas 5000 iteraciones	Media observada
Grupo 1	54.74	54.30
Grupo 2	73.29	79.95

En los dos algoritmos presentados hasta ahora,  $G_0$  se caracteriza por un mecanismo urna de Pólya que implica tomar muestras de la distribución a posteriori de un modelo jerárquico, que resulta en una regla de predicción explícita. La regla de predicción se define como la distribución condicional de una observación futura dados los valores previos muestreados de la a priori (Ishwaran & James, 2001). Las Ec. (2.14) en el algoritmo de Escobar & West (1995), y (2.17)-(2.18) del algoritmo de Neal (2000), describen la regla de predicción necesaria para el muestreo.

Aunque el muestreador Gibbs urna de Pólya es un método versátil para ajustar modelos jerárquicos Bayesianos, hay algunas limitaciones con esta aproximación: 1) la actualización del conjunto  $\theta$  uno a la vez; 2) calcular  $q_0$  de la Ec. (2.14), o equivalentemente, la integral en (2.18) es complicada en el caso no conjugado; 3) la inferencia sobre  $G$  está basada sólo en los valores  $\theta_i$  a posteriori (Ishwaran & James, 2001). Para evitar estos problemas, Ishwaran & James (2001) proponen un muestreo Gibbs por bloques, aplicable a modelos similares a (2.9), y basado en la representación *stick-breaking* del DP. El método se describe e ilustra en la siguiente sección.



**Figura 2.14:** Geyser Old Faithful (a) 272 observaciones; (b) Clusters: Algoritmo 2-Neal (2000)

### 2.4.3. Ejemplo 3: muestreo Gibbs por bloques

En la sección 1.3.1 se definió el DP, y en la 1.3.3 se describió el proceso *stick-breaking* como una forma de representar un DP. En estas secciones, las Ec. (2.2), (2.3) y (2.5) que definen al DP están dadas por una suma infinita de términos; en cambio, en el muestreo Gibbs por bloques (BGS) la a priori DP para el modelo (2.9) se asume de dimensión finita, lo que se traduce en un número finito de componentes de mezcla en la distribución de mezclas.

El DP de dimensión finita ( $DP_N$ ) está dado por la Ec. (2.5) con el límite de la suma igual a  $N$ , y donde  $\beta_N = 1$  garantiza que  $\sum_{k=1}^N \pi_k = 1$  con probabilidad 1. La dimensión finita de la a priori DP es la clave del BGS porque permite expresar al modelo completamente en términos de un número finito de variables aleatorias. Con una a priori de dimensión finita  $G \sim DP_N(\alpha, G_0)$ , el modelo (2.9) puede ser reescrito como (Ishwaran & James, 2001):

$$\begin{aligned}
 (y_i | \mathbf{Z}, \mathbf{K}) &\sim \pi(y_i | Z_{K_i}) \\
 K_i | \mathbf{p} &\sim \sum_{k=1}^N p_k \delta_k(\cdot) \\
 (\mathbf{p}, \mathbf{Z}) &\sim \pi(\mathbf{p}) \times G_0^N(\mathbf{Z})
 \end{aligned} \tag{2.21}$$

donde  $\mathbf{K} = (K_1, \dots, K_n)$ ,  $\mathbf{Z} = (Z_1, \dots, Z_N)$ ,  $\mathbf{p} = (p_1, \dots, p_N) \sim \text{Dir}(\alpha G_0)$ , y  $Z_k$  son iid  $G$ . En el modelo (2.21)  $Z_{K_i} = \theta_i$ , donde las  $K_i$  actúan como variables de clasificación para identificar a las  $Z_k$  asociadas con cada  $\theta_i$ .

Reescribiendo el modelo en la forma (2.21) permite al BGS muestrear de la distribución aposterior  $(\cdot | \mathbf{y})$  directamente. El método consiste en tomar valores iterativamente de las distribuciones condicionales de las variables en bloques:

$$(\mathbf{Z} | \mathbf{K}, \mathbf{y})$$

$$(\mathbf{K} | \mathbf{Z}, \mathbf{p}, \mathbf{y})$$

$$(\mathbf{p} | \mathbf{K})$$

lo que produce muestras de la distribución  $(\mathbf{Z}, \mathbf{K}, \mathbf{p} | \mathbf{y})$ . Cada muestra  $(\mathbf{Z}, \mathbf{K}, \mathbf{p})$  define una medida de probabilidad aleatoria

$$G(\cdot) = \sum_{k=1}^N p_k \delta_{Z_k}(\cdot)$$

la cual, iterativamente, produce muestras de la a posteriori.



## 2.4. Proceso Dirichlet para modelos de mezclas

Sea  $\{K_1^*, \dots, K_m^*\}$  el conjunto de  $m$  valores de  $\mathbf{K}$  distintos.

1. Condicional para  $\mathbf{Z}$ : Simular  $Z_k$  de la distribución a priori  $G_0$ .
2. Tomar  $Z_{K_j^*} | \mathbf{K}, \mathbf{y}$  de la distribución:

$$[Z_{K_j^*} | \mathbf{K}, \mathbf{y}] \propto G_0(Z_{K_j^*}) \prod_{i:K_i=K_j^*} [y_i | Z_{K_j^*}]$$

3. Condicional para  $\mathbf{K}$ : tomar valores de

$$K_i | \mathbf{Z}, \mathbf{p}, \mathbf{y} \sim \sum_{k=1}^N p_{k,i} \delta_k(\cdot), \quad i = 1, \dots, n$$

donde

$$(p_{1,i}, \dots, p_{N,i}) \propto (p_1 f(y_i | Z_1), \dots, p_N f(y_i | Z_N))$$

4. Condicional para  $\mathbf{p}$ :

$$\begin{aligned} p_1 &= V_1^* \\ p_k &= (1 - V_1^*)(1 - V_2^*) \cdots (1 - V_{k-1}^*) V_k^*, \quad k = 2, \dots, N - 1 \end{aligned}$$

donde

$$V_k^* \sim \text{Beta}(a_k + M_k, b_k + \sum_{l=k+1}^N M_l), \quad k = 1, \dots, N - 1$$

$M_k$ : número de  $K_i = k$ .

Para ilustrar el método, considere el problema de regresión tal que las observaciones  $y_i$  satisfacen:

$$y_i = g(x_i | \boldsymbol{\beta}_i) + \epsilon_i \tag{2.22}$$

La función  $g(\cdot | \boldsymbol{\beta}_i)$  es de alguna forma especificada, y los errores  $\epsilon_i$  pueden o no estar correlacionados, con media cero. Note que el vector de parámetros  $\boldsymbol{\beta}$  es específico para  $i$ , es decir, cada observación  $y_i$  es modelada por su propia distribución, definida por  $\boldsymbol{\beta}_i$ ; sin embargo, se asume que los  $\boldsymbol{\beta}_i$  tienen una forma funcional común.

## 2.4. Proceso Dirichlet para modelos de mezclas

Sea entonces  $y_1, \dots, y_n$  una muestra de  $n$  observaciones  $y_i$ , modelada como función de  $x_i$  a través de la siguiente relación:

$$y_i = \beta_{0i} + \beta_{1i}x_i + \epsilon_i, \quad \epsilon_i \sim N(0, 1/\tau_i) \quad (2.23)$$

donde los  $\epsilon_i$  son independientes, las  $x_i$  son conocidas y  $\tau_i$  es la precisión. En términos del modelo dado en (2.21),  $Z_{K_i} = \{\beta_i, \tau_i\}$ .

En este contexto, una forma conveniente de la media a priori es la normal-gamma:

$$\begin{aligned} \tau_i &\sim \text{Ga}(v/2, 2/v) \\ \beta_i | \tau_i &\sim N(\beta_0, 1/\tau_i B) \end{aligned} \quad (2.24)$$

donde  $v/2$  y  $2/v$  son parámetros de forma y escala, respectivamente. Para  $B^{-1}$  es común asumir una distribución *Wishart*( $u, V$ ).

Inicialmente se fijaron 4 grupos con igual probabilidad de pertenencia de las  $y_i$  en cada uno,  $i = 1, \dots, 200$ . Los valores  $\{\beta_j, \tau_j\}$  asociados a cada grupo se muestran en la Tabla 2.9. Los valores  $x$  se tomaron de la siguiente manera:  $x_1 \sim N(0, 1)$  y  $x_i = 0.7x_{i-1} + N(0, 1)$ ,  $i = 2, \dots, n$ , y las  $y_i$  se obtuvieron de una distribución normal de acuerdo con la Ec. (2.23). En la Tabla 2.9,  $w$  indica la proporción efectiva de observaciones.

**Tabla 2.9:** Parámetros de datos simulados para el Algoritmo BGS

Grupo	$\beta_{0j}$	$\beta_{1j}$	$\tau_j$	$w$
1	-1.00	-10.0	1.00	0.245
2	0.00	-5.00	1.33	0.260
3	1.00	-1.00	1.67	0.245
4	2.00	3.00	2.00	0.250

La Figura 2.15-(a) muestra la serie de observaciones simuladas; en la Figura 2.15-(b) las observaciones dentro de cada grupo se distinguen por color.

Dada la a priori (2.24), la distribución de muestro para el paso 2 del algoritmo es también Normal-Gamma:

$$\begin{aligned} \tau_i &\sim \text{Ga}(v_1/2, 2/v_2) \\ \beta_i | \tau_i &\sim N(\beta_0^*, 1/\tau_i B^*) \end{aligned}$$

## 2.4. Proceso Dirichlet para modelos de mezclas

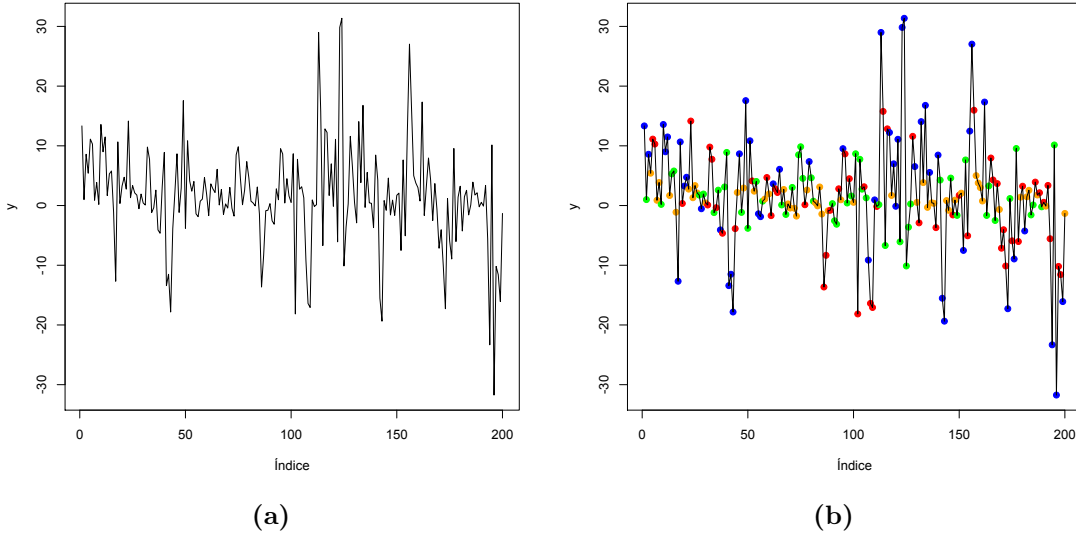


Figura 2.15: Datos simulados para el Algoritmo BGS

donde:

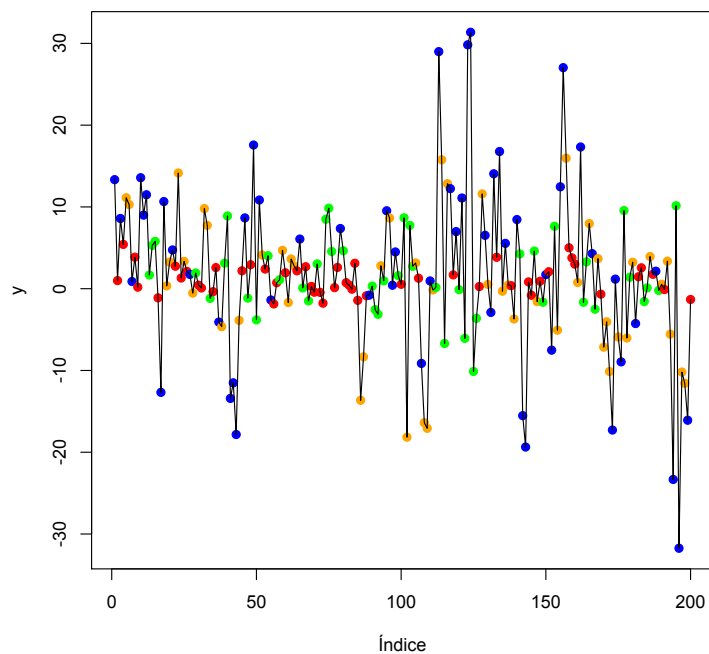
$$\begin{aligned}
 v_1 &= n_j + v, & n_j &= |\{i : K_i = K_j^*\}| \\
 v_2 &= v + (\mathbf{y}'\mathbf{y} + \boldsymbol{\beta}_0' B^{-1} \boldsymbol{\beta}_0) - (\mathbf{X}'\mathbf{y} + B^{-1} \boldsymbol{\beta}_0)' (B^{-1} + \mathbf{X}'\mathbf{X})^{-1} (\mathbf{X}'\mathbf{y} + B^{-1} \boldsymbol{\beta}_0) \\
 B^* &= (B^{-1} + \mathbf{X}'\mathbf{X})^{-1} \\
 \boldsymbol{\beta}_0^* &= B^* (\mathbf{X}'\mathbf{y} + B^{-1} \boldsymbol{\beta}_0)
 \end{aligned}$$

El vector  $\mathbf{y}$  y la matriz  $\mathbf{X}$  contienen sólo los elementos  $i$  tales que  $K_i = K_j^*$  (ver Apéndice A). El algoritmo se ejecutó 5000 veces, tomando como valores iniciales para los hiperparámetros:  $u = 2$  y  $V = u * \text{diag}(10)$  para los grados de libertad y la matriz de escala, respectivamente, de la distribución de  $B$ ;  $\boldsymbol{\beta}_0 = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$  para la media de la distribución condicional a priori de  $\boldsymbol{\beta}$ ;  $v = 2$  para los parámetros de forma y escala de la a priori para  $\tau$ ;  $\alpha = 1$  y  $N = 10$ , el parámetro de concentración y el límite de truncamiento, respectivamente, del DP. En el algoritmo se incluyó un paso de actualización para la matriz de escala  $V$  de su distribución condicional; para  $\alpha$ , siguiendo la propuesta de West (1992) y Escobar & West (1995); y para  $v$ , usando Metropolis-Hastings en el  $\log(v)$ . Los valores propuestos  $v^*$  son tomados de  $\exp(\eta)$ , donde  $\eta \sim N(\log(v), 0.5)$ .

La Figura 2.16 muestra la agrupación obtenida usando BGS. La Tabla 2.10 contiene las estimaciones para  $\boldsymbol{\beta}_j$ , el número de observaciones clasificadas en cada grupo, y el número de observaciones que se clasificaron incorrectamente en cada

## 2.4. Proceso Dirichlet para modelos de mezclas

grupo. Por ejemplo, el grupo 1 tiene 54 observaciones, pero 10 de ellas pertenecen en realidad a otro grupo. Todos los valores de la tabla son tomados de la última iteración.



**Figura 2.16:** Agrupación con el BGS

**Tabla 2.10:** Parámetros estimados con el Algoritmo BGS

Grupo	$\beta_{0j}$	$\beta_{1j}$	$n_j$	errores
1	-1.37	-9.97	54	10
2	0.14	-5.00	50	7
3	0.53	-1.06	51	10
4	1.70	3.02	45	5

Si bien el resultado de este método tuvo errores en la clasificación, la eficiencia no se puede comparar directamente con los algoritmos de los ejemplos anteriores, ya que en este ejemplo las distribuciones usadas para la simulación tienen evidentes traslapes entre grupos, por lo que el proceso de agrupamiento se dificulta.

## 2.5. Modelo de mezclas proceso Dirichlet jerárquico

Suponga que se tienen observaciones que están organizadas en *grupos*, y asuma que son intercambiables dentro de cada grupo y entre grupos. De manera específica, considere lo siguiente:

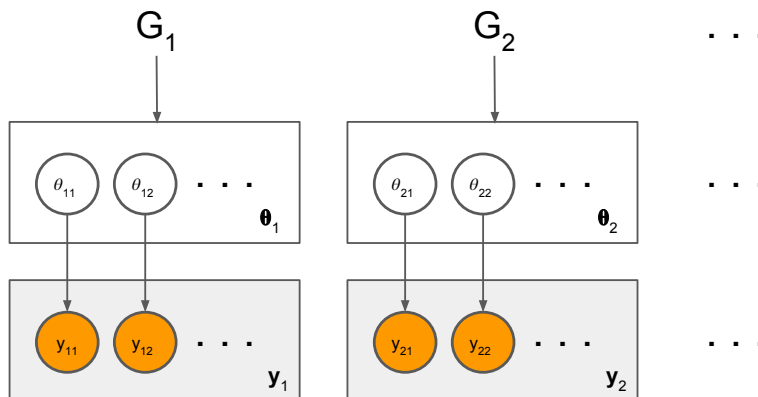
1.  $i$ : índice de las observaciones dentro de cada grupo.
2.  $j$ : índice de los grupos.
3. Se asume que las observaciones  $y_{j1}, y_{j2}, \dots$  son intercambiables dentro del grupo  $j$ .
4. Se asume que las observaciones son intercambiables a nivel de grupo, es decir, si  $\mathbf{y}_j = (y_{j1}, y_{j2}, \dots)$  denota todas las observaciones dentro del grupo  $j$ , entonces  $\mathbf{y}_1, \mathbf{y}_2, \dots$  son intercambiables.
5. Cada observación se toma independiente de un modelo de mezcla, es decir, hay un componente de mezcla asociado con cada observación.
6.  $\theta_{ji}$  es el parámetro que especifica el componente de mezcla asociado con la observación  $y_{ji}$  (las variables  $\theta_{ji}$  son llamadas *factores*).
7.  $F(\theta_{ji})$  es la distribución de  $y_{ji}$  dado el factor  $\theta_{ji}$ .
8.  $G_j$  es la distribución a priori para los factores  $\boldsymbol{\theta}_j = (\theta_{j1}, \theta_{j2}, \dots)$  asociados con el grupo  $j$ .
9. Se asume que los factores son condicionalmente independientes dado  $G_j$ , y que en cada grupo  $j$  la observación  $i$ ,  $y_{ji}$ , es condicionalmente independiente de las otras observaciones dado el factor  $\theta_{ij}$  (ver Figura 2.17).

Bajo los supuestos anteriores, el modelo de probabilidad descrito tiene la siguiente representación:

$$\begin{aligned} \theta_{ji}|G_j &\sim G_j \\ y_{ji}|\theta_{ji} &\sim F(\theta_{ji}) \end{aligned} \tag{2.25}$$

Note que si  $G_j$  se distribuye como DP, el modelo (2.25) es el DPMM (2.9) de la sección anterior para cada  $j$ . En el DPMM, el objetivo fue encontrar el número

## 2.5. Modelo de mezclas proceso Dirichlet jerárquico



**Figura 2.17:** Representación del HDPMM

de componentes en la mezcla y los parámetros que definen a cada uno de ellos. Sin embargo, ahora las observaciones están agrupadas, cada grupo  $j$  está asociado con un modelo de mezcla, y lo que se desea es *relacionar* esos modelos de mezcla. Bajo esta configuración, una a priori no paramétrica apropiada es el DP jerárquico (HDP), y el modelo se denomina modelo de mezclas proceso Dirichlet jerárquico (HDPMM).

Dicho de otro modo, una mezcla HDP se basa en múltiples mezclas DP, una para cada grupo. El objetivo de un HDP no sólo es conocer los clusters de cada grupo de manera individual, sino que el interés es la relación entre los clusters de diferentes grupos, y lo que se desea es conocer cómo los clusters se comparten entre varias mezclas DP.

En los modelos de mezclas los clusters se representan por sus parámetros, y los parámetros son seleccionados de una distribución base  $G_0$ , tal que, para permitir cualquier posible valor para los parámetros, generalmente se asume que  $G_0$  es una distribución continua. La idea de compartir clusters es equivalente a compartir parámetros, por lo que el supuesto de que  $G_0$  sea continua se contrapone con el objetivo de compartir clusters entre grupos. Es decir, aun si cada grupo comparte la misma distribución base  $G_0$ , siendo  $G_0$  una distribución continua, generará, con probabilidad uno, distintos valores de los parámetros. En cambio, si se induce a que  $G_0$  sea por sí misma un DP, se da oportunidad de compartir clusters entre los grupos.

### 2.5.1. Proceso Dirichlet jerárquico

El **HDP** es introducido por [Teh et al. \(2005\)](#) como una propuesta para tratar el problema que involucra datos agrupados, en donde cada observación dentro de un grupo se toma de un modelo de mezclas, y es deseable permitir que los grupos compartan clusters, es decir, componentes de mezclas. Aunque el término también es usado por [Beal et al. \(2002\)](#) y [Müller et al. \(2004\)](#), la representación dada aquí es la de [Teh et al. \(2005\)](#) y [Teh et al. \(2006\)](#).

Un **HDP** define un conjunto de medidas de probabilidad aleatorias  $G_j$ , una para cada grupo, y una medida de probabilidad aleatoria global  $G_0$ , tal que

$$G_0 | \gamma, H \sim \text{DP}(\gamma, H)$$

donde  $\gamma$  es el parámetro de concentración y  $H$  la medida de probabilidad base. Las medidas aleatorias  $G_j$  son condicionalmente independientes dado  $G_0$ , con distribuciones dadas por un **DP** con medida base  $G_0$ ,

$$G_j | \alpha, G_0 \sim \text{DP}(\alpha, G_0),$$

Los hiperparámetros del **HDP** consisten de la medida de probabilidad base  $H$  y los parámetros de concentración  $\alpha$  y  $\gamma$ . La distribución  $G_0$  varía alrededor de la a priori  $H$  de acuerdo con  $\gamma$ . La distribución  $G_j$  en el  $j$ -ésimo grupo se desvía de  $G_0$  de acuerdo con  $\alpha$ . Si se espera distinta variabilidad en diferentes grupos, entonces se puede usar un parámetro de concentración  $\alpha_j$  para cada grupo  $j$ .

Un **HDP** puede ser usado como la distribución a priori sobre los factores para datos agrupados, de manera que la Ec. (2.25) completa la definición de un HDPMM. El **HDP** puede ser extendido a más de dos niveles. Esto es, la medida base  $H$  puede ser tomada de un **DP**, y la jerarquía puede ser extendida para tantos niveles como sea útil.

De manera análoga a un **DP**, se encuentran en la literatura varias representaciones de un **HDP**. [Teh et al. \(2006\)](#) discuten el **HDP** en términos de un proceso *stick-breaking*, de una generalización del proceso de restaurant Chino, que refieren como franquicia de restaurant Chino, y de una aproximación truncada del **HDP**. Otra representación del **HDP** asociada al CRF puede consultarse en [Fox et al. \(2011b\)](#).

## 2.6. Conclusiones

- En este capítulo se ha introducido el proceso Dirichlet como una distribución a priori para los parámetros de los componentes de un modelo de mezclas. La distribución resulta conveniente porque está definida sobre un espacio muestral infinito, y la distribución a posteriori es analíticamente manejable. Estas propiedades son deseables en modelos no paramétricos.
- Se describieron e ilustraron tres métodos para muestrear de la distribución a posteriori de un modelo de mezclas proceso Dirichlet. Los métodos permiten a los datos determinar simultáneamente el número de componentes de mezcla (distribuciones) de los que fueron generados, y los parámetros que definen a cada distribución.
- El proceso Dirichlet puede ser extendido a dos o más niveles. El proceso resultante se denomina proceso Dirichlet jerárquico ([HDP](#)). En un contexto de modelos de mezclas se usa el [HDP](#) como distribución a priori, resultando en un modelo que se conoce como proceso Dirichlet jerárquico para modelos de mezclas. El objetivo de estos modelos es compartir clusters entre varias mezclas [DP](#).



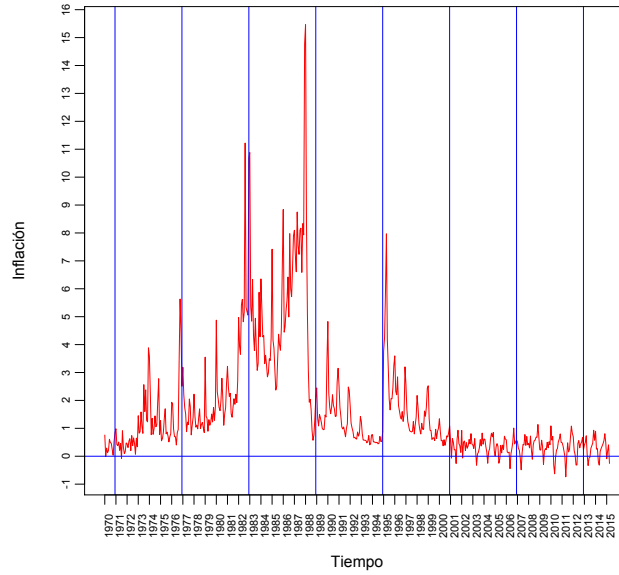
# Capítulo 3

## Modelos lineales dinámicos

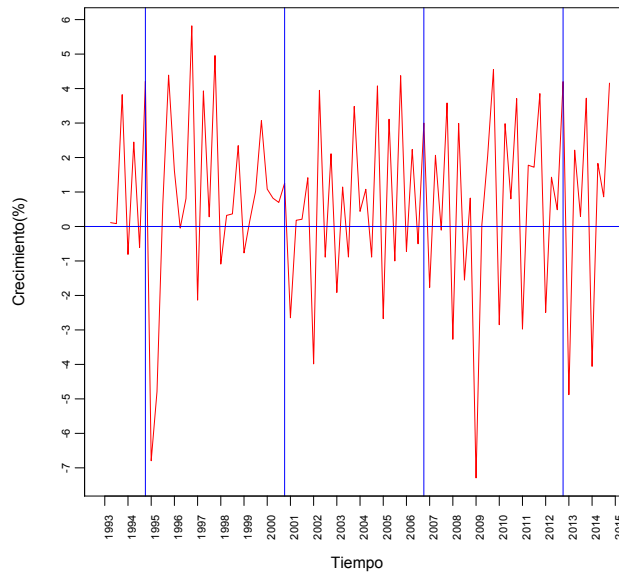
### 3.1. Introducción

El comportamiento irregular en el tiempo es una característica común de muchas variables macroeconómicas. La Figura 3.1 muestra las series mensuales de la inflación y el crecimiento trimestral del producto interno bruto (PIB) en México. Las líneas verticales separan periodos de tiempo correspondientes a los sexenios presidenciales. Hasta antes de 2000, se aprecia fácilmente la irregular variabilidad de la inflación. Después de ese año la serie fluctúa con cierta regularidad, debido a que el objetivo de política macroeconómica de los años recientes es mantener bajos niveles de inflación. También se distingue cierto componente estacional, con aumentos en el primero y último periodo del año. La serie del PIB tiene cambios de nivel abruptos durante todo el periodo; las disminuciones violentas son claras a finales de los sexenios, así como en periodos de crisis (1994-1995, 2008-2009).

### 3.1. Introducción



(a)



(b)

**Figura 3.1:** Series macroeconómicas en México. (a): inflación mensual, (b): crecimiento trimestral del producto interno bruto.

### 3.1. Introducción

---

Si se desea estudiar el comportamiento de series de tiempo no estacionarias, como las de la Figura (3.1), la metodología tradicional de [Box et al. \(2015\)](#) generalmente requiere una transformación preliminar de los datos para obtener una serie estacionaria. Una series es estacionaria cuando su media y varianza no cambian con el tiempo, y no sigue ninguna tendencia. La transformación de los datos puede ocasionar, por un lado, pérdida de información sobre el proceso que origina la serie, y por otro lado, puede dificultar la interpretación de los resultados.

Por otra parte, en muchos problemas de la estadística práctica los modelos de regresión juegan un papel fundamental. Con frecuencia, el objetivo de tales modelos se centra en proporcionar una descripción cuantitativa de las relaciones entre cantidades observables, tales como entre dos o más series de tiempo. Por ejemplo, es común el interés en la medida en que los cambios en la media de la variable respuesta son explicados a través del o los regresores. Para muchos propósitos una estimación global de esa medida (coeficientes de regresión) podría ser satisfactoria. Sin embargo, en el contexto de series de tiempo es poco probable que una estimación global describa adecuadamente la relación entre las variables conforme el tiempo evoluciona. Los modelos lineales dinámicos proporcionan flexibilidad al modelo al permitir la posibilidad de que los coeficientes de regresión varíen en el tiempo ([West & Harrison, 1997](#)).

En este capítulo se presentan las nociones básicas de los modelos de espacio-estado y su uso en el análisis de series de tiempo. Se hace énfasis en los dos tipos más importantes: (1) los modelos de Markov ocultos, cuya dinámica se describe en términos de transiciones de una variable aleatoria discreta; (2) los sistemas dinámicos lineales, en los que la variable aleatoria que describe la dinámica es Gaussiana. El principal objetivo es proveer del marco general para el estudio de sistemas dinámicos más complejos desarrollados en los capítulos 4 y 5; aunque la cobertura no es exhaustiva, se sugieren referencias que pueden ser consultadas si se desea ampliar la información.

Los modelos de espacio-estado asumen que las series de tiempo observadas fueron generadas a partir de una secuencia de variables ocultas, o no observadas, que evolucionan en el tiempo mediante una cadena de Markov, definida en la primera sección del capítulo. Cuando las variables ocultas son discretas, el modelo se conoce como [HMM](#); cuando las variables ocultas son Gaussianas, se trata de un [LDS](#). Uno de los principales problemas de interés a resolver en estos dos tipos de modelos de espacio-estado es encontrar la distribución marginal a posteriori del *estado* al tiempo  $t$ , dada una secuencia de observaciones  $y_{1:k}$ . Si las observaciones disponibles son tales que  $k = t$ , el procedimiento de inferencia se conoce como *filtering*; si se dispone de información adicional, de manera que  $k > t$ , el procedimiento de inferencia se conoce como *suavisamiento*. *Filtering* y *suavisamiento* se

## 3.2. Cadenas de Markov

---

llevan a cabo usando un algoritmo iterativo de dos pasos conocido como *forward-backward* que se describe en la sección 3.2.1 para un [HMM](#) y en la sección 3.2.2 para un [LDS](#). Extensiones de los [LDS](#) que incluyen parámetros desconocidos, como las matrices de covarianzas, se describen en la sección 3.2.3. Finalmente, la sección 3.3 resume las metodologías que se exponen en el capítulo.

## 3.2. Cadenas de Markov

Un modelo de Markov es un modelo estocástico que relaja el supuesto de observaciones secuenciales i.i.d., y expresa el efecto que ejerce un estado actual sobre un estado futuro. El modelo de Markov más simple es la cadena de Markov de primer orden; esta asume que una observación en el tiempo  $t$  sólo depende de la observación en  $t - 1$ , y no de lo ocurrido previamente. Esta característica de pérdida de memoria se conoce como propiedad de Markov.

De manera formal, una cadena de Markov se define sobre una secuencia de variables  $z_{1:T}$  discretas o continuas bajo el siguiente supuesto de independencia condicional ([Barber, 2012](#)):

$$p(z_t | z_1, \dots, z_{t-1}) = p(z_t | z_{t-L}, \dots, z_{t-1})$$

donde  $L \geq 1$  es el orden de la cadena de Markov y  $z_t = \emptyset$  para  $t < 1$ . Al conjunto de valores posibles de  $z_i$  se le llama espacio-estado de la cadena. Para una cadena de Markov de primer orden ( $L = 1$ ),

$$p(z_{1:T}) = p(z_1)p(z_2|z_1)p(z_3|z_2) \cdots p(z_T|z_{T-1}) \quad (3.1)$$

De la Ec. (3.1) y la regla del producto de probabilidad es fácil verificar que la distribución condicional para la observación  $z_t$ , dadas todas las observaciones previas al tiempo  $t$ , está dada por

$$p(z_t | z_1, \dots, z_{t-1}) = p(z_t | z_{t-1})$$

En muchas aplicaciones, las distribuciones condicionales  $p(z_t | z_{t-1})$  se restringen a ser iguales, es decir, la cadena se supone estacionaria. Dicho de otro modo, una cadena de Markov estacionaria, conocida también como cadena de Markov tiempo-homogénea, es tal que las transiciones  $p(z_t = s' | z_{t-1} = s) = f(s', s)$  son independientes del tiempo. Por ejemplo, si las distribuciones condicionales dependen de parámetros ajustables (cuyos valores pueden ser inferidos de un conjunto de datos de entrenamiento), entonces todas las distribuciones condicionales en la cadena compartirán los mismos valores de esos parámetros ([Bishop, 2006](#));

### 3.3. Modelos de espacio-estado

---

en cambio, en una cadena de Markov no estacionaria, el tiempo juega un papel central en la determinación de las distribuciones conjuntas, de tal manera que  $p(z_t = s' | z_{t-1} = s) = f(s', s, t)$ . Otras propiedades de las cadenas de Markov se pueden consultar ampliamente en [Berger \(1993\)](#); [Feldman & Valdez-Flores \(2010\)](#); [Schinazi \(2014\)](#).

### 3.3. Modelos de espacio-estado

El término *estado* es el concepto fundamental para describir a un sistema dinámico (lineal o no lineal). Intuitivamente, un estado se refiere a alguna información cuantitativa (un conjunto de números, una función, etc.) que es la menor cantidad de datos de la que se debe tener conocimiento sobre el comportamiento pasado del sistema para predecir su comportamiento futuro ([Kalman, 1960](#)).

Los modelos de espacio-estado son modelos que usan variables estado latentes para describir un sistema mediante un conjunto de ecuaciones diferenciales de primer orden. Las variables estado se reconstruyen a partir de los datos *input-output* medidos, pero no son observadas durante un experimento. La dinámica se describe en términos de transiciones de estado, es decir, cómo un estado se transforma en otro conforme el tiempo pasa.

La representación más general de un sistema dinámico lineal en la forma espacio-estado con  $p$  *inputs*,  $q$  *outputs*, y  $n$  variables de estado se escribe de la siguiente manera:

$$\dot{\mathbf{z}}(t) = A(t)\mathbf{z}(t) + B(t)\mathbf{u}(t) \quad (3.2)$$

$$\mathbf{y}(t) = C(t)\mathbf{z}(t) + D(t)\mathbf{u}(t) \quad (3.3)$$

donde:

$\mathbf{z}(\cdot)$  es el vector de estados,  $\mathbf{z}(t) \in \mathbb{R}^n$ ;

$\mathbf{y}(\cdot)$  es el vector de *output*,  $\mathbf{y}(t) \in \mathbb{R}^q$ ;

$\mathbf{u}(\cdot)$  es el vector de *input*, o vector control,  $\mathbf{u}(t) \in \mathbb{R}^p$ ,  $p \leq n$ ;

$A(\cdot)$  es la matriz de estados, de dimensión  $n \times n$ ;

$B(\cdot)$  es la matriz de *input*, de dimensión  $n \times p$ ;

$C(\cdot)$  es la matriz de *output*, de dimensión  $q \times n$ ,  $q \leq n$ ;

### 3.3. Modelos de espacio-estado

---

$D(\cdot)$  es la matriz *feedthrough*, de dimensión  $q \times p$ ;

$$\dot{\mathbf{z}}(t) := \frac{d}{dt}\mathbf{z}(t).$$

En las Ec. (3.2) y (3.3) todas las matrices dependen del tiempo; sin embargo, si todos los coeficientes de  $A(t)$ ,  $B(t)$ ,  $C(t)$ , y  $D(t)$  son constantes, se dice que el sistema dinámico es invariante en el tiempo o estacionario (Kalman, 1960). Adicionalmente, en casos donde el sistema no establece una conexión instantánea entre *input* y *output*, es decir, cuando el *input* al tiempo  $t$  no afecta el *output* al tiempo  $t$ ,  $D(t)$  es una matriz de ceros y el modelo se denomina puramente dinámico (Galar Pascual, 2015).

Los modelos de espacio-estado proveen el marco general para analizar sistemas dinámicos que se miden u observan a través de un proceso estocástico, y tienen muchas aplicaciones en áreas como ingeniería (Friedland, 1986), computación (Hangos et al., 2001) y economía (Zeng & Wu, 2013). Los dos ejemplos más importantes de los modelos de espacio-estado son: modelos de Markov ocultos (HMM), en los que las variables latentes son discretas, y sistemas dinámicos lineales (LDS), en los que las variables latentes son Gaussianas.

#### 3.3.1. Modelos de Markov ocultos

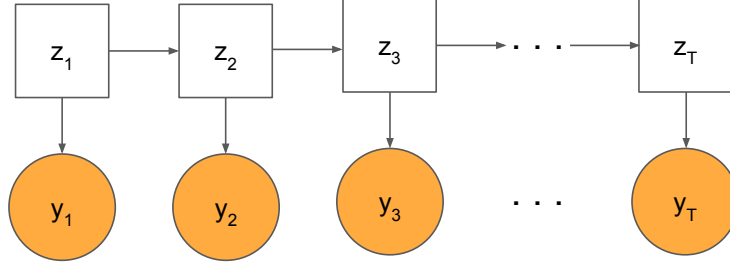
Un HMM es una extensión de un modelo de mezclas (ver por ejemplo Ec. 2.6) en el que la elección del componente de mezcla para cada observación no se selecciona independiente, sino que depende de la elección del componente para la observación previa (Bishop, 2006). Considere como ejemplo la Figura 3.2. Cada forma geométrica representa una variable aleatoria que puede adoptar cualquier valor. Los nodos cuadrados denotan una variable aleatoria discreta ( $z$ ), y los círculos una variable aleatoria continua ( $y$ ). El color enfatiza que la variable es observada<sup>1</sup>. Las flechas denotan dependencia condicional. La variable aleatoria  $z_t$  (variable latente) es el estado oculto al tiempo  $t$ . La variable aleatoria  $y_t$  es la observación al tiempo  $t$ . La figura ilustra que en un HMM la distribución de probabilidad condicional de la variable oculta  $z_t$ , dados los valores de  $z$  en todo  $t$ , depende sólo del valor de la variable en el tiempo previo,  $z_{t-1}$ , y no del valor en  $t - 2$  ni de los anteriores (propiedad de Markov). Mientras que el valor de la variable observada  $y_t$  sólo depende del valor de la variable oculta  $z_t$ . Es decir, la distribución de cada observación en un HMM es una densidad de la mezcla; las variables ocultas, que controlan el componente de la mezcla que se selecciona para cada observación ( $p(y|z)$ ), están

---

<sup>1</sup>La variable observada puede ser también discreta.

### 3.3. Modelos de espacio-estado

relacionadas a través de un proceso de Markov en lugar de ser independientes ( $p(z_t|z_{t-1})$ ).



**Figura 3.2:** HMM de T pasos en el tiempo.

Note que en los modelos de Markov más simples el estado es directamente visible para el observador. En un HMM el estado no es visible directamente, pero el resultado, que depende del estado, es visible. El HMM clásico, ilustrado en la Figura 3.2, es formalizado de la siguiente manera:

$$z_t|z_{t-1} \sim \pi_{z_{t-1}} \quad (3.4)$$

$$y_t|z_t \sim F(\theta_{z_t}) \quad (3.5)$$

donde  $F(\cdot)$  es una familia de distribuciones de emisión indexada,  $\theta_i$  son los parámetros de emisión para el estado  $i$ ,  $z_t$  denota el estado de la cadena de Markov al tiempo  $t$ ,  $t = 1, \dots, T$ , y  $\pi_j$  especifica la distribución de transición para el estado  $j$ . En el contexto Bayesiano, (3.4) define la distribución a priori del proceso  $\{z_{1:T}\}$  y (3.5) define la función de verosimilitud. Dado el estado  $z_t$ , la observación  $y_t$  es condicionalmente independiente de las observaciones y estados a otros tiempos, por lo que la distribución conjunta de las variables latentes y las variables observadas está dada por:

$$p(y_{1:T}, z_{1:T}) = p(z_1)p(y_1|z_1) \prod_{t=2}^T p(y_t|z_t)p(z_t|z_{t-1}) \quad (3.6)$$

$$\begin{aligned} &= p(z_1)p(z_2|z_1) \cdots p(z_T|z_{T-1})p(y_1|z_1)p(y_2|z_2) \cdots p(y_T|z_T) \\ &= p(z_{1:T})p(y_{1:T}|z_{1:T}) \end{aligned} \quad (3.7)$$

donde:  $y_{1:T} = \{y_1, \dots, y_T\}$ ,  $z_{1:T} = \{z_1, \dots, z_T\}$  y  $p(z_1)$  es la distribución a priori de la variable latente inicial. La verosimilitud marginal se obtiene de:

### 3.3. Modelos de espacio-estado

---

$$p(y_{1:T}) = \sum_{z_{1:T}} p(y_{1:T}, z_{1:T})$$

Aunque se ha señalado que las variables latentes son discretas, en estadística es común que el término **HMM** se refiera a cualquier modelo con la estructura de independencia dada en la Ec. (3.6) (Barber, 2012), sin embargo, en este trabajo los modelos en los que la variable latente es continua son referidos como **LDS**.

Una sencilla ilustración de un **HMM** se presenta en el siguiente ejemplo. En la práctica, el uso de los **HMM** es bastante amplio, y se pueden encontrar diversas aplicaciones en campos como: reconocimiento de voz (Rabiner, 1989; Juang & Rabiner, 1991), análisis de secuencias biológicas (Krogh et al., 1994, 2001) y en finanzas (Bhar & Hamori, 2004; Mamon & Elliott, 2014).

Ejemplo: suponga que Paco es un estudiante de doctorado y Sara, su mamá, una viuda jubilada. Ellos viven en países distintos. Paco llama con frecuencia a Sara, quien platica con su hijo sobre sus tres actividades principales: tomar café con sus amigas, sacar a pasear a su perro y ver en su casa alguna película. La actividad que Sara elija depende sólo del estado del clima (soleado o con lluvia). Con base en lo que Sara le cuenta a Paco, él trata de adivinar el clima de un día en particular. Paco no observa directamente los dos estados (soleado o con lluvia), esto es, los valores que puede asumir la variable oculta  $z_t$  (clima en el tiempo  $t$ ) del modelo (3.4)-(3.5). Cada día hay cierta probabilidad, que depende del estado del clima, de que Sara realice cada una de las tres actividades. Estas probabilidades son conocidas como *probabilidades de emisión*, y determinan las distribuciones  $F(\cdot)$  en (3.5). Las probabilidades de transición  $\pi_j$  representan los cambios en el estado del clima, por ejemplo, la probabilidad de que hoy llueva dado que ayer llovió.

Los problemas clásicos de inferencia en los **HMM** se pueden resumir en encontrar: (1) la distribución marginal a posteriori del estado en un particular momento  $t$  del tiempo, condicional a una secuencia de datos, es decir,  $p(z_t|y_{1:k})$ . Si  $k < t$ , el proceso se conoce como predicción; si  $k = t$ , el proceso se conoce como *filtering* y; si  $k > t$ , el proceso se suele referir como *suavisamiento*. (2) la verosimilitud  $p(y_{1:T})$ ; (3) la trayectoria oculta más probable  $\operatorname{argmax}_{z_{1:T}} p(z_{1:T}|y_{1:T})$ . Los procesos *filtering* y *suavisamiento* se abordan usando un algoritmo iterativo de dos pasos conocido como *forward-backward* (Rabiner, 1989). La verosimilitud  $p(y_{1:T})$  se puede obtener marginalizando la distribución conjunta usada en el proceso *filtering*. La secuencia más probable de estados dada una secuencia de observaciones se encuentra mediante el algoritmo de Viterbi (1967). Las derivaciones del Cap. 4 están basadas en los procesos *filtering* y *suavisamiento*, y en el algoritmo *forward-backward*, por



### 3.3. Modelos de espacio-estado

---

lo que estos procedimientos se describen brevemente en los siguiente apartados. Una revisión más completa se puede consultar, por ejemplo, en [Rabiner \(1989\)](#), [Bishop \(2006\)](#) y [Barber \(2012\)](#). El problema de predicción de los estados no es abordado en esta tesis.

#### Filtering

Dada una secuencia de observaciones y los parámetros del modelo, el objetivo de *filtering* es calcular la distribución de la variable latente al final de la secuencia, es decir, se desea calcular  $p(z_t|y_{1:t})$ . La distribución deseada se obtiene por normalización de la distribución conjunta marginal  $p(z_t, y_{1:t})$ :

$$\begin{aligned}
 p(z_t, y_{1:t}) &= \sum_{z_{t-1}} p(z_t, z_{t-1}, y_{1:t-1}, y_t) \\
 &= \sum_{z_{t-1}} p(y_t|y_{1:t-1}, z_t, z_{t-1})p(z_t|y_{1:t-1}, z_{t-1})p(y_{1:t-1}, z_{t-1}) \\
 &\quad \text{(por el supuesto de independencia condicional del modelo)} \\
 &= \sum_{z_{t-1}} p(y_t|z_t)p(z_t|z_{t-1})p(z_{t-1}, y_{1:t-1})
 \end{aligned}$$

Definiendo  $\alpha(z_t) = p(z_t, y_{1:t})$ :

$$\alpha(z_t) = p(y_t|z_t) \sum_{z_{t-1}} p(z_t|z_{t-1})\alpha(z_{t-1}), \quad t > 1 \quad (3.8)$$

$$\alpha(z_1) = p(z_1, y_1) = p(y_1|z_1)p(z_1) \quad (3.9)$$

Note que para calcular  $p(z_t, y_{1:t})$  directamente es necesario marginalizar sobre todas las posibles secuencias de estados  $\{z_{1:t-1}\}$ . En lugar de eso, se toma ventaja de la independencia condicional implicada en el [HMM](#) para realizar el cálculo recursivamente, de manera que se puede encontrar  $\alpha(z_t)$  fácilmente de  $\alpha(z_{t-1})$ .

Las probabilidades  $p(y_t|z_t)$  y  $p(z_t|z_{t-1})$  de (3.8) y (3.9) están dadas por la distribución de emisión y las probabilidades de transición, respectivamente. El primer término del lado derecho de (3.8) se conoce como corrector, y la parte de la sumatoria como predictor. La distribución condicional de la variable latente  $z_t$  es entonces proporcional a  $\alpha(z_t)$ :

$$\begin{aligned}
 p(z_t|y_{1:t}) &\propto \alpha(z_t) \\
 &\propto \sum_{z_{t-1}} p(y_t|z_t)p(z_t|z_{t-1})p(z_{t-1}|y_{1:t-1}), \quad t > 1
 \end{aligned} \quad (3.10)$$

### 3.3. Modelos de espacio-estado

---

Y la normalización es tal que  $\sum_{z_t} \alpha(z_t) = 1$ . La distribución  $p(z_{t-1}|y_{1:t-1})$  incorpora toda la información previa observada. Su interpretación es que representa una nueva a priori modificada por  $z_t$ , es decir, por la dinámica (de un paso en el tiempo) del modelo. El término  $p(y_t|z_t)p(z_t|z_{t-1})$  se puede interpretar como una varosímilitud; la a priori y la nueva evidencia de datos dan lugar a la posteriori conjunta  $p(z_t, z_{t-1}|y_{1:t})$ . En el siguiente paso, la a posteriori se convierte en la nueva a priori, y así sucesivamente. El problema de *filtering* se puede abordar eficientemente usando un algoritmo recursivo conocido como *forward*, descrito más adelante.

#### Smoothing

El objetivo del *smoothing* es similar al de *filtering*, pero el interés ahora es sobre la distribución de la variable latente en cualquier punto  $t$  de tiempo de la secuencia, menor al tiempo  $u$  final. Es decir, se desea encontrar:  $p(z_t|y_{1:u})$  (*smoothed posterior*), tal que  $t < u$ . El método más común en la literatura de [HMM](#) para encontrar la distribución se conoce como *parallel*. Este consiste en separar la distribución *smoothed posterior* en contribuciones del pasado y el futuro mediante la distribución conjunta marginal ([Barber, 2012](#)):

$$\begin{aligned} p(z_t, y_{1:T}) &= p(z_t, y_{1:t}, y_{t+1:T}) = p(z_t, y_{1:t})p(y_{t+1:T}|z_t, y_{1:t}) \\ &= p(z_t, y_{1:t})p(y_{t+1:T}|z_t) = \alpha(z_t)\beta(z_t) \end{aligned}$$

El término  $\alpha(z_t)$  es la contribución del pasado, y se obtiene por el procedimiento recursivo *forward* anterior. El término  $\beta(z_t)$  es la contribución del futuro, y se puede obtener usando un procedimiento recursivo conocido como *backward*. Note que  $z_t$  separa el pasado del futuro. Los procedimientos *forward* y *backward* recursivos son independientes por lo que se ejecutan en paralelo. Combinando los resultados se obtiene la *smoothed posterior* de la siguiente manera:

$$\begin{aligned} p(y_{t+1:T}|z_t) &= \sum_{z_{t+1}} p(y_{t+1}, y_{t+2:T}, z_{t+1}|z_t) \\ &= \sum_{z_{t+1}} p(y_{t+1}|y_{t+2:T}, z_{t+1}, z_t)p(y_{t+2:T}, z_{t+1}|z_t) \\ &= \sum_{z_{t+1}} p(y_{t+1}|z_{t+1})p(y_{t+2:T}|z_{t+1}, z_t)p(z_{t+1}|z_t) \\ &= \sum_{z_{t+1}} p(y_{t+1}|z_{t+1})p(y_{t+2:T}|z_{t+1})p(z_{t+1}|z_t) \\ \beta(z_t) &= \sum_{z_{t+1}} p(y_{t+1}|z_{t+1})p(z_{t+1}|z_t)\beta(z_{t+1}), \quad 1 \leq t \leq T \end{aligned}$$

### 3.3. Modelos de espacio-estado

---

donde:  $\beta(z_t) \equiv p(y_{t+1:T}|z_t)$  y  $\beta(z_T) = 1$ . Entonces, la *smoothed posterior* está dada por:

$$p(z_t|y_{1:T}) = \frac{\alpha(z_t)\beta(z_t)}{\sum_{z_t} \alpha(z_t)\beta(z_t)}$$

Por otra parte, en algunas aplicaciones es necesario muestrear la trayectoria conjunta  $z_{1:T}$  de la a posteriori  $p(z_{1:T}|y_{1:T})$ . Una manera práctica es usar la siguiente representación condicional (ver [Fruhwirth-Schnatter, 1994](#)):

$$p(z_{1:T}|y_{1:T}) = \prod_{t=0}^{T-1} p(z_t|z_{t+1}, \dots, z_T, y_{1:T})p(z_T|y_{1:T}) \quad (3.11)$$

Para obtener (3.11) se comienza muestreando el estado  $z_T$  de la marginal  $p(z_T|y_{1:T})$ . Entonces, se muestrea hacia atrás en el tiempo de las densidades condicionales  $p(z_t|z_{t+1}, \dots, z_T, y_{1:T})$ ,  $t = T - 1, \dots, 0$ .

#### Algoritmo *forward-backward*

El objetivo de *filtering* es actualizar el conocimiento del sistema conforme el tiempo evoluciona y se adiciona una nueva observación. Es decir, las observaciones se incorporan en la secuencia de datos conforme van ocurriendo, y cada observación se traduce en información de la que se aprende antes de descartarla y considerar la siguiente. Mientras el *filtering* estima la distribución del estado en el tiempo  $t$  de un [HMM](#), con base en las observaciones recibidas hasta el momento  $t$ , el suavizado estima la distribución del estado a un particular tiempo  $t$ , dadas todas las observaciones hasta algún tiempo  $k$  posterior a  $t$ . La trayectoria estimada por suavizado tiende a ser más *suave*, resultado de la información adicional incorporada.

El algoritmo *forward-backward* es un algoritmo para inferencia en los [HMM](#) que calcula las marginales a posteriori de todos los estados ocultos dada una secuencia de observaciones. El algoritmo consiste en tres pasos: (1) calcular la probabilidad *forward* de terminar en cualquier estado, dadas las primeras  $t$  observaciones, esto es,  $p(z_t|y_{1:t})$ ,  $t = 1, \dots, T$ ; (2) calcular la probabilidad *backward* de observar la secuencia de datos restante, dado el estado en cualquier punto inicial  $t$ , esto es,  $p(y_{t+1:T}|z_t)$ ; (3) combinar (1) y (2) para obtener la distribución de los estados en

### 3.3. Modelos de espacio-estado

---

cualquier punto específico del tiempo, dada la secuencia completa de observaciones.

De manera concreta, considere a  $\mathbf{T}$  una matriz de probabilidades de transición,  $\mathbf{E}$  una matriz de probabilidades de emisión y  $\mathbf{O}$  una matriz diagonal de probabilidad de observación, con elementos extraídos de  $\mathbf{E}$ . Asumiendo un vector de probabilidades de estados inicial  $\pi_0$  ( $t = 0$ ), se calcula

$$\mathbf{f}_{0:0} = \pi_0 \mathbf{T} \mathbf{O}_0$$

donde  $\mathbf{O}_t$ , que se conoce como matriz de observación, es una matriz diagonal que contiene las probabilidades del evento observado al tiempo  $t$  dado cada estado. Este calculo se realiza *forward* en tiempo incorporando observaciones adicionales:

$$\mathbf{f}_{0:t} = \mathbf{f}_{0:t-1} \mathbf{T} \mathbf{O}_t$$

$\mathbf{f}_{0:t}$  se interpreta como el vector de probabilidades *forward* no normalizado, con la  $i$ -ésima entrada dada por:  $\mathbf{f}_{0:t}(i) = p(y_{1:t}, z_t = i | \pi)$ . En cada paso, se introduce un factor de escala  $c_t$  tal que el vector probabilidades *forward* sume 1:

$$\hat{\mathbf{f}}_{0:t} = c_t^{-1} \mathbf{f}_{0:t} \mathbf{T} \mathbf{O}_t \tag{3.12}$$

donde  $\hat{\mathbf{f}}_{0:t-1}$  es el vector normalizado del paso previo. Note que el producto de los factores de escala es la probabilidad total de observar los eventos dados (la secuencia observada hasta el momento  $t$ ), con independencia de los estados finales:

$$p(y_{1:t} | \pi) = \prod_{k=1}^t c_k$$

Lo anterior implica que el vector de probabilidades escalado está dado por:

$$\hat{\mathbf{f}}_{0:t}(i) = \frac{\mathbf{f}_{0:t}(i)}{\prod_{k=1}^t c_k} = \frac{p(y_{1:t}, z_t = i | \pi)}{p(y_{1:t} | \pi)} = p(z_t = i | y_{1:t}, \pi)$$

Es decir, la probabilidad escalada es la probabilidad de estar en el estado  $i$  en el tiempo  $t$ .

Para calcular las probabilidades *backward* se asume que se inicia en un particular estado  $i$  ( $z_t = i$ ). El interés es en la probabilidad de observar los eventos futuros de ese estado, es decir:

$$\mathbf{b}_{t:T}(i) = p(y_{t+1:T} | z_t = i)$$

### 3.3. Modelos de espacio-estado

---

Debido a que el estado inicial se asume como dado, se comienza con:

$$\mathbf{b}_{T:T} = [111 \cdots]'$$

y el vector (columna) de probabilidades *backward* se obtiene de:

$$\mathbf{b}_{t-1:T}(i) = \mathbf{TO}_t \mathbf{b}_{t:T} \quad (3.13)$$

Note que cada elemento  $i$  del vector en (3.13) es la probabilidad de la secuencia de eventos futuros, dado el estado inicial  $i$ . Normalizar este vector es equivalente a encontrar la verosímilitud de cada estado inicial, dados los eventos futuros. Sin embargo, es más común escalar el vector usando la misma constante  $c_t$  de la probabilidad *forward*:

$$\hat{\mathbf{b}}_{t-1:T} = c_t^{-1} \mathbf{TO}_t \hat{\mathbf{b}}_{t:T} \quad (3.14)$$

donde  $\hat{\mathbf{b}}_{t:T}$  representa el vector escalado previo. Note que:

$$\hat{\mathbf{b}}_{t:T}(i) = \frac{\mathbf{b}_{t:T}(i)}{\prod_{k=t+1}^T c_k}$$

Y en conjunto, las probabilidades *forward* y *backward* permiten encontrar la probabilidad total de cada estado a un tiempo  $t$ :

$$p(z_t = i | y_{1:T}, \pi) = \frac{p(y_{1:T}, z_t = i | \pi)}{p(y_{1:T} | \pi)} = \frac{\mathbf{f}_{0:t}(i) \cdot \mathbf{b}_{t:T}(i)}{\prod_{k=1}^T c_k} = \hat{\mathbf{f}}_{0:t}(i) \cdot \hat{\mathbf{b}}_{t:T}(i) \quad (3.15)$$

La probabilidad en (3.15) se conoce como probabilidad *smoothed*. El numerador  $\mathbf{f}_{0:t}(i) \cdot \mathbf{b}_{t:T}(i)$  es la probabilidad de observar los eventos dados de manera que pasa por el estado  $i$  al tiempo  $t$ . Al dividirlo por la probabilidad total de la secuencia observada, se extrae la probabilidad de que  $z_t = i$ , es decir, la probabilidad de encontrarse en el estado  $i$  al tiempo  $t$ . Estas probabilidades son útiles para determinar el estado más probable en cualquier  $t$ . Sin embargo, note que la secuencia de estados más probables individualmente no produce la secuencia más probable. Esto se debe a que se calculan las probabilidades para cada  $t$  independientes unas de otras, y no se toma en cuenta las probabilidades de transición entre estados. La secuencia más probable de estados que produce una secuencia de observaciones se puede encontrar usando el algoritmo de [Viterbi \(1967\)](#).

### 3.3. Modelos de espacio-estado

---

#### Ejemplo

Siguiendo con el ejemplo de Paco y Sara, considere las siguientes matrices:

$$\text{matriz de probabilidades de transición } (\mathbf{T}) = \begin{bmatrix} & \text{sol} & \text{lluvia} \\ \text{sol} & 0.8 & 0.2 \\ \text{lluvia} & 0.4 & 0.6 \end{bmatrix}$$

$$\text{matriz de probabilidades de emisión } (\mathbf{E}) = \begin{bmatrix} & (1) & (2) & (3) \\ \text{sol} & 0.4 & 0.4 & 0.2 \\ \text{lluvia} & 0.3 & 0.1 & 0.6 \end{bmatrix}$$

Suponga que Paco tiene la siguiente secuencia de observaciones:  $y = \{\text{ver película, ver película, café, pasear al perro}\}$ . En forma matricial, la probabilidad de observar el evento  $j$  en el tiempo  $t$  está dada por:  $\mathbf{O}_{j(t)} = \text{diag}(\mathbf{E}[, j])$ . Es decir, cada elemento de la diagonal es la probabilidad del evento observado dado cada estado. Por ejemplo, para  $j = 3$  (ver película):

$$\mathbf{O}_{3(t)} = \begin{bmatrix} 0.2 & 0 \\ 0 & 0.6 \end{bmatrix}$$

donde  $t = 1, 2$  de acuerdo con las observaciones de Paco. Para calcular las probabilidades *forward*, suponga que no se tiene información sobre el estado del tiempo inicial ( $t = 0$ ), por lo que se considera la siguiente a priori:

$$\mathbf{f}_{0:0} = (0.5 \quad 0.5)$$

De la Ec. (3.12):

$$\begin{aligned} \hat{\mathbf{f}}_{0:1} &= c_1^{-1} \mathbf{f}_{0:0} \mathbf{T} \mathbf{O}_{3(1)} = c_1^{-1} (0.5 \quad 0.5) \begin{bmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{bmatrix} \begin{bmatrix} 0.2 & 0 \\ 0 & 0.6 \end{bmatrix} \\ &= c_1^{-1} (0.12 \quad 0.24) = (0.333 \quad 0.667) \\ \hat{\mathbf{f}}_{0:2} &= c_2^{-1} \hat{\mathbf{f}}_{0:1} \mathbf{T} \mathbf{O}_{3(2)} = c_2^{-1} (0.333 \quad 0.667) \begin{bmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{bmatrix} \begin{bmatrix} 0.2 & 0 \\ 0 & 0.6 \end{bmatrix} \\ &= c_2^{-1} (0.107 \quad 0.280) = (0.276 \quad 0.724) \\ \hat{\mathbf{f}}_{0:3} &= c_3^{-1} \hat{\mathbf{f}}_{0:2} \mathbf{T} \mathbf{O}_{1(3)} = c_3^{-1} (0.276 \quad 0.724) \begin{bmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{bmatrix} \begin{bmatrix} 0.4 & 0 \\ 0 & 0.3 \end{bmatrix} \\ &= c_3^{-1} (0.2042 \quad 0.1469) = (0.5816 \quad 0.4184) \\ \hat{\mathbf{f}}_{0:4} &= c_4^{-1} \hat{\mathbf{f}}_{0:3} \mathbf{T} \mathbf{O}_{2(4)} = c_4^{-1} (0.5816 \quad 0.4184) \begin{bmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{bmatrix} \begin{bmatrix} 0.4 & 0 \\ 0 & 0.1 \end{bmatrix} \\ &= c_4^{-1} (0.2531 \quad 0.0367) = (0.8734 \quad 0.1266) \end{aligned}$$

### 3.3. Modelos de espacio-estado

Note que  $c_1 = 0.36$ ,  $c_2 = 0.387$ ,  $c_3 = 0.3511$  y  $c_4 = 0.2898$ . Para obtener las probabilidades *backward* se comienza con:

$$\mathbf{b}_{4:4} = \begin{bmatrix} 1.0 \\ 1.0 \end{bmatrix}$$

De la Ec. (3.14):

$$\begin{aligned} \hat{\mathbf{b}}_{3:4} &= c_4^{-1} \mathbf{T} \mathbf{O}_{2(4)} \mathbf{b}_{4:4} = 3.451 \begin{bmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{bmatrix} \begin{bmatrix} 0.4 & 0 \\ 0 & 0.1 \end{bmatrix} \begin{bmatrix} 1.0 \\ 1.0 \end{bmatrix} = \begin{bmatrix} 1.173 \\ 0.759 \end{bmatrix} \\ \hat{\mathbf{b}}_{2:4} &= c_3^{-1} \mathbf{T} \mathbf{O}_{1(3)} \hat{\mathbf{b}}_{3:4} = 2.848 \begin{bmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{bmatrix} \begin{bmatrix} 0.4 & 0 \\ 0 & 0.3 \end{bmatrix} \begin{bmatrix} 1.173 \\ 0.759 \end{bmatrix} = \begin{bmatrix} 1.199 \\ 0.924 \end{bmatrix} \\ \hat{\mathbf{b}}_{1:4} &= c_2^{-1} \mathbf{T} \mathbf{O}_{3(2)} \hat{\mathbf{b}}_{2:4} = 2.584 \begin{bmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{bmatrix} \begin{bmatrix} 0.2 & 0 \\ 0 & 0.6 \end{bmatrix} \begin{bmatrix} 1.199 \\ 0.924 \end{bmatrix} = \begin{bmatrix} 0.782 \\ 1.107 \end{bmatrix} \\ \hat{\mathbf{b}}_{0:4} &= c_1^{-1} \mathbf{T} \mathbf{O}_{3(1)} \hat{\mathbf{b}}_{1:4} = 2.778 \begin{bmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{bmatrix} \begin{bmatrix} 0.2 & 0 \\ 0 & 0.6 \end{bmatrix} \begin{bmatrix} 0.782 \\ 1.107 \end{bmatrix} = \begin{bmatrix} 0.717 \\ 1.281 \end{bmatrix} \end{aligned}$$

Los valores del vector  $\hat{\mathbf{b}}_{t-1:4}$ ,  $t = 4, \dots, 1$ , no suman 1 porque se ha escalado usando el factor  $c_t$  obtenido en el paso *forward*. Como se mencionó, las probabilidades *backward* se pueden normalizar de manera que sumen 1; en tal caso, se interpretan como la verosimilitud de cada estado al tiempo  $t$ , dadas las observaciones futuras. Finalmente, se calculan las probabilidades *smoothed* de acuerdo con la Ec. (3.15):

$$\begin{aligned} p(z_4|y_{1:4}) &= \hat{\mathbf{f}}_{0:4} \circ \hat{\mathbf{b}}_{4:4} = (0.8734 \quad 0.1266) * (1 \quad 1) = (0.87 \quad 0.13) \\ p(z_3|y_{1:4}) &= \hat{\mathbf{f}}_{0:3} \circ \hat{\mathbf{b}}_{3:4} = (0.5816 \quad 0.4184) * (1.173 \quad 0.759) = (0.68 \quad 0.32) \\ p(z_2|y_{1:4}) &= \hat{\mathbf{f}}_{0:2} \circ \hat{\mathbf{b}}_{2:4} = (0.276 \quad 0.724) * (1.199 \quad 0.924) = (0.33 \quad 0.67) \\ p(z_1|y_{1:4}) &= \hat{\mathbf{f}}_{0:1} \circ \hat{\mathbf{b}}_{1:4} = (0.333 \quad 0.667) * (0.782 \quad 1.107) = (0.26 \quad 0.74) \\ p(z_0|y_{1:4}) &= \hat{\mathbf{f}}_{0:0} \circ \hat{\mathbf{b}}_{0:4} = (0.5 \quad 0.5) * (0.717 \quad 1.281) = (0.36 \quad 0.64) \end{aligned}$$

donde "\*" denota el producto elemento por elemento. Los valores obtenidos representan las probabilidades de cada estado (*sol*, *lluvia*) a diferentes tiempos. Los resultados dan evidencia de que el estado del clima más probable para los dos últimos días es soleado, mientras que para los primeros es lluvia.

#### 3.3.2. Sistemas dinámicos lineales

Los **HMM** corresponden a modelos de espacio-estado en los que las variables latentes son discretas, mientras que las variables observadas pueden ser discretas o continuas. En cambio, si las variables observadas y las variables latentes son Gaussianas, entonces se trata de un sistema dinámico lineal Gaussiano, o simplemente

### 3.3. Modelos de espacio-estado

---

sistema dinámico lineal cuando no hay ambigüedad. Adicionalmente, los sistemas dinámicos lineales (LDS) puede ser de tiempo discreto, por ejemplo como los que se caracterizan en Kalman (1960); West & Harrison (1997); Bishop (2006); Caron et al. (2008); Petris et al. (2009); o de tiempo continuo, como los que define Kalman (1963) (aunque en este caso el espacio-estado es real-lineal- $n$ -dimensional, pero no necesariamente Gaussiano). Los LDS caracterizados en lo que resta de este capítulo son Gaussianos y de tiempo discreto, y serán referidos como LDS; sin embargo, esta misma clase de modelos se encuentran en la literatura como *dynamic linear models* (DLM), por ejemplo en West & Harrison (1997) y Petris et al. (2009).

Los LDS relacionan a un vector de observaciones  $y_t$  de tamaño  $(r \times 1)$ ,  $t = 1, 2, \dots$ , con un vector de parámetros (estados)  $z_t$  de tamaño  $(p \times 1)$  a través de las siguientes ecuaciones:

$$y_t = C_t' z_t + w_t, \quad w_t \sim N[0, R_t] \quad (3.16)$$

$$z_t = A_t z_{t-1} + e_t, \quad e_t \sim N[0, \Sigma_t] \quad (3.17)$$

donde:

- $C_t$  es una matriz diseño conocida, de tamaño  $(p \times r)$
- $A_t$  es una matriz de evolución conocida, de tamaño  $(p \times p)$
- $R_t$  es una matriz de covarianzas conocida, de tamaño  $(r \times r)$
- $\Sigma_t$  es una matriz de covarianzas conocida, de tamaño  $(p \times p)$
- $w_t$  es el vector de error observacional, de tamaño  $(r \times 1)$
- $e_t$  es el vector de error de evolución, de tamaño  $(p \times 1)$

La ecuación (3.16) se conoce como *ecuación de observación* y la ecuación (3.17) como *ecuación de evolución* o de *estados*.  $z_t$  contiene los estados no observados del sistema que se asume evolucionan con el tiempo de acuerdo con la matriz de evolución  $A_t$ . Por convención, la serie se desarrolla igualmente espaciada en tiempo  $(y_1, y_2, \dots)$ ; sin embargo, este supuesto no es necesario. El carácter dinámico del modelo hace que este sea sólo localmente apropiado para cada  $t$ . En la ecuación (3.17), el modelo es apropiado en el tiempo  $t$  hasta que  $y_t$  es observada. Entonces el modelo se actualiza con la nueva información. En cada actualización, adicional a la observación más reciente, la nueva información puede consistir de información



### 3.3. Modelos de espacio-estado

---

externa relevante; sin embargo, el modelo se considera *cerrado* a información de este tipo (West & Harrison, 1997).

Adicionalmente, se asume que  $z_0$  tiene distribución Gaussiana con  $m_0$  y  $S_0$  como los momentos a priori:

$$z_0 \sim N(m_0, S_0). \quad (3.18)$$

El conjunto  $\{C_t, A_t, \Sigma_t, R_t\}$  representa los parámetros del modelo (3.16)-(3.17), donde cada uno de los elementos puede o no estar indexado por el tiempo. Las secuencias de errores  $w_t$  y  $e_t$  son interna y mutuamente independientes, e independientes de  $z_0$ . Note que si no hay error de evolución ( $e_t = 0, \Sigma_t = 0$ ) y  $A_t = I$ , entonces el sistema dinámico se reduce a un sistema lineal estático ( $z_t = z$  para toda  $t$ ). Adicionalmente, si  $C_t' = (X_{t1}, \dots, X_{tp})$ , donde  $X_{t,1:p}$  es una colección de variables independientes al tiempo  $t$ , conocidas como regresores en el contexto de *regresión*, el LDS definido por  $\{I, C_t, \Sigma_t, R_t\}$  es un modelo de regresión dinámica, con respuesta media  $\mu_t$  dada por  $\mu_t = C_t' z_t = \sum_{i=1}^p z_{ti} X_{ti}$ . En este caso, los estados  $z_t$  tienen la interpretación de los coeficientes de regresión, es decir, representan la contribución de los regresores a los cambios en la media de la variable respuesta  $y_t$ ; estos coeficientes de regresión varían con el tiempo, contrario a un modelo de regresión estática.

La formulación de un LDS puede ser vista como un caso especial de un modelo jerárquico con tres niveles: las observaciones, descritas por la ecuación de observación; el proceso de evolución de los estados, descrito por la ecuación de evolución; y los parámetros que definen el modelo  $\{A_t, C_t, \Sigma_t, R_t\}$ . Adicionalmente, la formulación del modelo se puede extender a modelos no lineales y errores no Gaussianos, y modelos espacio-temporales.

Los LDS tienen la misma estructura de independencia de los HMM dada en la Ec. (3.6), por lo que la forma general de los algoritmos para inferencia es igual, excepto que las sumatorias sobre las variables latentes se reemplazan por integrales. En el paso *forward* se obtiene la probabilidad a posteriori de los estados  $z_t$  dadas las observaciones  $y_{1:t}$ ; las ecuaciones resultantes son análogas a las obtenidas en el HMM, pero se conocen como ecuaciones *Kalman filter* (Kalman, 1960). En el paso *backward*, las ecuaciones análogas al término  $\beta(\cdot)$  en el HMM se conocen como ecuaciones *Kalman smoother*.

Debido a que un LDS es un modelo Gaussiano, la distribución conjunta de todas las variables, así como las distribuciones marginales y condicionales implicadas, son Gaussianas. De esto se sigue que la secuencia de variables latentes individuales más probables es la misma que la secuencia latente más probable (ver Barber, 2012), por lo que no es necesario un algoritmo análogo al algoritmo Viterbi para

### 3.3. Modelos de espacio-estado

---

LDS.

#### Kalman filter

El *Kalman filter* se usa para estimar el estado actual,  $z_t$ , condicionado a las observaciones disponibles hasta el tiempo  $t$ , es decir, la distribución de probabilidad de interés es  $p(z_t|y_{1:t})$ . Esto se logra mediante un procedimiento recursivo *forward* en tiempo que involucra dos fases: (1) antes de observar  $y_t$  (predicción) y; (2) después de observar  $y_t$  (actualización). Estas fases son consecuencia inmediata de la aplicación del teorema de Bayes:

$$p(z_t|y_{1:t}) = \frac{p(y_t|z_t, y_{1:t-1})p(z_t|y_{1:t-1})}{\int p(y_t|z_t, y_{1:t-1})p(z_t|y_{1:t-1})dz_t} = \frac{p(y_t|z_t)p(z_t|y_{1:t-1})}{\int p(y_t|z_t)p(z_t|y_{1:t-1})dz_t} \quad (3.19)$$

En la Ec. (3.19) la expresión del lado izquierdo,  $p(z_t|y_{1:t})$ , denota la distribución a posteriori de  $z$  al tiempo  $t$ . Del lado derecho, los términos del numerador representan la verosimilitud y la distribución a priori, respectivamente; el denominador es un factor de normalización.

Para derivar el procedimiento recursivo se comienza asumiendo que en el momento  $t - 1$  el conocimiento que se tiene sobre  $z_t$  está dado por:

$$p(z_{t-1}|y_{1:t-1}) = N(f_{t-1}, F_{t-1}) \quad (3.20)$$

Adicionalmente, note que la distribución de probabilidad asociada con la fase de predicción está dada por:

$$p(z_t|y_{1:t-1}) = \int p(z_t|z_{t-1})p(z_{t-1}|y_{1:t-1})dz_{t-1} \quad (3.21)$$

donde la distribución de probabilidad asociada con la transición de estados del momento  $t - 1$  al momento  $t$  es normal, con media  $A_t z_{t-1}$  y matriz de varianzas  $\Sigma_t$ , de acuerdo con la ecuación de evolución (3.17). Combinando las dos distribuciones

### 3.3. Modelos de espacio-estado

se tiene:

$$\begin{aligned}
p(z_t|y_{1:t-1}) &\propto \int \exp \left\{ -\frac{1}{2}(z_t - A_t z_{t-1})' \Sigma_t^{-1} (z_t - A_t z_{t-1}) \right\} \\
&\quad \cdot \exp \left\{ -\frac{1}{2}(z_{t-1} - f_{t-1})' F_{t-1}^{-1} (z_{t-1} - f_{t-1}) \right\} dz_{t-1} \\
&\propto \int \exp \left\{ -\frac{1}{2}(z_t' \Sigma_t^{-1} z_t - z_{t-1}' A_t' \Sigma_t^{-1} z_t - z_t' \Sigma_t^{-1} A_t z_{t-1} + z_{t-1}' A_t' \Sigma_t^{-1} A_t z_{t-1}) \right\} \\
&\quad \cdot \exp \left\{ -\frac{1}{2}(z_{t-1}' F_{t-1}^{-1} z_{t-1} - f_{t-1}' F_{t-1}^{-1} z_{t-1} - z_{t-1}' F_{t-1}^{-1} f_{t-1}) \right\} dz_{t-1} \\
&= \int \exp \left\{ -\frac{1}{2} \begin{bmatrix} z_t \\ z_{t-1} \end{bmatrix}' \begin{bmatrix} \Sigma_t^{-1} & -\Sigma_t^{-1} A_t \\ -A_t' \Sigma_t^{-1} & A_t' \Sigma_t^{-1} A_t \end{bmatrix} \begin{bmatrix} z_t \\ z_{t-1} \end{bmatrix} \right\} \\
&\quad \cdot \exp \left\{ -\frac{1}{2} \begin{bmatrix} z_t \\ z_{t-1} \end{bmatrix}' \begin{bmatrix} 0 & 0 \\ 0 & F_{t-1}^{-1} \end{bmatrix} \begin{bmatrix} z_t \\ z_{t-1} \end{bmatrix} + \begin{bmatrix} z_t \\ z_{t-1} \end{bmatrix}' \begin{bmatrix} 0 \\ F_{t-1}^{-1} f_{t-1} \end{bmatrix} \right\} dz_{t-1} \\
&= \int \exp \left\{ -\frac{1}{2} \begin{bmatrix} z_t \\ z_{t-1} \end{bmatrix}' \begin{bmatrix} \Sigma_t^{-1} & -\Sigma_t^{-1} A_t \\ -A_t' \Sigma_t^{-1} & A_t' \Sigma_t^{-1} A_t + F_{t-1}^{-1} \end{bmatrix} \begin{bmatrix} z_t \\ z_{t-1} \end{bmatrix} \right\} \\
&\quad \cdot \exp \left\{ \begin{bmatrix} z_t \\ z_{t-1} \end{bmatrix}' \begin{bmatrix} 0 \\ F_{t-1}^{-1} f_{t-1} \end{bmatrix} \right\} dz_{t-1}
\end{aligned}$$

De las propiedades de la función de densidad de probabilidad normal se sigue que  $p(z_t|y_{1:t-1}) \propto N(m_t, S_t)$ , donde  $m_t$  y  $S_t$  son tales que:

$$\begin{aligned}
S_t &= [\Sigma_t^{-1} - \Sigma_t^{-1} A_t (A_t' \Sigma_t^{-1} A_t + F_{t-1}^{-1})^{-1} A_t' \Sigma_t^{-1}]^{-1} \\
m_t &= S_t \Sigma_t^{-1} A_t (A_t' \Sigma_t^{-1} A_t + F_{t-1}^{-1})^{-1} F_{t-1}^{-1} f_{t-1}
\end{aligned}$$

Es fácil probar que la varianza  $S_t$  se puede reescribir, por el Teorema de Woodbury, como ([Henderson & Searle, 1981](#)):

$$S_t = \Sigma_t + A_t F_{t-1} A_t'$$

Usando la identidad anterior y los resultados conocidos para la inversión de suma de matrices (ver [Henderson & Searle, 1981](#)),  $m_t$  se puede simplificar como:

$$\begin{aligned}
m_t &= (\Sigma_t + A_t F_{t-1} A_t') \Sigma_t^{-1} A_t [I - F_{t-1} A_t' \Sigma_t^{-1} A_t (I + F_{t-1} A_t' \Sigma_t^{-1} A_t)^{-1}] f_{t-1} \\
&= A_t [I + F_{t-1} A_t' \Sigma_t^{-1} A_t] [I - F_{t-1} A_t' \Sigma_t^{-1} A_t (I + F_{t-1} A_t' \Sigma_t^{-1} A_t)^{-1}] f_{t-1} \\
&= \{A_t [I + F_{t-1} A_t' \Sigma_t^{-1} A_t] - A_t F_{t-1} A_t' \Sigma_t^{-1} A_t\} f_{t-1} \\
&= A_t f_{t-1}
\end{aligned}$$

Alternativamente, se puede usar directamente la ecuación de estados ([3.17](#)) para obtener los parámetros de la densidad predictiva  $p(z_t|y_{1:t-1})$ :

$$\begin{aligned}
E(z_t|y_{1:t-1}) &= E(A_t z_{t-1} + e_t|y_{1:t-1}) = A_t f_{t-1} \\
V(z_t|y_{1:t-1}) &= V(A_t z_{t-1} + e_t|y_{1:t-1}) = A_t F_{t-1} A_t' + \Sigma_t
\end{aligned}$$

### 3.3. Modelos de espacio-estado

Una vez que se observa  $y_t$ , el objetivo es encontrar la distribución a posteriori para  $z_t$ . Introduciendo (3.16) y el resultado de (3.21) en (3.19):

$$\begin{aligned}
p(z_t|y_{1:t}) &= \frac{p(y_t|z_t)p(z_t|y_{1:t-1})}{\int p(y_t|z_t)p(z_t|y_{1:t-1})dz_t} \propto p(y_t|z_t)p(z_t|y_{1:t-1}) \\
&\propto \exp \left\{ -\frac{1}{2}(y_t - C'_t z_t)' R_t^{-1} (y_t - C'_t z_t) \right\} \exp \left\{ -\frac{1}{2}(z_t - m_t)' \right. \\
&\quad \left. S_t^{-1} (z_t - m_t) \right\} \\
&\propto \exp \left\{ -\frac{1}{2}(y'_t R_t^{-1} y_t - y'_t R_t^{-1} C'_t z_t - z'_t C_t R_t^{-1} y_t + z'_t C_t R_t^{-1} C'_t z_t \right. \\
&\quad \left. + z'_t S_t^{-1} z_t + m'_t S_t^{-1} m_t - z'_t S_t^{-1} m_t - m'_t S_t^{-1} z_t) \right\} \\
&\propto \exp \left\{ -\frac{1}{2} \left[ z_t - (S_t^{-1} + C_t R_t^{-1} C'_t)^{-1} (C_t R_t^{-1} y_t + S_t^{-1} m_t) \right]' \right. \\
&\quad \left. (S_t^{-1} + C_t R_t^{-1} C'_t) \left[ z_t - (S_t^{-1} + C_t R_t^{-1} C'_t)^{-1} (C_t R_t^{-1} y_t + S_t^{-1} m_t) \right] \right. \\
&\quad \left. + \frac{1}{2} (C_t R_t^{-1} y_t + S_t^{-1} m_t)' (S_t^{-1} + C_t R_t^{-1} C'_t)^{-1} (C_t R_t^{-1} y_t + S_t^{-1} m_t) \right. \\
&\quad \left. - \frac{1}{2} y'_t R_t^{-1} y_t \right\} \tag{3.22}
\end{aligned}$$

De (3.22) se sigue que  $p(z_t|y_{1:t})$  es  $N(f_t, F_t)$ , donde:

$$F_t = (S_t^{-1} + C_t R_t^{-1} C'_t)^{-1} = S_t - S_t C_t (C'_t S_t C_t + R_t)^{-1} C'_t S_t \tag{3.23}$$

$$f_t = F_t (C_t R_t^{-1} y_t + S_t^{-1} m_t) = m_t + S_t C_t (C'_t S_t C_t + R_t)^{-1} (y_t - C'_t m_t) \tag{3.24}$$

Las segundas igualdades de las Ec. (3.23) y (3.24) se pueden verificar siguiendo los resultados de [Henderson & Searle \(1981\)](#). Por otra parte, en muchas aplicaciones es de interés conocer la distribución predictiva para las observaciones, representada por el denominador  $\int p(y_t|z_t)p(z_t|y_{1:t-1})dz_t = p(y_t|y_{1:t-1})$  en la Ec. (3.19). Esta

### 3.3. Modelos de espacio-estado

distribución se obtiene al integrar  $z_t$  en (3.22):

$$\begin{aligned}
& \int p(y_t|z_t)p(z_t|y_{1:t-1})dz_t \propto \int \exp \left\{ -\frac{1}{2} \left[ z_t - (S_t^{-1} + C_t R_t^{-1} C_t')^{-1} \right. \right. \\
& \left. \left. (C_t R_t^{-1} y_t + S_t^{-1} m_t) \right]' (S_t^{-1} + C_t R_t^{-1} C_t') \left[ z_t - (S_t^{-1} + C_t R_t^{-1} C_t')^{-1} (C_t R_t^{-1} y_t + S_t^{-1} m_t) \right] \right. \\
& \left. + \frac{1}{2} (C_t R_t^{-1} y_t + S_t^{-1} m_t)' (S_t^{-1} + C_t R_t^{-1} C_t')^{-1} (C_t R_t^{-1} y_t + S_t^{-1} m_t) - \frac{1}{2} y_t' R_t^{-1} y_t \right\} dz_t \\
& \propto \exp \left\{ \frac{1}{2} (C_t R_t^{-1} y_t + S_t^{-1} m_t)' (S_t^{-1} + C_t R_t^{-1} C_t')^{-1} (C_t R_t^{-1} y_t + S_t^{-1} m_t) - \frac{1}{2} y_t' R_t^{-1} y_t \right\} \\
& \propto \exp \left\{ \frac{1}{2} (C_t R_t^{-1} y_t + S_t^{-1} m_t)' [S_t - S_t C_t (R_t + C_t' S_t C_t)^{-1} C_t' S_t] (C_t R_t^{-1} y_t + S_t^{-1} m_t) \right. \\
& \quad \left. - \frac{1}{2} y_t' R_t^{-1} y_t \right\} \\
& \propto \exp \left\{ \frac{1}{2} \left[ y_t' R_t^{-1} C_t' S_t C_t R_t^{-1} y_t + m_t' S_t^{-1} S_t C_t R_t^{-1} y_t + y_t' R_t^{-1} C_t' S_t S_t^{-1} m_t \right. \right. \\
& \quad \left. \left. - (y_t' R_t^{-1} C_t' S_t C_t (R_t + C_t' S_t C_t)^{-1} C_t' S_t C_t R_t^{-1} y_t) - 2 y_t' R_t^{-1} C_t' S_t C_t (R_t + C_t' S_t C_t)^{-1} \right. \right. \\
& \quad \left. \left. C_t' S_t S_t^{-1} m_t \right] - \frac{1}{2} y_t' R_t^{-1} y_t \right\} \\
& \propto \exp \left\{ -\frac{1}{2} y_t' \left[ R_t^{-1} + R_t^{-1} C_t' S_t C_t (R_t + C_t' S_t C_t)^{-1} C_t' S_t C_t R_t^{-1} - R_t^{-1} C_t' S_t C_t R_t^{-1} \right] y_t \right. \\
& \quad \left. - y_t' R_t^{-1} C_t' S_t C_t (R_t + C_t' S_t C_t)^{-1} C_t' m_t + y_t' R_t^{-1} C_t' m_t \right\} \\
& \propto \exp \left\{ -\frac{1}{2} y_t' \left[ R_t - R_t^{-1} C_t' S_t [I - C_t (R_t + C_t' S_t C_t)^{-1} C_t' S_t] C_t R_t^{-1} \right] y_t \right. \\
& \quad \left. - y_t' R_t^{-1} C_t' S_t C_t (R_t + C_t' S_t C_t)^{-1} C_t' m_t + y_t' R_t^{-1} C_t' m_t \right\} \\
& \propto \exp \left\{ -\frac{1}{2} y_t' \left[ R_t - R_t^{-1} C_t' S_t (I + C_t R_t^{-1} C_t' S_t)^{-1} C_t R_t^{-1} \right] y_t \right. \\
& \quad \left. - y_t' R_t^{-1} C_t' S_t C_t (R_t + C_t' S_t C_t)^{-1} C_t' m_t + y_t' R_t^{-1} C_t' m_t \right\} \\
& \propto \exp \left\{ -\frac{1}{2} \left[ y_t' (R_t + C_t' S_t C_t)^{-1} y_t - 2 y_t' [R_t^{-1} - R_t^{-1} C_t' S_t C_t (R_t + C_t' S_t C_t)^{-1}] C_t' m_t \right] \right\} \\
& \propto \exp \left\{ -\frac{1}{2} (y_t - C_t' m_t)' (R_t + C_t' S_t C_t)^{-1} (y_t - C_t' m_t) \right\} \tag{3.25}
\end{aligned}$$

Entonces, la distribución predictiva de un paso, es decir, de un punto adelante en el tiempo, es  $N(a_t, Q_t)$ , donde:

$$\begin{aligned}
a_t &= C_t' m_t \\
Q_t &= R_t + C_t' S_t C_t
\end{aligned}$$

El filtro de Kalman permite calcular las densidades predictivas recursivamente. Comenzando con  $z_0|y_0 \sim N(f_0, F_0)$  calcular  $p(z_1|y_{1:1})$  y proceder de manera recursiva conforme se dispone de nueva información. Note que la media *filter* de la Ec.

### 3.3. Modelos de espacio-estado

---

(3.24) es igual a la media de predicción  $m_t$  más una corrección dependiendo de qué tanto la nueva información  $y_t$  difiere de su predicción  $C_t' m_t$  ( $y_t - C_t' m_t$  es el error de pronóstico). El peso del término de corrección está dado por  $S_t C_t (C_t' S_t C_t + R_t)^{-1}$ , que depende de la matriz de varianzas de las observaciones  $R_t$  y de la varianza  $S_t = V[z_t | y_{1:t-1}]$ .

La siguiente sección ilustra las principales características de un **LDS** con un modelo de primer orden con varianzas conocidas. El desarrollo de modelos más generales se puede consultar en [West & Harrison \(1997\)](#).

#### Ejemplo: **LDS** de primer orden con varianzas conocidas

Para cada  $t$ , el **LDS** de primer orden, caracterizado por  $\{1, 1, R_t, \Sigma_t\}$ , se describe de la siguiente manera:

$$y_t = \mu_t + w_t, \quad w_t \sim N[0, R_t] \tag{3.26}$$

$$\mu_t = \mu_{t-1} + e_t, \quad e_t \sim N[0, \Sigma_t] \tag{3.27}$$

$$[\mu_0 | D_0] \sim N[f_0, F_0] \tag{3.28}$$

donde para toda  $t$  y  $s$  con  $t \neq s$ ,  $w_t$  y  $w_s$  son independientes,  $e_t$  y  $e_s$  son independientes, y  $w_t$  y  $e_s$  son independientes. Adicionalmente, los errores son independientes de la información inicial. En esta formulación, el *nivel*,  $\mu_t$ , es modelado como una caminata aleatoria:  $e_t$  representa cambios puramente aleatorios en el nivel entre el tiempo  $t - 1$  y  $t$ . Note también que  $R_t$  y  $\Sigma_t$  son escalares y no matrices, puesto que  $y_t$  y  $\mu_t$  son escalares. Las ecuaciones (3.26), (3.27) y (3.28) definen el modelo más simple posible, referido comúnmente en la literatura como modelo estable (*steady model*). Para cada incremento de tiempo, las siguientes distribuciones describen la actualización con respecto a la nueva información:

### 3.3. Modelos de espacio-estado

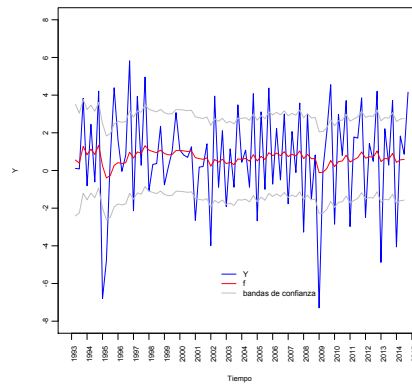
**Tabla 3.1:** Distribuciones para el proceso de actualización:  $\{1, 1, R_t, \Sigma_t\}$

(a) <b>A posteriori para <math>\mu_{t-1}</math>:</b>	$(\mu_{t-1} y_{1:t-1}) \sim N[f_{t-1}, F_{t-1}]$
(b) <b>A priori para <math>\mu_t</math>:</b>	$(\mu_t y_{1:t-1}) \sim N[f_{t-1}, S_t]$ donde $S_t = F_{t-1} + \Sigma_t$
(c) <b>Pronóstico de 1-paso:</b>	$(y_t y_{1:t-1}) \sim N[a_t, Q_t]$ donde $a_t = f_{t-1}$ y $Q_t = S_t + R_t$
(d) <b>A posteriori para <math>\mu_t</math>:</b>	$(\mu_t y_{1:t}) \sim N[f_t, F_t]$ con $f_t = f_{t-1} + V_t e_t$ y $F_t = V_t R_t$ donde $V_t = S_t / (S_t + R_t) = S_t / Q_t$ y $e_t = y_t - f_{t-1}$

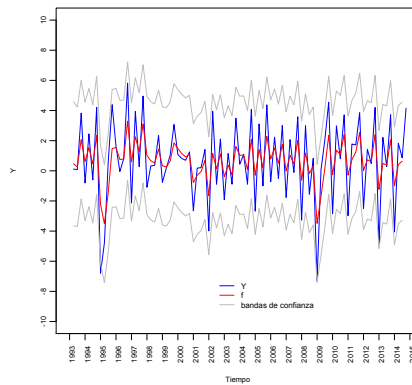
La media de  $\mu_t|y_{1:t}$  es calculada como la media previa corregida por el error de pronóstico, ponderado por  $V_t$ , un valor entre 0 y 1 conocido como el coeficiente adaptativo (West & Harrison, 1997). Para ver con más claridad el papel de  $V_t$ , la media se puede representar como  $f_t = (1 - V_t)f_{t-1} + V_t y_t$  que muestra a  $f_t$  como un promedio ponderado entre la estimación a priori,  $f_{t-1}$ , y la observación  $y_t$ . Entre más grande sea la varianza observacional  $R_t$  con respecto a la varianza de la apriori de  $\mu_t$ ,  $S_t$ , entonces  $V_t \rightarrow 0$ , y la media de la distribución a posteriori está más concentrada cerca de la media de la distribución a priori. Si  $V_t$  es cercano a 1, la a priori es menos informativa que los datos (o la verosimilitud).

En la Figura 3.3 se muestra el conjunto de datos del crecimiento del producto interno bruto trimestral ( $Y$ ) y el pronóstico ( $f$ ) suponiendo una caminata aleatoria y asumiendo a  $R_t$  y  $\Sigma_t$  conocidas y constantes para toda  $t$  (es decir  $\{1, 1, R, \Sigma\}$ , conocido como modelo constante). Distintos valores de  $\Sigma$  fueron usados para ilustrar el papel del coeficiente adaptativo en el pronóstico. Los valores iniciales asumidos en todos los casos fueron  $f_0 = F_0 = 1$  y  $\mu_0$  se tomó de una  $N(f_0 = 1, F_0 = 1)$ . La línea azul representa los valores observados, la línea roja la media  $a_t$ , y las líneas gris las bandas de confianza al 95 %, calculadas como es usual:  $a_t \pm 1.96\sqrt{Q_t}$ . En (a) la varianza de evolución,  $\Sigma = 0.01$ , es muy pequeña comparada con la varianza observacional,  $R = 1$ , lo que produce un valor muy pequeño para  $V_t$  y como consecuencia una media de pronóstico casi constante; en (b) hay más variación en el nivel  $\mu_t$  debido a que la varianza  $\Sigma$  no es tan pequeña comparada con  $R$ ; y en (c)  $\Sigma$  es dos veces  $R$ , lo que resulta en  $V_t$  cercana a 1, dando mucho peso a cada información adicional observada.

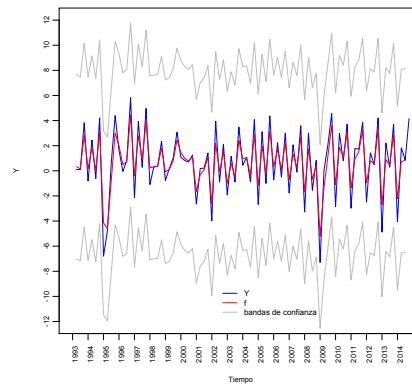
### 3.3. Modelos de espacio-estado



(a)



(b)



(c)

**Figura 3.3:** Valores pronosticados y bandas de confianza de 95% para el crecimiento del producto interno bruto en México. (a)  $\Sigma = 0.01$ ; (b)  $\Sigma = 0.5$ ; (c)  $\Sigma = 2$ . En todos los casos  $R = 1$ .

El caso particular  $V_t = 1$  para  $t \geq 2$  corresponde a  $R = 0$ . Entonces  $f_t = y_t$  y el modelo es de poca utilidad para predicción. La distribución de pronóstico descrita en la Tabla 3.1 corresponde a 1-paso, es decir, un punto en el tiempo más allá de  $t$ . Es importante considerar que este LDS es útil sólo para pronósticos de corto plazo, y particularmente en casos en los que la varianza de observación,  $R_t$ , es considerablemente más grande que la varianza del nivel,  $\Sigma_t$  (West & Harrison, 1997).



### 3.3. Modelos de espacio-estado

#### Kalman smoother

En análisis de series de tiempo es común tener observaciones,  $y_t$ , para un cierto periodo ( $t = 1, \dots, T$ ), y se desea reconstruir retrospectivamente el comportamiento del sistema que generó las observaciones. Dicho de otra manera, el objetivo es encontrar las densidades condicionales de  $z_t|y_{1:T}$ ,  $t < T$ , para estimar toda la historia de estados recursivamente *backward* en el tiempo. Estas densidades se obtienen marginalizando  $z_t$  de  $p(z_t, z_{t+1}|y_{1:T})$ :

$$\begin{aligned} p(z_t|y_{1:T}) &= \int p(z_t, z_{t+1}|y_{1:T}) dz_{t+1} = \int p(z_{t+1}|y_{1:T}) p(z_t|z_{t+1}, y_{1:T}) dz_{t+1} \\ &= \int p(z_{t+1}|y_{1:T}) p(z_t|z_{t+1}, y_{1:t}) dz_{t+1} \\ &= \int p(z_{t+1}|y_{1:T}) \frac{p(z_{t+1}|z_t, y_{1:t}) p(z_t|y_{1:t})}{p(z_{t+1}|y_{1:t})} dz_{t+1} \\ &= \int p(z_{t+1}|y_{1:T}) \frac{p(z_{t+1}|z_t) p(z_t|y_{1:t})}{p(z_{t+1}|y_{1:t})} dz_{t+1}. \end{aligned}$$

Se comienza suponiendo que  $p(z_{t+1}|y_{1:T}) = N(b_{t+1}, B_{t+1})$ . La distribución  $p(z_{t+1}|z_t)$  está definida por la ecuación de estados, de tal manera que  $p(z_{t+1}|z_t) = N(A_{t+1}z_t, \Sigma_{t+1})$ . La distribución  $p(z_t|y_{1:t})$  es la obtenida en el filtro de Kalman y  $p(z_{t+1}|y_{1:t}) = N(m_{t+1}, S_{t+1})$  es la distribución de predicción de los estados en  $t + 1$ , de acuerdo con la Ec. (3.21). Entonces, de las propiedades de la distribución Normal,  $p(z_t|y_{1:T})$  es también Normal. Una manera fácil de encontrar los parámetros que la definen es usar la ley de la esperanza total y la ley de la varianza total:

$$\begin{aligned} E(z_t|y_{1:T}) &= E(E(z_t|z_{t+1}, y_{1:T})|y_{1:T}) \\ V(z_t|y_{1:T}) &= V(E(z_t|z_{t+1}, y_{1:T})|y_{1:T}) + E(V(z_t|z_{t+1}, y_{1:T})|y_{1:T}). \end{aligned}$$

Y la distribución  $p(z_t|z_{t+1}, y_{1:T})$  se puede obtener como sigue:

$$\begin{aligned} p(z_t|z_{t+1}, y_{1:T}) &= \frac{p(z_{t+1}|z_t) p(z_t|y_{1:t})}{p(z_{t+1}|y_{1:t})} \propto p(z_{t+1}|z_t) p(z_t|y_{1:t}) \\ &\propto \exp \left\{ -\frac{1}{2} (z_{t+1} - A_{t+1}z_t)' \Sigma_{t+1}^{-1} (z_{t+1} - A_{t+1}z_t) \right\} \\ &\quad \cdot \exp \left\{ -\frac{1}{2} (z_t - f_t)' F_t^{-1} (z_t - f_t) \right\} \\ &\propto \exp \left\{ -\frac{1}{2} \left[ z_t' (A_{t+1}' \Sigma_{t+1}^{-1} A_{t+1} + F_t^{-1}) z_t \right. \right. \\ &\quad \left. \left. - 2z_t' (A_{t+1}' \Sigma_{t+1}^{-1} z_{t+1} + F_t^{-1} f_t) \right] \right\} \end{aligned}$$

### 3.3. Modelos de espacio-estado

---

Por tanto,

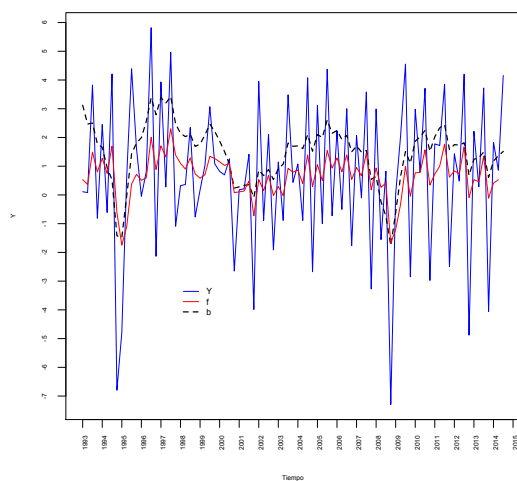
$$\begin{aligned}
V(z_t|z_{t+1}, y_{1:T}) &= (A'_{t+1}\Sigma_{t+1}^{-1}A_{t+1} + F_t^{-1})^{-1} \\
&= F_t - F_t A'_{t+1} (\Sigma_{t+1} + A_{t+1} F_t A'_{t+1})^{-1} A_{t+1} F_t \\
&= F_t - F_t A'_{t+1} S_{t+1}^{-1} A_{t+1} F_t \\
E(z_t|z_{t+1}, y_{1:T}) &= (F_t - F_t A'_{t+1} S_{t+1}^{-1} A_{t+1} F_t) (A'_{t+1} \Sigma_{t+1}^{-1} z_{t+1} + F_t^{-1} f_t) \\
&= F_t A'_{t+1} S_{t+1}^{-1} (S_{t+1} - A_{t+1} F_t A'_{t+1}) \Sigma_{t+1}^{-1} z_{t+1} \\
&\quad - F_t A'_{t+1} S_{t+1}^{-1} A_{t+1} f_t + f_t \\
&= F_t A'_{t+1} S_{t+1}^{-1} z_{t+1} - F_t A'_{t+1} S_{t+1}^{-1} m_{t+1} + f_t \\
&= f_t + F_t A'_{t+1} S_{t+1}^{-1} (z_{t+1} - m_{t+1}).
\end{aligned}$$

Entonces, la distribución de interés  $p(z_t|y_{1:T})$  es  $N(b_t, B_t)$ , donde:

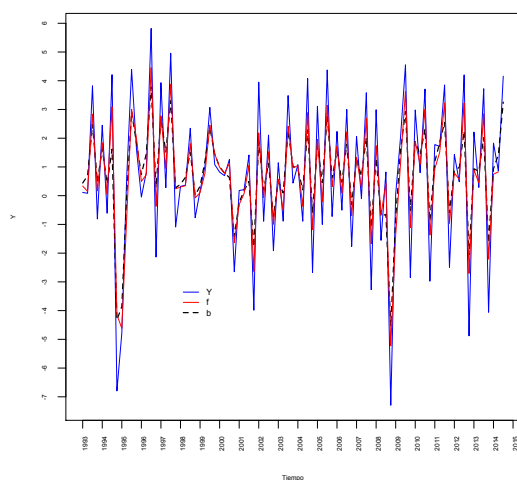
$$\begin{aligned}
b_t &= E(z_t|y_{1:T}) = E(E(z_t|z_{t+1}, y_{1:T})|y_{1:T}) = f_t + F_t A'_{t+1} S_{t+1}^{-1} (b_{t+1} - m_{t+1}) \\
B_t &= V(z_t|y_{1:T}) = V(E(z_t|z_{t+1}, y_{1:T})|y_{1:T}) + E(V(z_t|z_{t+1}, y_{1:T})|y_{1:T}) \\
&= F_t A'_{t+1} S_{t+1}^{-1} B_{t+1} S_{t+1}^{-1} A_{t+1} F_t + F_t - F_t A'_{t+1} S_{t+1}^{-1} A_{t+1} F_t \\
&= F_t + F_t A'_{t+1} S_{t+1}^{-1} (B_{t+1} - S_{t+1}) S_{t+1}^{-1} A_{t+1} F_t
\end{aligned}$$

El *Kalman smoother* permite calcular  $z_t|y_{1:T}$  comenzando con  $z_T|y_{1:T} \sim N(b_T = f_T, B_T = F_T)$  y después proceder *backward* en  $t$ . La Figura 3.4 muestra las medias *forward* ( $f_t$ ) y *backward* ( $b_t$ ) para los datos del crecimiento del producto interno bruto trimestral (Y) y el modelo definido por las Ec. (3.26)-(3.28), con  $\{R_t = R, \Sigma_t = \Sigma\}$  para toda  $t$ . En la gráfica (a):  $\{R = 1, \Sigma = 0.1\}$ ; en la gráfica (b):  $\{R = 1, \Sigma = 2\}$ . Los valores iniciales fueron como en el ejemplo anterior,  $f_0 = F_0 = 1$ . El ejemplo ilustra que las trayectorias estimadas por *Kalman smoother* tienden a ser más *suaves* que las obtenidas por el filtro de Kalman, esto como resultado de la información adicional disponible. Entonces, si no se requieren estimaciones de los estados al tiempo  $t$  instantáneamente, es conveniente utilizar las observaciones posteriores disponibles para obtener una mejor estimación.

### 3.3. Modelos de espacio-estado



(a)



(b)

**Figura 3.4:** Medias *forward*,  $f$ , y *backward*,  $b$ , para el crecimiento del producto interno bruto en México. (a)  $\Sigma = 0.1$ ,  $R = 1$ ; (b)  $\Sigma = 2$ ,  $R = 1$ .

#### 3.3.3. Sistemas dinámicos lineales con matrices de covarianzas desconocidas

El modelo representado por las Ec. (3.16)-(3.17) asume conocidas las matrices de covarianzas  $R_t$  y  $\Sigma_t$ ; sin embargo, en la práctica, raramente estas matrices son

### 3.3. Modelos de espacio-estado

completamente conocidas. Un caso simple es suponer que  $R_t$  y  $\Sigma_t$  son desconocidas solo por un factor de escala común, es decir,  $R_t = \sigma^2 \tilde{R}_t$  y  $\Sigma_t = \sigma^2 \tilde{\Sigma}_t$ , con  $\sigma^2$  desconocido y  $\{\tilde{R}_t, \tilde{\Sigma}_t\}$  conocidos. Un clásico ejemplo para la matriz de covarianzas del error observacional es  $R_t = \sigma^2 I$ , donde  $I$  es la matriz identidad, y se elige una distribución a priori conjugada para  $1/\sigma^2$  para llevar a cabo la inferencia. Es decir, suponga que se satisface lo siguiente:

$$\begin{aligned} R_t &= \sigma^2 \tilde{R}_t = \sigma^2 \\ \Sigma_t &= \sigma^2 \tilde{\Sigma}_t \\ F_0 &= \sigma^2 \tilde{F}_0 \\ \phi &= 1/\sigma^2 \\ (z_0, \phi) &\sim NG(f_0, F_0, \alpha_0, \beta_0). \end{aligned}$$

La elección a priori para  $(z_0, \phi)$  como Normal-Gamma es conveniente para los cálculos. Se comienza suponiendo que

$$z_{t-1}, \phi | y_{1:t-1} \sim NG(f_{t-1}, \tilde{F}_{t-1}, \alpha_{t-1}, \beta_{t-1}).$$

Usando un procedimiento semejante al de la sección anterior para encontrar  $z_t | y_{1:t-1}$ , es fácil verificar que  $z_t, \phi | y_{1:t-1}$  es  $NG(m_t, \tilde{S}_t, \alpha_{t-1}, \beta_{t-1})$ , donde

$$\begin{aligned} m_t &= A_t f_{t-1} \\ \tilde{S}_t &= A_t \tilde{F}_{t-1} A_t' + \tilde{\Sigma}_t. \end{aligned}$$

Entonces,  $z_t, \phi | y_{1:t}$  es  $NG(f_t, \tilde{F}_t, \alpha_t, \beta_t)$ , donde:

$$\tilde{F}_t = \tilde{S}_t - \tilde{S}_t C_t (C_t' \tilde{S}_t C_t + \tilde{R}_t)^{-1} C_t' \tilde{S}_t = \tilde{S}_t - \tilde{S}_t C_t \tilde{Q}_t^{-1} C_t' \tilde{S}_t \quad (3.29)$$

$$f_t = m_t + \tilde{S}_t C_t \tilde{Q}_t^{-1} (y_t - C_t' m_t) \quad (3.30)$$

$$\alpha_t = \alpha_{t-1} + \frac{p}{2} \quad (3.31)$$

$$\beta_t = \beta_{t-1} + \frac{1}{2} (y_t - C_t' m_t)' \tilde{Q}_t^{-1} (y_t - C_t' m_t). \quad (3.32)$$

Finalmente, asumiendo  $W_t = \beta_t / \alpha_t$ , la distribución  $z_t | y_{1:t}$  se obtiene marginalizando  $\phi$  de  $z_t, \phi | y_{1:t}$ . Es fácil verificar que  $z_t | y_{1:t} \sim T_{2\alpha_t}(f_t, F_t)$ , donde  $T_\alpha(b_t, B_t)$  es una distribución  $t$ -student no estandarizada con  $\alpha$  grados de libertad, y parámetros de localización y escala  $b_t$  y  $B_t$ , respectivamente. En este caso, el parámetro  $F_t$  está dado por:

$$F_t = \tilde{F}_t W_t$$

### 3.3. Modelos de espacio-estado

---

Casos más generales en los que la varianza del error observacional es estocástica y dependiente del tiempo se pueden consultar en [West & Harrison \(1997\)](#). Cuando el interés es muestrear de la distribución conjunta condicional  $z_{1:T}|\psi, y_{1:T}$ , donde  $\psi$  representa los parámetros desconocidos presentes en la especificación del modelo, el algoritmo *forward-filtering, backward-sampling* (FFBS) de [Fruhwirth-Schnatter \(1994\)](#), una versión del proceso recursivo de suavizado, es una aproximación eficiente para simular de esa distribución.

Como se ha señalado,  $\Sigma_t$  tiene un rol determinante en la estimación de los *estados*: si es grande, hay mucha incertidumbre en la evolución de los *estados* y se pierde gran cantidad de información al pasar de  $z_{t-1}$  a  $z_t$ ; es decir, la información de  $z_{t-1}$  transmitida por las observaciones pasadas  $y_{1:t-1}$ , es de poca relevancia en el pronóstico de  $z_t$ , por lo que  $y_t$  determina principalmente la estimación de  $z_t|y_{1:t}$ . Entonces, resulta fundamental para el modelo determinar la magnitud de la matriz de covarianzas  $\Sigma_t$ . Una sugerencia práctica es especificar la matriz  $\Sigma_t$  usando un factor de descuento ([West & Harrison, 1997](#)). El factor de descuento es fácil de entender cuando se asocia con la precisión en la predicción del vector de estados.

En el filtro de Kalman, la incertidumbre sobre  $z_{t-1}$  dados los datos  $y_{1:t-1}$  se resume en la matriz de covarianzas condicional  $V(z_{t-1}|y_{1:t-1}) = F_{t-1}$ ; al moverse de  $z_{t-1}$  a  $z_t$  mediante la ecuación de estados  $z_t = A_t z_{t-1} + e_t$ , la incertidumbre se incrementa de manera que  $V(z_t|y_{1:t-1}) = S_t = A_t F_{t-1} A_t' + \Sigma_t$ . Entonces, si  $\Sigma_t = 0$ , es decir, si no hay error en la ecuación de estados, se tiene  $V(z_t|y_{1:t-1}) = A_t F_{t-1} A_t' = P_t$ . El término  $P_t$  se puede ver como una varianza a priori para el [LDS](#) dado por  $\{C_t, A_t, R_t, 0\}$ , esto es, un vector de estados sin cambios estocásticos. Conforme  $V(z_t|y_{1:t-1})$  se incrementa ( $V(z_t|y_{1:t-1}) = P_t + \Sigma_t$ ),  $\Sigma_t$  expresa la pérdida de información de pasar de  $z_{t-1}$  a  $z_t$  debido al componente de error estocástico en la ecuación de estados; la pérdida depende de la magnitud de  $\Sigma_t$  con respecto a  $P_t$ . Expresando  $\Sigma_t$  como proporción de  $P_t$  se tiene:

$$\Sigma_t = \frac{1 - \delta}{\delta} P_t, \quad (3.33)$$

donde  $\delta \in (0, 1]$ . El caso  $\delta = 1$  corresponde al modelo estático. De (3.33) se sigue que  $S_t = 1/\delta P_t$ , con  $1/\delta > 1$ . El parámetro  $\delta$  se conoce como factor de descuento, debido a que descuenta o disminuye la precisión  $S_t^{-1}$  (aumenta la varianza) en una proporción  $\delta$  mediante  $\delta P_t^{-1}$ . En la práctica,  $\delta$  se elige entre 0.9 y 0.99; entre más alto es el factor de descuento, el modelo es más permanente en el sentido de que, en cualquier momento  $t$ , no hay cambios estocásticos, y por tanto incertidumbre, en la precisión asociada a  $z_t$ , dado  $y_{1:t-1}$  ([West & Harrison, 1997](#)).

### 3.3. Modelos de espacio-estado

---

#### Ejemplo: uso del factor de descuento

Para ilustrar el uso del factor de descuento considere de nuevo la serie de tiempo del incremento del producto interno bruto (Y) trimestral de 1993 a 2015, modelada por el LDS definido por las Ec. (3.26) a (3.28) con matrices  $R_t$  y  $\Sigma_t$  desconocidas, tales que:

$$\begin{aligned} R_t &= \sigma^2 \tilde{R}_t = \sigma^2 \\ \Sigma_t &= \sigma^2 \tilde{\Sigma}_t. \end{aligned}$$

Sea  $\phi = 1/\sigma^2$  y  $(z_0, \phi) \sim NG(f_0, \tilde{F}_0, \alpha_0, \beta_0)$ , con  $F_0 = \sigma^2 \tilde{F}_0$  y parámetros  $\alpha_0, \beta_0$  de manera que  $E(1/\phi)$  estima la varianza observacional  $\sigma^2$ :

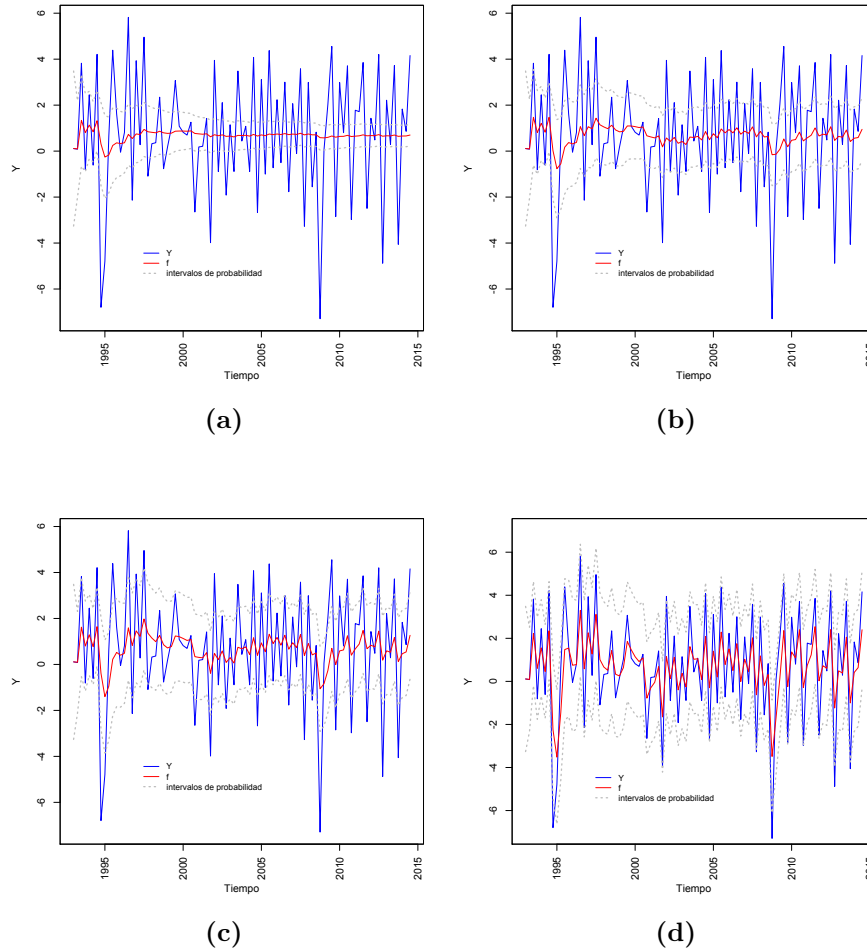
$$E(1/\phi) = \beta_0 / (\alpha_0 - 1).$$

Los cálculos de  $\alpha_t$  y  $\beta_t$  se hacen recursivamente usando (3.31) y (3.32):

$$\begin{aligned} \alpha_t &= \alpha_0 + \frac{t}{2} \\ \beta_t &= \beta_0 + \frac{1}{2} \sum_{i=1}^t (y_i - C'_i m_i)^2 \tilde{Q}_i^{-1}. \end{aligned}$$

Se usaron como valores iniciales para la distribución a posteriori de los *estados*:  $f_0 = 0$ ,  $\tilde{F}_0 = 10^6$  (la varianza a priori debe ser muy grande si hay mucha incertidumbre sobre la media a priori  $f_0$ ); y para la a priori de  $\sigma^2$ :  $\alpha_0 = 2$ ,  $\beta_0 = 7$ , tal que la estimación inicial para la varianza observacional es 7. Se examinaron 4 modelos definidos por distintos valores del *factor de descuento*: 1.0, 0.9, 0.8 y 0.5. La Figura (3.5) muestra los datos y las estimaciones *filter* de la Ec. (3.30), con intervalos de probabilidad de 90 %. Note sobreajuste conforme  $\delta$  disminuye.

### 3.4. Conclusiones



**Figura 3.5:** Valores *filtering* y bandas de confianza de 90 % para el crecimiento del producto interno bruto en México con distintos valores del *factor de descuento*. (a)  $\delta = 1.0$ ; (b)  $\delta = 0.9$ ; (c)  $\delta = 0.8$ ; (d)  $\delta = 0.5$ .

### 3.4. Conclusiones

En este capítulo se describieron las metodologías para estudiar dos tipos de modelos de espacio-estado: los modelos de Markov ocultos y los sistemas dinámicos lineales. Uno de los principales problemas de interés de estos modelos es encontrar la distribución marginal a posteriori de los estados, dada una secuencia de observaciones. Este problema se aborda mediante un algoritmo iterativo conocido como *forward-backward*. El desarrollo del algoritmo explota la estructura de in-

### 3.4. Conclusiones

---

dependencia del modelo. El paso *forward* estima la distribución de los estados al tiempo  $t$ , dada una secuencia de observaciones disponibles hasta  $t$ ; el paso *backward* estima la distribución de los estados al tiempo  $t$ , pero considerando una secuencia de observaciones que se extiende más allá de  $t$ . La trayectoria estimada en el paso *backward* tiende a ser más suave, como resultado de la información adicional empleada.

La formulación de los sistemas dinámicos lineales en la forma espacio-estado es consistente con los modelos de regresión dinámica, esto es, modelos de regresión en los que los coeficientes asociados a las variables independientes varían con el tiempo. Esta formulación es de gran utilidad para modelar series de tiempo no estacionarias: es una alternativa a la metodología tradicional que asume series de tiempo estacionarias y permite describir adecuadamente la relación entre variables conforme el tiempo evoluciona. Sin embargo, los sistemas dinámicos lineales son útiles sólo para pronósticos de corto plazo, y en casos en los que la varianza de observación es más grande que la varianza de los estados. Una manera práctica de determinar la magnitud de la varianza de los estados es usando un factor de descuento, un valor entre 0 y 1 que tiene el efecto de aumentar la varianza del pronóstico de los estados en una proporción igual al inverso de su valor, por lo que entre mayor sea el factor de descuento el modelo es más estable, en términos de la incertidumbre asociada al pronóstico de los estados.



# Capítulo 4

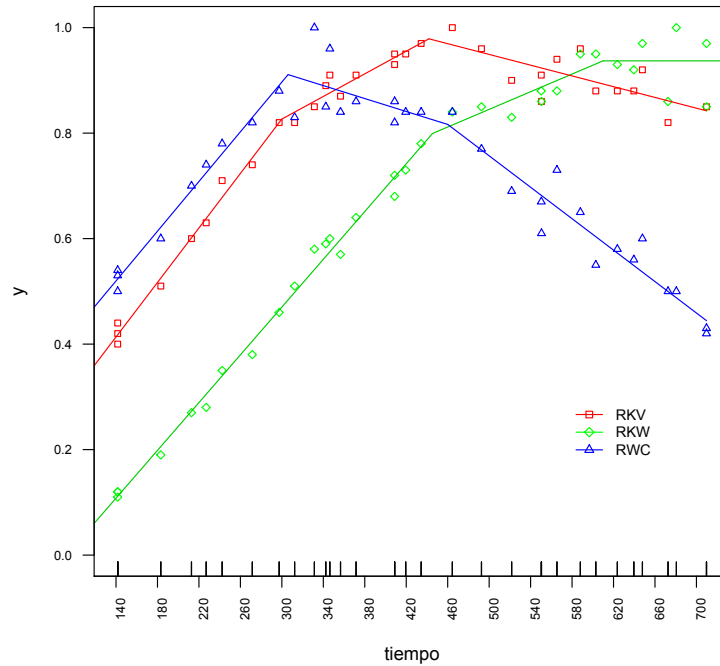
## SLDS para problemas de regresión

### 4.1. Introducción

La naturaleza dinámica de muchos sistemas y procesos demanda que los modelos usados para representarlos reconozcan la incertidumbre debida al paso del tiempo, de manera que la forma del modelo pueda cambiar y adaptarse a diversas circunstancias (West & Harrison, 1997). Considere la Figura 4.1. Las gráficas muestran las mediciones en el tiempo de tres atributos (RKV, RKW, RWC) del órgano de una planta. Los datos están disponibles en la librería *segmented* en R y fueron usados por Mugeo (2008) para ilustrar el ajuste de modelos de regresión con relaciones lineales segmentadas. En cada atributo se puede apreciar una relación lineal entre la respuesta y el tiempo que se ajusta mejor por segmentos. En este tipo de modelos de regresión en donde la relación entre la respuesta y una o más variables explicativas son lineales por piezas, es decir, representadas por dos o más segmentos de líneas rectas conectadas, es de interés conocer los puntos de quiebre (o de cambio) y las pendientes de la relación. Sistemas más complejos, con puntos de cambio llamados *switch*, son representados por alguna clase de sistemas dinámicos lineales de cambio de régimen, que consisten en una combinación de los modelos de Markov ocultos con un conjunto de sistemas dinámicos lineales que capturan la evolución en el tiempo del proceso que genera los datos, identifican los segmentos en los que las observaciones tienen características en común y proporcionan estimaciones de tales características.

## 4.1. Introducción

---



**Figura 4.1:** Mediciones en el tiempo del órgano de una planta.

Es común que el comportamiento de muchos sistemas sea estructuralmente complejo, y que no se pueda describir adecuadamente por un único modelo, pero que se pueda aproximar mediante una secuencia de modelos de un conjunto de sistemas lineales. Es decir, el sistema dinámico se divide en segmentos, cada uno modelado por un sistema dinámico lineal. Distintos segmentos pueden compartir un mismo modelo, pero modelos de segmentos adyacentes difieren. Estos cambios se pueden especificar de manera probabilística con base en una variable latente, de espacio discreto, que identifica el estado del sistema, llamado *modo*, en cada segmento. Cuando la variable latente es un proceso de Markov de tiempo discreto, el modelo se conoce como sistema dinámico lineal de cambio de régimen (**SLDS**). Un **SLDS** es una extensión de los modelos de Markov ocultos (**HMM**), pero a diferencia de estos, en los que se tienen observaciones condicionalmente independientes dada la secuencia de modos, en un **SLDS** cada modo está asociado con un proceso dinámico lineal.

El objetivo del presente capítulo es desarrollar un modelo de regresión dinámica para series de tiempo que identifique cambios de modo. El interés es motivado principalmente por tres aspectos: (1) el análisis de regresión es la herramienta estadística más usada en la práctica para investigar relaciones causales entre va-

## 4.1. Introducción

---

riables; (2) las técnicas clásicas de análisis de regresión de series de tiempo son apropiadas para series estacionarias; sin embargo, en la práctica, es común enfrentarse con series no estacionarias. Una solución frecuente es hacer transformaciones a los datos para obtener una serie estacionaria. El costo asumido por esto es la pérdida de información sobre el proceso que origina la serie, y posiblemente la dificultad para interpretar los resultados; (3) los modelos de regresión con *change points* (llamados también *regímenes*), esto es, en los que la relación entre la variable respuesta y las variables explicativas es lineal por partes (por ejemplo la Figura 4.1), son relevantes en teoría y en práctica. Se pueden encontrar en la literatura diversas propuestas de métodos de estimación y detección de *change points* (Chen et al., 2011 discuten varios de ellos), pero la mayoría asume conocido el número de *change points*.

En estudios prácticos, el modelo propuesto permitirá tratar tres objetivos de interés: (1) identificar los cambios de modo en la serie; (2) estimar el vector de estados, esto es, los coeficiente de pendiente en análisis de regresión que cuantifican el efecto que las covariables asociadas tienen sobre las observaciones; (3) el ajuste de la serie de tiempo que se deriva de (1) y (2).

El capítulo está organizado de la siguiente manera: el apartado 4.2 introduce al **SLDS**, una generalización de los **HMM** que admiten espacios de estados continuos, y en el que cada estado está asociado con un proceso dinámico lineal. La sección 4.3 define el **HDP** de la sección 2.5.1 como a priori para el **HMM**, lo que permite considerar un número infinito de estados. Sin embargo, con la a priori **HDP** los estados tienen distribuciones de transición similares, por lo que un estado puede transitar a cualquier otro con frecuencia, y siempre existe la posibilidad de transitar a uno nuevo. Esto puede ocasionar grupos redundantes y muchos grupos con poca información de datos. Una alternativa a este problema es introducir un parámetro que incrementa la probabilidad esperada de permanecer en un mismo estado. El modelo, que se conoce como sticky **HDP-HMM**, es útil para muchas aplicaciones en las que es razonable asumir persistencia temporal en los estados. La sección 4.4 generaliza el **SLDS** para permitir describir una serie de datos mediante una relación funcional con una o más covariables, y hacer la ecuación de observaciones explícitamente dependiente de los modos. La propuesta es una extensión del trabajo de Fox et al. (2011a). Al menos dos ventajas se distinguen en un modelo de regresión dinámico no paramétrico Bayesiano con respecto a un método clásico: (1) permite que el efecto del cambio de una variable sobre otra evolucione con el tiempo; (2) permite inferir sobre el número de modos dinámicos. El desempeño del modelo se examina mediante simulación, y se ilustra con dos estudios de caso: el tipo de cambio en México de 1970 a 2016, y los niveles de ozono en dos estaciones de monitoreo de la Ciudad de México.

## 4.2. Modelos de Markov ocultos y sistemas dinámicos lineales de cambio de régimen

---

En análisis de regresión clásico de series de tiempo se requiere que las variables sean estacionarias para que se mantengan los resultados de los estimadores de mínimos cuadrados ordinarios. Con series no estacionarias es práctica común calcular  $d$  diferencias entre observaciones consecutivas para obtener una serie estacionaria de orden  $d$ . Sin embargo, si las variables son no cointegradas pueden producir una regresión espuria. Probar la hipótesis de que hay una relación estadísticamente significativa entre variables es equivalente a probar cointegración entre las series. Regresión espuria y cointegración son descritos en la sección 4.5. La sección concluye con un modelo de regresión lineal dinámica para explorar los efectos de la inflación y el crédito privado en el crecimiento económico de México. La sección 4.6 presenta las conclusiones finales del capítulo.

## 4.2. Modelos de Markov ocultos y sistemas dinámicos lineales de cambio de régimen

En muchas aplicaciones es deseable poder predecir un valor en una serie de tiempo dado un conjunto de observaciones previas. Intuitivamente, se espera que sea más probable que las observaciones más recientes proporcionen mayor información en predecir el valor futuro que aquellas más lejanas en el tiempo. Esto nos lleva a considerar los modelos de Markov, en los que se asume que la predicción de un valor futuro es independiente de todas las observaciones previas, excepto de la más reciente.

Como se señaló en el capítulo 3, un modelo de Markov es un modelo estocástico usado para modelar sistemas que cambian estocásticamente, en el que se asume que un estado en el tiempo  $t + 1$  depende sólo del estado en el tiempo  $t$  y no de lo ocurrido en cualquier otro tiempo previo. Extensiones de los modelos de Markov se obtienen introduciendo variables latentes, por ejemplo, los modelos de Markov ocultos (HMM), en los que las variables latentes son discretas, y modelos dinámicos lineales, en los que las variables latentes son, generalmente, Gaussianas.

En los HMM estándar, el espacio de estados de la variable oculta es discreta, mientras que las observaciones pueden ser discretas o continuas. Sin embargo, los HMM pueden ser generalizados para permitir espacios de estado continuos; por ejemplo, los modelos en donde el proceso de Markov asociado a las variables ocultas es un sistema dinámico lineal (LDS).

Los LDS son útiles para describir fenómenos dinámicos como series de tiempo financieras (Kim, 1994; Carvalho & Lopes, 2007; West, 2013; McAlinn & West,

## 4.2. Modelos de Markov ocultos y sistemas dinámicos lineales de cambio de régimen

2016), movimiento humano (Bregler, 1997; Pavlović et al., 2001) y riesgos medioambientales (Conrad Lamon III et al., 1998; Huerta et al., 2004; Velasco Cruz et al., 2012). Sin embargo, muchos fenómenos dinámicos complejos no son adecuadamente descritos por un solo modelo dinámico lineal, pero se pueden aproximar mediante *cambios* de modelos, en diferentes periodos, de un conjunto de sistemas lineales. Estos cambios se especifican de manera probabilística con base en una variable latente de espacio discreto. Cuando la variable latente es un proceso de Markov de tiempo discreto, el modelo se conoce como sistema dinámico lineal de cambio de régimen (SLDS).

Un SLDS se puede ver como una extensión de los HMM en el que cada estado está asociado con un proceso dinámico lineal. Es decir, dada la secuencia de estados, los HMM tienen observaciones condicionalmente independientes, mientras que un SLDS tiene un modelo dinámico lineal (ver Figura 4.2), lo que permite capturar dependencias temporales más complejas. De manera formal, un SLDS se puede describir usando el siguiente conjunto de ecuaciones de acuerdo con el modelo de Fox et al. (2011a):

$$\begin{aligned} z_t | z_{t-1} &\sim \pi_{z_t-1} \\ \boldsymbol{\beta}_t &= A^{(z_t)} \boldsymbol{\beta}_{t-1} + \mathbf{e}_t^{(z_t)} \\ y_t &= C \boldsymbol{\beta}_t + w_t \end{aligned} \tag{4.1}$$

donde  $\boldsymbol{\beta}_t \in \mathbb{R}^P$  denota el estado oculto del LDS,  $\mathbf{e}_t$  es el ruido del proceso,  $y_t$  es la variable respuesta, y  $w_t$  es el ruido de observación. Se asume que  $\mathbf{e}_t^{(z_t)} \sim N(0, \Sigma^{(z_t)})$  y  $w_t \sim N(0, R)$ . En este trabajo se considera a  $y_t$  un escalar, pero puede ser también un vector de observaciones, y  $w_t$  una matriz de covarianzas. Los componentes  $A$  y  $C$  son los típicos parámetros del LDS, como en el capítulo anterior: la matriz de evolución (o de transición) y la matriz diseño, respectivamente. El HMM es un proceso de Markov discreto de primer orden, con variables de estado  $z_t$ , llamadas *modos*<sup>1</sup>, de un conjunto de  $K$  modos, y distribuciones de transición  $\pi_j$ ,  $j = 1, \dots, K$ . El LDS y el HMM se relacionan mediante la dependencia que los parámetros  $A$  y  $\Sigma$  tienen del modo  $z_t$ . Aunque se ha asumido que el ruido de observación y la matriz  $C$  no dependen de  $z_t$ , este supuesto se puede modificar para permitir dependencia. La sección 4.4 extiende el SLDS para permitir dependencia de  $w_t$  de  $z_t$ . El capítulo 5 extiende la dependencia de las observaciones sobre  $z_t$  a través de la matriz diseño.

---

<sup>1</sup>Para diferenciar entre los estados ocultos  $z_t$  y  $\boldsymbol{\beta}_t$ , los estados HMM que indexan a los parámetros dinámicos son referidos como *modos*.

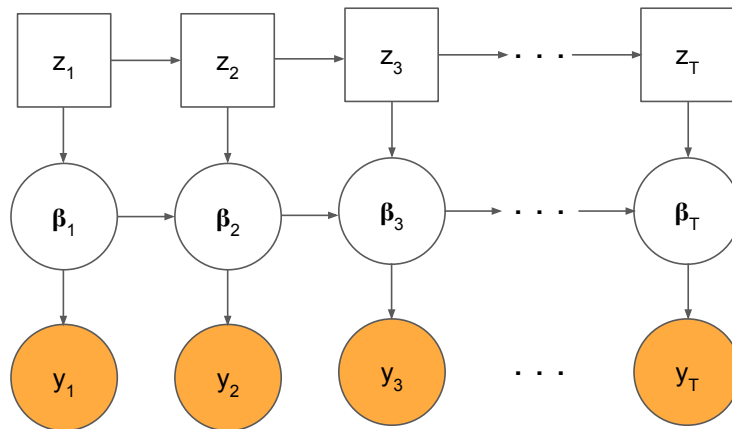


Figura 4.2: SLDS de T pasos en el tiempo.

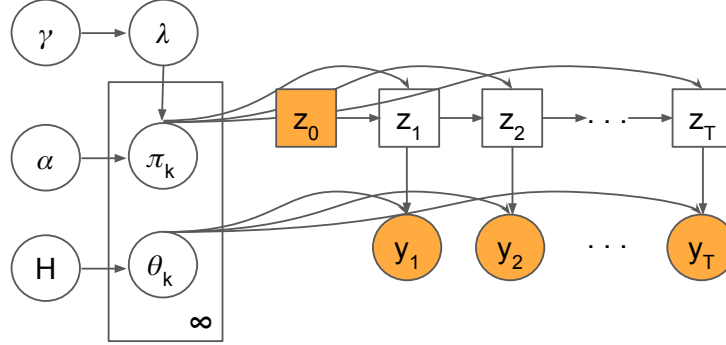
### 4.3. Modelos de Markov ocultos y procesos Dirichlet jerárquicos

La Figura 3.2 y el modelo (3.4)-(3.5) de la sección anterior representan a un HMM como una cadena de Markov en el que una secuencia de variables de estado discretas  $\{z_1, z_2, \dots, z_T\}$  está relacionada a través de una matriz de transición, y cada elemento  $y_t$  de una secuencia de observaciones  $\{y_1, y_2, \dots, y_T\}$  se toma independiente de las otras observaciones, condicionada a  $z_t$ . Esto es esencialmente un modelo de mezclas finito, en el que cada componente de mezcla corresponde a un valor del estado. Sin embargo, un HMM involucra a un conjunto de modelos de mezclas, uno para cada posible valor del estado. Esto es, el estado actual  $z_t$  indexa una fila específica de la matriz de transición. Las probabilidades de esa fila sirven como proporciones de la mezcla para la elección del siguiente estado  $z_{t+1}$ . Dado  $z_{t+1}$ , la observación  $y_{t+1}$  se toma del componente de mezcla indexado por  $z_{t+1}$ . Teh et al. (2006) proponen una variante no paramétrica del HMM, definiendo el HDP como a priori para el HMM. Esto permite considerar un conjunto infinito de estados, es decir, de modelos de mezclas, y garantiza que el soporte del DP en cada fila de la matriz de transición sea el mismo. El modelo que resulta se conoce como proceso Dirichlet jerárquico para modelos de Markov ocultos (HDP-HMM).

La Figura 4.3 ilustra la representación que Teh et al. (2006) hacen del HDP-HMM usando *stick-breaking*. La variable  $z$  denota los estados discretos del sistema, distribuidos de acuerdo con  $\pi$ . Por sí misma,  $\pi$  se distribuye como un DP con parámetro de concentración  $\alpha$  y media  $\lambda$ . La distribución base  $\lambda$  es un proceso *stick-breaking* con parámetro  $\gamma$ . Las observaciones, denotadas por  $y_t, t = 1, \dots, T$ , son condicionalmente independientes dado  $z_t$  y  $\theta_k$ , donde  $\theta_k$  tiene distribución  $H$ .

### 4.3. Modelos de Markov ocultos y procesos Dirichlet jerárquicos

El número de estados es infinito, y hay un parámetro  $\theta$  asociado a cada estado,  $k = 1, 2, \dots$ . Esta descripción es formalizada en el modelo jerárquico (4.2).



**Figura 4.3:** Representación gráfica de un HDP-HMM.

$$\begin{aligned}
 G_0 &= \sum_{k=1}^{\infty} \lambda_k \delta_{\theta_k} & \lambda | \gamma &\sim \text{GEM}(\gamma) \\
 G_j &= \sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta_k} & \pi_j | \alpha, \lambda &\sim \text{DP}(\alpha, \lambda) \\
 \theta_k | H &\sim H & & (4.2) \\
 z_t | z_{t-1}, \{\pi_k\}_{k=1}^{\infty} &\sim \pi_{z_{t-1}} \\
 y_t | z_t, \{\theta_k\}_{k=1}^{\infty} &\sim F(\theta_{z_t})
 \end{aligned}$$

donde  $k = 1, 2, \dots$ ;  $t = 1, \dots, T$ ;  $\pi_j = [\pi_{j1} \ \pi_{j2} \ \dots]$ , y  $\text{GEM}(\cdot)$  hace referencia al proceso *stick-breaking*. El HDP define una colección de medidas de probabilidad  $\{G_j\}$  sobre el mismo soporte  $\{\theta_1, \theta_2, \dots, \dots\}$ , asumiendo que cada medida discreta  $G_j$  es una variación de una medida discreta global  $G_0$ . Específicamente, el modelo jerárquico Bayesiano toma  $G_j \sim \text{DP}(\alpha, G_0)$ , con  $G_0$  muestreada de  $\text{DP}(\gamma, H)$  (Fox et al., 2011a). Los HDP-HMM han mostrado utilidad en aplicaciones como *speech recognition* (Teh et al., 2006), genética (Sohn & Xing, 2007) y música (Hoffman et al., 2008).

En el modelo (4.2),  $\lambda$  se toma de un proceso *stick-breaking* con parámetro de concentración  $\gamma$ ;  $\lambda$  es la distribución base del DP que produce a  $\pi_k$  para todos

### 4.3. Modelos de Markov ocultos y procesos Dirichlet jerárquicos

---

los estados (o modos) ocultos ( $\lambda$  se comparte entre todos los estados);  $\pi_k$  es una distribución multinomial correspondiente a cada estado oculto. Las distribuciones multinomiales corresponden a una matriz de probabilidades de transición en un HMM. Esto significa que el HDP-HMM tiene matrices con dimensiones potencialmente infinitas. Un estado oculto puede transitar a cualquier otro estado estocásticamente, pero debido a que  $\pi_k$  no tiene un parámetro de auto transición, es decir, de permanencia en un mismo estado, las transiciones a otros estados ocurren frecuentemente. Dicho de otro modo, al definir  $\pi_j \sim \text{DP}(\alpha, \lambda)$ , la a priori HDP permite que los estados tengan distribuciones de transición similares. Siguiendo las propiedades del DP, estas distribuciones de transición son idénticas en esperanza:

$$E(\pi_{jk}|\lambda) = \lambda_k \quad (4.3)$$

La Ec. (4.3) precisa que la media de transición desde cualquier estado al  $k$ -ésimo estado es la misma, por lo que varios grupos pueden tener similares parámetros de emisión (grupos redundantes) y puede haber muchos grupos con poca información de datos (el HDP-HMM permite considerar un número infinito de estados).

**Ejemplo:** Se desea modelar con HDP-HMM valores diarios de un índice bursátil. Los valores diarios son las observaciones  $y_t$ ; los estados  $z_t$  identifican condiciones de mercado que determinan propiedades, por ejemplo la volatilidad del mercado; estas propiedades son los valores de  $\theta$ , de manera que  $F(\theta_{z_t})$  determina el comportamiento de  $y_t$ . Dado el valor  $z_t = j$ , las probabilidades  $\pi_j$  son las probabilidades de transición para la elección del siguiente estado,  $z_{t+1}$ , es decir, las condiciones de mercado en el tiempo  $t + 1$  dependen de las condiciones prevalecientes en el tiempo  $t$ . El modelo establece dos hechos: (1) como las probabilidades se distribuyen DP, siempre existe la posibilidad de transitar a un nuevo estado; (2) la media de la distribución de transición de un estado  $j$  a otro cualquiera no depende del estado  $j$ , y no considera la permanencia en un estado ( $E[\pi_{jk}|\lambda] = \lambda_k$  y  $E[\pi_{kk}|\lambda] = \lambda_k$ ). Sin embargo, para el índice bursátil es razonable asumir que las condiciones de mercado no cambian a diario, por lo que este modelo no sería apropiado.

Una alternativa de solución a los problemas que el HDP-HMM presenta en muchas aplicaciones relacionadas a la del ejemplo anterior es la extensión conocida como *sticky HDP-HMM* (Fox et al., 2011b). La idea básica es introducir un parámetro que incrementa la probabilidad esperada de permanecer en el mismo estado, o auto transición (*self-transition*), y colocar una distribución a priori para ese parámetro.



#### 4.3.1. *Sticky* HDP-HMM

El sticky HDP-HMM fue propuesto por Fox et al. (2011b) en una aplicación de *speaker diarization*, donde las grabaciones de audio se segmentan en intervalos de tiempo asociados con *speakers* individuales y se identifica el número de *speakers*. Un aspecto clave de *speaker diarization* es que generalmente un individuo repite su participación (*speech*) en múltiples intervalos de tiempo disjuntos. Resulta entonces natural en esta aplicación considerar los estados ocultos como correspondientes a los *speakers*. Adicionalmente, el objetivo es identificar el número de *speakers* y la transición entre ellos, y el HDP-HMM tiende a cambiar rápidamente entre estados redundantes. Al incorporar un parámetro de auto transición es posible modelar la persistencia temporal de estados. Otros trabajos también han propuesto parámetros auto transición para HMM con espacios de estado infinito (Beal et al., 2002; Sohn & Xing, 2007). La descripción presentada aquí es la propuesta por Fox et al. (2011b).

Como se mencionó en la sección anterior, al definir la distribución de transición como un DP ( $\pi_j \sim \text{DP}(\alpha, \lambda)$ ), la a priori HDP permite que los estados tengan distribuciones de transición similares (Ec. (4.3)). Sin embargo, en muchas aplicaciones es razonable suponer que los cambios dinámicos ocurren lentamente y que hay una alta probabilidad de persistencia en los estados. Para tratar con los inconvenientes del HDP-HMM, Fox et al. (2011b) propusieron muestrear  $\pi_j$  como:

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

donde  $\text{GEM}(\gamma)$  es un proceso *stick-breaking*;  $\delta_j = 1$  para el parámetro correspondiente a la auto transición, y  $\delta_j = 0$  de otro modo;  $(\alpha\lambda + \kappa\delta_j)$  indica que una cantidad  $\kappa > 0$  se agrega al  $j$ -ésimo componente de  $\alpha\lambda$ . Esto incrementa la probabilidad esperada de una transición hacia el mismo modo en una cantidad proporcional a  $\kappa$ . Es decir,

$$E(\pi_{jk} | \lambda, \kappa) = \frac{\alpha}{\alpha + \kappa} \lambda_k + \frac{\kappa}{\alpha + \kappa} \delta(j, k),$$

donde  $\delta(j, k)$  denota la delta de Kronecker. Debido a que el valor  $\kappa$  incrementa  $E(\pi_{jj} | \lambda)$ , este modelo se denomina *sticky* HDP-HMM. Cuando  $\kappa = 0$  se regresa al original HDP-HMM de Teh et al. (2006).

El sticky HDP-HMM extiende la metáfora del CRF de Teh et al. (2006) a res-

### 4.3. Modelos de Markov ocultos y procesos Dirichlet jerárquicos

---

taurantes con clientes *leales* para ilustrar el proceso. La representación se conoce como franquicia de restaurant Chino con clientes leales, y puede consultarse ampliamente en [Fox et al. \(2011b\)](#).

#### 4.3.2. HDP-SLDS

La sección 4.1 describe un [SLDS](#) mediante un modelo jerárquico de cuatro niveles: las observaciones, descritas mediante una relación lineal con una variable de estado oculta a través de la ecuación de observación; el proceso dinámico o de evolución de los estados, descrito a través de la ecuación de evolución; los parámetros que definen al modelo; y la variable latente que indexa a dichos parámetros para especificar cuál modelo, o modo, del conjunto de [SLDS](#) será usado, descrita mediante una dinámica de transición que satisface la propiedad de Markov. La inferencia en un modelo [SLDS](#) implica determinar la distribución a posteriori de los estados dinámicos y modos [HMM](#) ocultos dada la secuencia de observaciones; es decir, el objetivo es encontrar  $p(\beta_{1:T}, z_{1:T} | y_{1:T})$ . Si el modelo no tuviera cambios de modo (*switch*), la inferencia sobre  $\beta_{1:T}$  dado  $y_{1:T}$  usando [LDS](#) sería analíticamente fácil y computacionalmente manejable. Sin embargo, la presencia de la variable latente  $z_t$  hace complicada la inferencia. Aunque la derivación algebraica de la distribución es fácil, el tamaño de las mezclas incrementa exponencialmente con  $t$ , lo que hace intratable la implementación numérica (ver [Pavlović et al., 2001](#); [Oh et al., 2008](#); [Barber, 2012](#)).

Se encuentran en la literatura muchos esfuerzos por derivar métodos de inferencia eficientes sobre los parámetros de un [SLDS](#) como el dado en (4.1); sin embargo, la mayoría asume que el número de modos es conocido ([Pavlović et al., 1999, 2001](#); [Oh et al., 2008](#)) o bien se modela la serie de tiempo con el objetivo de detectar el número y localización de los cambios de modo con técnicas conocidas como *segmentation* o *change point detection*, en las que cada segmento o cambio de modo se considera como un nuevo modo nunca antes visitado ([Fearnhead, 2005, 2006](#); [Xuan & Murphy, 2007](#)). Un planteamiento más robusto para aprender de un modelo [SLDS](#) es el de [Fox et al. \(2011a\)](#); los autores proponen una aproximación no paramétrica Bayesiana que permite inferir sobre el número de modos dinámicos y sobre los componentes del vector de estados, al tiempo que permite regresar a comportamientos dinámicos (modos) anteriormente observados. La propuesta extiende la formulación del [HDP-HMM](#) de [Teh et al. \(2006\)](#) para [SLDS](#), considerando un *sticky* en la distribución de transición de los modos que incrementa la probabilidad esperada de permanencia en un modo en una cantidad proporcional al sticky (ver Sección 4.2.1). El sticky [HDP-HMM](#) se usa como a priori del proceso de Markov  $\{z_t\}$ . El modelo que resulta se conoce como [HDP-SLDS](#). Las Ec. (4.4a)

### 4.3. Modelos de Markov ocultos y procesos Dirichlet jerárquicos

y (4.4b) describen el proceso, y la Figura 4.4 ilustra la representación.

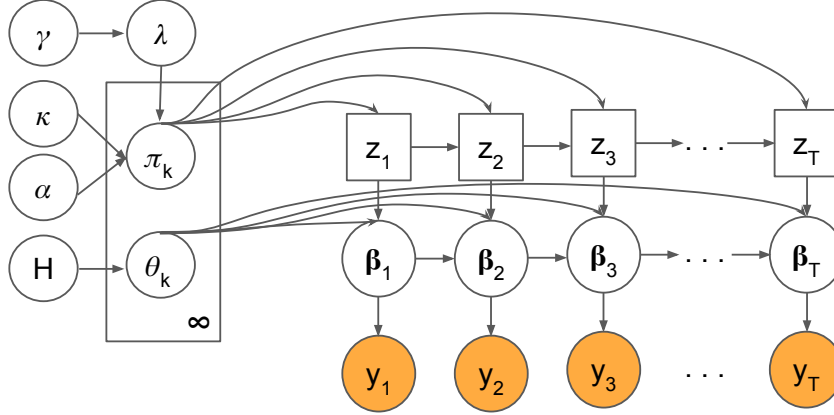


Figura 4.4: Representación gráfica de un HDP-SLDS.

$$G_0 = \sum_{k=1}^{\infty} \lambda_k \delta_{\theta_k} \quad \lambda | \gamma \sim \text{GEM}(\gamma)$$

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta_k} \quad \pi_j | \alpha, \kappa, \lambda \sim \text{DP}\left(\alpha + \kappa, \frac{\alpha \lambda + \kappa \delta_j}{\alpha + \kappa}\right)$$

$$\theta_k | H \sim H \quad (4.4a)$$

$$z_t | z_{t-1}, \{\pi_k\}_{k=1}^{\infty} \sim \pi_{z_{t-1}}$$

$$\beta_t = A^{(z_t)} \beta_{t-1} + \mathbf{e}_t^{(z_t)}$$

$$y_t = C \beta_t + w_t \quad (4.4b)$$

donde  $k = 1, 2, \dots$ ;  $t = 1, \dots, T$ ; y  $\pi_j = [\pi_{j1} \ \pi_{j2} \ \dots]$  es generada de un DP con parámetro de concentración  $\alpha + \kappa$ ;  $\kappa$  es un parámetro de auto transición, que incrementa la probabilidad esperada de transición de un estado hacia sí mismo y  $H$  es una medida base apropiada para los parámetros dinámicos del modelo  $\theta_k = \{A^{(k)}, \Sigma^{(k)}\}$ . Los parámetros  $\gamma$ ,  $\lambda$  y las medidas  $G_0$  y  $G_j$  son las mismas de la Ec. (4.2). El modelo (4.4b) es el SLDS (4.1) y se puede extender para que  $C$  y  $w_t$  estén indexados por  $z_t$ .

Para inferencias basadas en el modelo (4.4a)-(4.4b), Fox et al. (2011a) propusieron un algoritmo del tipo muestreo Gibbs por bloques (ver Ishwaran & James, 2001) que acelera el proceso de muestreo. El algoritmo itera entre el muestreo de la secuencia de estados  $\beta_{1:T}$  y, condicionados a  $\beta_{1:T}$ , el muestreo de la secuencia

de modos  $z_{1:T}$ , del conjunto de parámetros dinámicos y de parámetros del sticky HDP-HMM.

## 4.4. Regresión con SLDS

En muchas aplicaciones es de interés describir el comportamiento de una serie de datos a través de una relación funcional con una o más series de variables observadas. Por ejemplo, podría ser de interés el efecto de la temperatura del aire en la temperatura del agua (Velasco Cruz et al., 2012); el efecto de la inflación y el desarrollo financiero sobre el producto interno bruto (Tinoco Zermeño et al., 2014); el efecto de las intervenciones del banco central en la volatilidad del tipo de cambio (Hung, 1997; Huang, 2007; Mondal, 2013) o; el efecto de la temperatura en la concentración de ozono (Huerta et al., 2004). En este contexto, los modelos de regresión representan una importante herramienta estadística para el estudio de relaciones causales de una variable, comúnmente llamada regresor o variable explicativa, sobre la variable de respuesta o dependiente.

En la práctica, la construcción de modelos es guiada por ciertos objetivos específicos al problema bajo estudio. Para algunos propósitos un modelo estático podría ser satisfactorio, pero podría ser inadecuado para describir relaciones cambiantes entre las variables conforme el tiempo transcurre. Los modelos dinámicos agregan flexibilidad en la modelación, ya que permiten la posibilidad de que el efecto de una variable sobre otra evolucione con el tiempo (ver Cap. 3).

En el modelo (4.4a)-(4.4b) una serie de tiempo  $y_{1:T}$  (o múltiples si  $y_t$  es un vector) se modela mediante una relación lineal con interceptos aleatorios. Es decir,

$$y_t = C' \beta_t + w_t, \quad (4.5)$$

donde el vector de estados  $\beta_t$  representa la evolución del efecto promedio que cada componente de  $C$  tiene sobre  $y_t$  a través del tiempo, y la matriz de emisión de las observaciones,  $C$ , se supone fija para todos los modos. Aunque se ha señalado que este supuesto puede ser modificado para permitir que  $C$  sea específica para cada modo,  $C^{(z_t)}$ , Fox et al. (2011a) argumentan que tal elección no siempre es necesaria o apropiada para ciertas aplicaciones, y que puede tener implicaciones sobre la identificabilidad del modelo. En el capítulo 5 se propone una manera de hacer la ecuación de observaciones específica a cada modo, pero en un contexto de regresión. En esta sección, el interés se centra en establecer a la ecuación (4.5) como un modelo de regresión en donde  $C$  se reemplaza por una matriz diseño indexada

#### 4.4. Regresión con SLDS

por  $t$ , compuesta de covariables asociadas a  $y_t$ . Adicionalmente, el SLDS de Fox et al. (2011a) se extiende para permitir que el ruido de medición sea específico a cada modo. Es decir, el modelo de regresión HDP-HMM se define como:

$$\begin{aligned} z_t | z_{t-1}, \{\pi_k\}_{k=1}^\infty &\sim \pi_{z_{t-1}} \\ \boldsymbol{\beta}_t &= A^{(z_t)} \boldsymbol{\beta}_{t-1} + \mathbf{e}_t^{(z_t)} \\ y_t &= X_t' \boldsymbol{\beta}_t + w_t^{(z_t)}, \end{aligned} \tag{4.6}$$

donde  $z_t$ ,  $\pi_k$ ,  $\boldsymbol{\beta}_t$ ,  $A^{(k)}$  y  $\mathbf{e}^{(k)}$  son como antes,  $w_t^{(k)} \sim N(0, R^{(k)})$ , y  $X_t' = \{1, x_{t1}, \dots, x_{t,p-1}\}$  es un vector de covariables.

Para establecer claramente las implicaciones de introducir  $X_t$  en el modelo, es necesario especificar los supuestos en la elección de la matriz  $C$  del modelo (4.4a)-(4.4b). Debido a que los componentes de  $C$  están asociados a los componentes del vector de estados  $\boldsymbol{\beta}_t$ , y a su vez la dependencia funcional entre los componentes de  $\boldsymbol{\beta}_t$  está determinada por los parámetros dinámicos  $\{A^{(k)}, \Sigma^{(k)}\}$ , la especificación de  $C$  está estrechamente relacionada con la elección de la distribución a priori sobre estos parámetros. Fox et al. (2011a) proponen dos alternativas: *automatic relevance determination* (ARD) para la matriz dinámica  $A^{(z_t)}$ , y una distribución a priori conjugada para la verosimilitud del modelo, la matriz-normal<sup>2</sup> inversa-Wishart (MNIW) para el conjunto  $\{A^{(k)}, \Sigma^{(k)}\}$ . El uso de la a priori ARD permite identificar componentes del vector de estados que son irrelevantes para el modelo, haciendo tender a cero la columna de la matriz  $A^{(k)}$  correspondiente al componente que no contribuye en la ecuación de estados. Esto implica que  $C$  debe ser elegida de manera que si un modo  $k$  tiene una columna de ceros en  $A^{(k)}$ , entonces la realización debe tener el  $j$ -ésimo componente de  $C$  (o columna, si  $C$  es matriz) igual a cero. De otro modo, se puede considerar un modelo más general donde la ecuación de las observaciones sea específica a cada modo, y colocar una a priori sobre  $C^{(k)}$  en lugar de ser fijo, pero este problema no es abordado en el trabajo de Fox et al. (2011a). En el capítulo 5 se propone un método de selección de variables en cada modo para el modelo de regresión HDP-SLDS, lo que puede ser visto como una generalización a la restricción impuesta a  $C$ . La generalización implica hacer cero los componentes del vector de estados asociados a las covariables que no sean significativas a las realizaciones de ese modo. Aunque en esta tesis se

<sup>2</sup>La distribución matriz-normal es una generalización de la distribución normal multivariada. Se suele denotar como  $MN_{n,p}(M, V, U)$ , y está relacionada a la distribución normal multivariada de la siguiente manera:

$$\mathbf{A} \sim MN_{n \times p}(M, V, U)$$

si y sólo si

$$\text{vec}(\mathbf{A}) \sim N_{np}(\text{vec}(M), U \otimes V)$$

donde  $\text{vec}(M)$  denota la vectorización de  $M$  y  $\otimes$  el producto Kronecker.

## 4.4. Regresión con SLDS

---

usa solamente ARD, más detalles sobre las implicaciones de las dos alternativas a priori, y las derivaciones de las distribuciones condicionales para el muestreo, se dan en el Anexo B como referencia. Adicionalmente, en el mismo Anexo B se presenta la derivación de la distribución condicional para el muestreo de la varianza del error de medición. El resto de las distribuciones condicionales para hacer inferencia sobre el modelo de regresión (4.6) se presentan en la siguiente sección. Las derivaciones se basan en los resultados presentados por Fox et al. (2011a), sin embargo, la notación difiere ligeramente.

### 4.4.1. Muestreo Gibbs para el modelo de regresión HDP-SLDS

Para hacer inferencia sobre el modelo de regresión HDP-SLDS (4.6) el muestreo Gibbs itera entre: (1) el muestreo de la secuencia de estados  $\beta_{1:T}$ , condicionados a las secuencias de modos y observaciones; (2) el muestreo de la secuencia de modos, condicionados a las observaciones; (3) el muestreo del conjunto de parámetros dinámicos, y de los parámetros del sticky HDP-HMM. Adicionalmente, se incluye en la iteración el muestreo de la secuencia de modos  $z_{1:T}$  marginalizando sobre la secuencia de estados para mejorar la convergencia.

#### Muestreo de los parámetros dinámicos

Dada la secuencia de estados  $\beta_{1:T}$ , de modos  $z_{1:T}$ , y un conjunto de valores iniciales  $\{A^{(k)}, \Sigma^{(k)}, \alpha^{(k)}\}$ , la actualización de los parámetros dinámicos  $\{A^{(k)}, \Sigma^{(k)}, R^{(k)}\}$  usando la distribución a priori ARD sigue los pasos del Algoritmo 1 (ver Anexo B para las derivaciones).

Algoritmo 1. Actualización de parámetros dinámicos con a priori ARD

Asumiendo distribuciones Gaussianas independientes sobre las columnas de la matriz  $A^{(k)}$ , tal que  $p(A^{(k)}|\boldsymbol{\alpha}^{(k)}) = \prod_{j=1}^p N(\mathbf{0}, 1/\alpha_j^{(k)} I_p)$ . Para cada  $k \in \{1, \dots, K\}$ :

1. Construir  $\tilde{\mathbf{B}}_t$  tal que

$$\begin{aligned}\boldsymbol{\beta}_t &= [\beta_{t-1,1} I_p \quad \beta_{t-1,2} I_p \quad \cdots \quad \beta_{t-1,p} I_p] \text{vec}(A^{(k)}) + \mathbf{e}_t^{(k)} \quad \forall t | z_t = k \\ &= \tilde{\mathbf{B}}_{t-1} \text{vec}(A^{(k)}) + \mathbf{e}_t^{(k)}\end{aligned}$$

2. Construir  $\Sigma_0^{(k)} = \text{diag}(\alpha_1^{(k)}, \dots, \alpha_1^{(k)}, \dots, \alpha_p^{(k)}, \dots, \alpha_p^{(k)})^{-1}$  y actualizar la matriz dinámica de la siguiente manera:

$$\begin{aligned}\text{vec}(A^{(k)}) | \Sigma^{(k)}, \boldsymbol{\alpha}^{(k)} &\sim N(\boldsymbol{\mu}, \Lambda) \\ \boldsymbol{\mu} &= \Lambda^{-1} \sum_{t|z_t=k} \tilde{\mathbf{B}}_{t-1}^T (\Sigma^{(k)})^{-1} \boldsymbol{\beta}_t \\ \Lambda &= \left[ (\Sigma_0^{(k)})^{-1} + \sum_{t|z_t=k} \tilde{\mathbf{B}}_{t-1}^T (\Sigma^{(k)})^{-1} \tilde{\mathbf{B}}_{t-1} \right]^{-1}\end{aligned}$$

3. Para cada  $l \in \{1, \dots, p\}$ , actualizar el parámetro de precisión:

$$\alpha_l^{(k)} | A^{(k)} \sim \text{Gam} \left( a + p/2, b + \sum_{i=1}^p (a_{il}^{(k)})^2 / 2 \right)$$

4. Calcular

$$\begin{aligned}n_k &= |\{t | z_t = k, t = 1, \dots, T\}| \\ S_{\beta|\beta_{-1}} &= \sum_{t|z_t=k} (\boldsymbol{\beta}_t - A^{(k)} \boldsymbol{\beta}_{t-1}) (\boldsymbol{\beta}_t - A^{(k)} \boldsymbol{\beta}_{t-1})^T \\ S_{R^{(k)}} &= \frac{1}{2} \sum_{t:z_t=k} (y_t - X_t' \boldsymbol{\beta}_t)^2\end{aligned}$$

- a) Actualizar la matriz de covarianzas

$$\Sigma^{(k)} | A^{(k)} \sim IW \left( n_0 + n_k, S_0 + S_{\beta|\beta_{-1}} \right)$$

- b) Actualizar la varianza del ruido de medición

$$1/R^{(k)} \sim \text{Gam} \left( \frac{n_k}{2} + a_r, b_r + S_{R^{(k)}} \right)$$

## 4.4. Regresión con SLDS

---

### Block sampling $z_{1:T}$

Para aproximar la distribución  $p(z_t|v_{1:T})$  de los estados ocultos en los modelos de Markov ocultos (HMM), se utiliza comúnmente *parallel smoothing* (Barber, 2012). Este método separa la distribución en dos componentes, uno que representa la contribución de la información pasada y presente, y otro la contribución del futuro para inferir el pasado:

$$p(z_t, v_{1:T}) = p(z_t, v_{1:t}, v_{t+1:T}) = p(z_t, v_{1:t})p(v_{t+1:T}|z_t) = \alpha(z_t)\beta(z_t), \quad (4.7)$$

donde  $z_t$  representa la variable oculta o latente al tiempo  $t$ , y  $v_{1:T}$  las variables visibles u observables. El término  $\alpha(z_t)$  hace inferencia sobre el presente mediante un mecanismo *forward* en tiempo, conocido como *filtering*. El término  $\beta(z_t)$  toma en cuenta la evidencia futura para mejorar la estimación del pasado, mediante un mecanismo *backward* en tiempo, conocido como suavizado. El forward y el backward son independientes y pueden ser ejecutados en forma paralela para obtener la a posteriori deseada (ver sección 3.2.1).

El algoritmo *forward-backward* explota la estructura de independencia de los HMM, que permite expresar la distribución conjunta de las variables latentes y las variables observadas como:

$$p(z_{1:T}, v_{1:T}) = p(v_1|z_1)p(z_1) \prod_{t=2}^T p(v_t|z_t)p(z_t|z_{t-1}),$$

donde  $p(v_t|z_t)$  modela el proceso de generación de las variables observadas, y  $p(z_t|z_{t-1})$  modela la transición de las variables ocultas. Dada la secuencia  $z_{1:T}$ ,  $v_t$  es condicionalmente independiente de las observaciones pasadas y futuras.

Los SLDS son una extensión de los HMM, y es posible derivar un procedimiento *forward-backward* para inferir sobre  $z_{1:T}$ . En los HMM se tienen observaciones condicionalmente independientes dada la secuencia de modos, mientras que en los SLDS cada modo está asociado a un proceso dinámico lineal con espacio de estados continuo. Sin embargo, los estados obedecen a una estructura Markoviana que, junto con la regla de la cadena, permite descomponer la distribución conjunta de la secuencia de modos en términos del producto de probabilidades condicionales, y separar las contribuciones del pasado y el futuro para inferir sobre  $z_{1:T}$ , similar a (4.7). Para el modelo (4.1) usado por Fox et al. (2011a), la distribución conjunta de la secuencia de modos, dados los estados  $\beta_{1:T}$ , las probabilidades de transición  $\pi$  y el conjunto de parámetros dinámicos  $\theta = \{\theta_k\}_{k=1}^K$ , donde  $\theta_k = \{A^{(k)}, \Sigma^{(k)}\}$ ,



#### 4.4. Regresión con SLDS

---

se puede expresar como:

$$\begin{aligned}
 p(z_{1:T}|\beta_{1:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) = & \\
 & p(z_T|z_{T-1}, \beta_{1:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) p(z_{T-1}|z_{T-2}, \beta_{1:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) \\
 & \cdots p(z_2|z_1, \beta_{1:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) p(z_1|\beta_{1:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) \quad (4.8)
 \end{aligned}$$

A su vez, cada término del producto en la expresión anterior se puede descomponer de manera equivalente a (4.7) y, por tanto, emplear un mecanismo *forward-backward* para muestrear de la distribución deseada. Para cada  $t$ , note que:

$$\begin{aligned}
 & p(z_t|z_{t-1}, \beta_{1:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) \\
 & \propto p(z_t|\pi_{z_{t-1}}) p(\beta_1|z_t, z_{t-1}, \boldsymbol{\pi}, \boldsymbol{\theta}) p(\beta_2|\beta_1, z_t, z_{t-1}, \boldsymbol{\pi}, \boldsymbol{\theta}) \cdots p(\beta_T|\beta_{T-1}, z_t, z_{t-1}, \boldsymbol{\pi}, \boldsymbol{\theta}) \\
 & \propto p(z_t|\pi_{z_{t-1}}) \prod_{i=1}^t p(\beta_i|\beta_{i-1}, z_t, z_i, \boldsymbol{\pi}, \boldsymbol{\theta}) \prod_{j=t+1}^T p(\beta_j|\beta_{j-1}, z_t, z_j, \boldsymbol{\pi}, \boldsymbol{\theta}) \\
 & \propto p(\beta_1, \dots, \beta_t, z_t|z_{t-1}, \boldsymbol{\pi}, \boldsymbol{\theta}) \cdot p(\beta_{t+1}, \dots, \beta_T|z_t, \boldsymbol{\pi}, \boldsymbol{\theta}) \quad (4.9)
 \end{aligned}$$

Para derivar el procedimiento y muestrear conjuntamente la secuencia de estados  $z_{1:T}$ , la expresión (4.8) sugiere muestrear primero  $z_1$ ; entonces, condicionado a ese valor, muestrear  $z_2$ , y así sucesivamente. En la práctica, es común considerar una aproximación truncada del HDP, lo que permite expresar el modelo completo en términos de un número finito de variables aleatorias, pero que no implica suponer que se conoce la cardinalidad del espacio de estados. El límite de truncamiento del HDP es heredado al mecanismo *forward-backward*, en el que representa el número posible de estados HMM, por lo que reduce la dimensionalidad del problema.

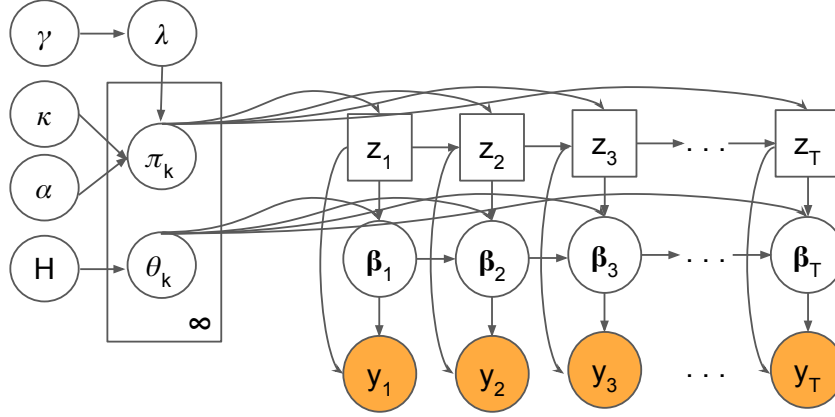
Un algoritmo eficiente que emplea un proceso *stick-breaking* truncado para aproximar el DP fue propuesto por Ishwaran & James (2001). Ishwaran & Zarepour (2002) aproximan el DP mediante una suma finita de pesos Dirichlet, el método se conoce como *finite-dimensional Dirichlet prior*. Cuestiones sobre determinar un nivel de truncamiento apropiado y evaluar la convergencia de la aproximación truncada a su distribución límite se pueden consultar en Ishwaran & James (2001), Ishwaran & Zarepour (2002) y, en el contexto de mezclas finitas de normales, en Ishwaran & James (2002). Por su simplicidad y eficiencia computacional Fox et al. (2011a) emplean *finite-dimensional Dirichlet prior* sobre  $\lambda$  del modelo (4.4a)-(4.4b), lo que induce también a una distribución a priori Dirichlet finita sobre  $\pi_j$ .

Una idea semejante se puede usar para inferir sobre  $z_{1:T}$  del modelo (4.6). Pero ahora el ruido de medición  $w_t^{(z_t)}$  también depende de los modos ocultos (ver Figura

#### 4.4. Regresión con SLDS

4.5), por lo que el modelo define una distribución conjunta dada por

$$p(z_{1:T}, \boldsymbol{\beta}_{1:T}, y_{1:T} | \boldsymbol{\pi}, \boldsymbol{\theta}). \quad (4.10)$$



**Figura 4.5:** Representación gráfica de un HDP-SLDS con observaciones dependientes de los modos.

La distribución conjunta (4.10) se puede descomponer de manera semejante a (4.8) como:

$$\begin{aligned} p(z_{1:T} | \boldsymbol{\beta}_{1:T}, y_{1:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) = \\ p(z_T | z_{T-1}, \boldsymbol{\beta}_{1:T}, y_{1:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) p(z_{T-1} | z_{T-2}, \boldsymbol{\beta}_{1:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) \\ \cdots p(z_2 | z_1, \boldsymbol{\beta}_{1:T}, y_{1:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) p(z_1 | \boldsymbol{\beta}_{1:T}, y_{1:T}, \boldsymbol{\pi}, \boldsymbol{\theta}). \end{aligned} \quad (4.11)$$

Sin embargo, tanto (4.11) como (4.8) dependen de la secuencia de estados ocultos  $\boldsymbol{\beta}_{1:T}$ . En muchos problemas hay fuerte dependencia entre las variables que determinan los modos y los estados, causando que el muestreo Gibbs converja lentamente o incluso que no alcance convergencia (Carter & Kohn, 1996). Inspirado en Carter & Kohn (1996), el algoritmo de Fox et al. (2011a) incluye un paso que consiste en el muestreo secuencial de los modos, sin condicionar a los estados, de la distribución:

$$p(z_t | z_{-t}, y_{1:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) \propto p(z_t = k | z_{-t}, \boldsymbol{\pi}) p(y_{1:T} | z_t = k, z_{-t}, \boldsymbol{\theta}), \quad (4.12)$$

donde  $z_{-t} = \{z_1, \dots, z_{t-1}, z_{t+1}, \dots, z_T\}$ . Es decir, el muestreo secuencial actualiza cada  $z_t$  individualmente dados  $z_{-t}, y_{1:T}, \boldsymbol{\pi}$  y  $\boldsymbol{\theta}$ . En cambio, el muestreo conjunto, que se conoce como muestreo Gibbs por bloques (*blocked Gibbs sampling*), actualiza toda la secuencia  $z_{1:T}$  de la distribución conjunta condicional dados  $y_{1:T}, \boldsymbol{\pi}, \boldsymbol{\theta}$

#### 4.4. Regresión con SLDS

y  $\beta_{1:T}$ . Debido a que el muestreo secuencial es condicionado a menos información porque se marginaliza  $\beta_{1:T}$ , es probable que alcance convergencia más rápido que el muestreo conjunto (Carter & Kohn, 1996). Sin embargo, el muestreo secuencial es computacionalmente intensivo. En esta tesis, se propone muestrear  $z_{1:T}$  conjuntamente de un modelo marginal en  $\beta_{1:T}$ . Específicamente, se usa la distribución predictiva de las observaciones, que implica integrar sobre la secuencia de estados. Haciendo uso de los resultados sobre la distribución normal multivariada, es fácil probar que la distribución predictiva es Gaussiana (ver Sec. 3.3.2), tal que es suficiente calcular sus momentos:

$$\begin{aligned} E(y_{t+1}|y_{1:t}, \boldsymbol{\theta}) &= E(E(y_{t+1}|\beta_{t+1}, \boldsymbol{\theta})|y_{1:t}) = X'_{t+1} E(\beta_{t+1}|y_{1:t}, \boldsymbol{\theta}) \\ &= X'_{t+1} \mathbf{f}_{t,t+1} \end{aligned} \quad (4.13a)$$

$$\begin{aligned} V(y_{t+1}|y_{1:t}, \boldsymbol{\theta}) &= V(E(y_{t+1}|\beta_{t+1}, \boldsymbol{\theta})|y_{1:t}) + E(V(y_{t+1}|\beta_{t+1}, \boldsymbol{\theta})|y_{1:t}) \\ &= X'_{t+1} V(\beta_{t+1}|y_{1:t}, \boldsymbol{\theta}) X_{t+1} + E(R^{(z_{t+1})}|y_{1:t}) \\ &= X'_{t+1} \mathbf{F}_{t,t+1} X_{t+1} + R^{(z_{t+1})} \end{aligned} \quad (4.13b)$$

donde:

$$\begin{aligned} \mathbf{f}_{t,t+1} &= A^{(z_{t+1})} \mathbf{f}_{t|t}^f \\ \mathbf{F}_{t,t+1} &= \Sigma^{(z_{t+1})} + A^{(z_{t+1})} \mathbf{F}_{t|t}^f A^{(z_{t+1})'}, \end{aligned}$$

con  $\mathbf{f}_{t|t}^f$  y  $\mathbf{F}_{t|t}^f$  el vector de medias y la matriz de covarianzas, respectivamente, de la distribución *filtering* del vector de estados  $\beta_t$ , especificada más adelante. Note que (4.13a)-(4.13b) son análogos a los momentos de la distribución predictiva (3.21) de la Sec. 3.3.2.

Dada la estructura condicional del modelo, se propone usar directamente la distribución  $p(y_t|y_{1:t-1}, \boldsymbol{\theta})$  para inferir sobre  $z_{1:T}$ . De manera explícita, note que la distribución conjunta de  $z_{1:T}$ , dada la secuencia de observaciones y el resto de parámetros, se puede descomponer de la siguiente manera:

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

Usando un procedimiento forward-backward (ver Cap. 3), se muestrea  $z_1$  de  $p(z_1|y_{1:T}, \boldsymbol{\pi}, \boldsymbol{\theta})$ , después se muestrea  $z_2$  condicionado en  $z_1$  de  $p(z_2|z_1, y_{1:T}, \boldsymbol{\pi}, \boldsymbol{\theta})$ , y así sucesiva-

#### 4.4. Regresión con SLDS

mente. Note que para cada  $t$ , la distribución condicional está dada por:

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

Si  $t = 1$ , entonces

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

Cuando  $t = 2$ :

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

En general, la distribución condicional de  $z_t$ , para todo  $t$ , se puede descomponer como sigue:

$$p(z_t | z_{t-1}, y_{1:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) \propto p(z_t | \pi_{z_{t-1}}) p(y_t | y_{1:t-1}, z_t, \boldsymbol{\theta}) m_{t+1,t}(z_t), \quad (4.14)$$

donde,

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

El término  $m_{t,t-1}(z_{t-1})$  se conoce como *backward message* en la literatura de *machine learning* (ver por ejemplo Bishop, 2006). Usando una aproximación truncada

#### 4.4. Regresión con SLDS

para el sticky HDP, el muestreo consiste en calcular  $m_{t,t-1}(k)$  para cada  $k$  *backward* en  $t$ , comenzando con  $m_{T,T-1}(k) = 1 \forall t$ , y después tomar cada  $z_t$  de (4.14) *forward* en  $t$ . El algoritmo se resume de la siguiente manera:

Algoritmo 3. Actualización de la secuencia de modos  $z_{1:T}$

Dado un conjunto de valores para  $\boldsymbol{\pi}$ ,  $\boldsymbol{\theta}$  y  $K$ :

1. Iniciar con  $\mathbf{f}_{0|0}^f = \mathbf{0}$  y  $\mathbf{F}_{0|0}^f = \mathbf{I}$ . Para cada  $t = 1, \dots, T$ , calcular:

$$\begin{aligned}\mathbf{f}_{t-1,t} &= A^{(z_t)} \mathbf{f}_{t-1|t-1}^f \\ \mathbf{F}_{t-1,t} &= \Sigma^{(z_t)} + A^{(z_t)} \mathbf{F}_{t-1|t-1}^f A^{(z_t)'} \\ \mathbf{f}_{t|t}^f &= \mathbf{F}_{t|t}^f \left( 1/R^{(z_t)} X_t' y_t + \mathbf{F}_{t-1,t}^{-1} \mathbf{f}_{t-1,t} \right) \\ \mathbf{F}_{t|t}^f &= \left( 1/R^{(z_t)} X_t' X_t + \mathbf{F}_{t-1,t}^{-1} \right)^{-1}.\end{aligned}$$

2. *Backward*. Hacer  $m_{T+1,T}(k) = 1$  y calcular  $m_{t,t-1}(k)$  para cada  $t \in \{T, \dots, 1\}$  y  $k \in \{1, \dots, K\}$  como sigue:

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

donde

$$\begin{aligned}\tilde{a}_t &= X_t' \mathbf{f}_{t-1,t} \\ \tilde{Q}_t^{(j)} &= X_t' \mathbf{F}_{t-1,t} X_t + R^{(j)}\end{aligned}$$

3. *Forward*. Para cada  $t \in \{1, \dots, T\}$ :

- a) Calcular para cada  $k \in \{1, \dots, K\}$ :

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

- b) Muestrear el modo  $z_t$  de:

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

4. Actualizar  $\boldsymbol{\lambda}$ ,  $\boldsymbol{\pi}$ , y los hiperparámetros del modelo de regresión HDP-SLDS (4.6) como en Fox et al. (2011b).

#### 4.4. Regresión con SLDS

---

##### Block sampling $\beta_{1:T}$

El algoritmo de Fox et al. (2011a) para inferir sobre el modelo (4.1) usa un esquema de muestreo similar al descrito para la secuencia de modos  $z_{1:T}$  para generar muestras de la secuencia de estados  $\beta_{1:T}$ . Un LDS tiene la misma estructura de independencia que un HMM, con la salvedad de que las variables ocultas son continuas, en lugar de discretas. El mecanismo *filtering-smoothing* para muestrear  $\beta_{1:T}$  se produce entonces reemplazando las sumas por integrales. Debido a que las distribuciones de los estados y las observaciones se asumen Gaussianas, las integrales son analíticamente manejables y resultan también Gaussianas (ver sección 3.2.2). Considere la siguiente distribución condicional (*filtering*) para  $t > 1$ ,

$$\begin{aligned} p(\beta_t | y_{1:t}, z_t, \theta) &\propto \int p(y_t | \beta_t, z_t) p(\beta_t | \beta_{t-1}, z_t, \theta) p(\beta_{t-1} | y_{1:t-1}, z_{t-1}, \theta) d\beta_{t-1} \\ &= p(y_t | \beta_t, z_t) \int p(\beta_t | \beta_{t-1}, z_t, \theta) p(\beta_{t-1} | y_{1:t-1}, z_{t-1}, \theta) d\beta_{t-1}. \end{aligned} \quad (4.15)$$

Debido a que el producto de dos densidades Gaussianas es otra Gaussiana, y la integral de una densidad Gaussiana es otra Gaussiana,  $p(\beta_t | y_{1:t})$  es también Gaussiana. Usando esta propiedad de cerradura, se puede representar  $p(\beta_{t-1} | y_{1:t-1}) = N(\mathbf{f}_{t-1|t-1}^f, \mathbf{F}_{t-1|t-1}^f)$ . El objetivo es encontrar un mecanismo recursivo para actualizar  $\mathbf{f}_{t|t}^f$  y  $\mathbf{F}_{t|t}^f$  de  $p(\beta_t | y_{1:t}, z_t, \theta) = N(\mathbf{f}_{t|t}^f, \mathbf{F}_{t|t}^f)$  en términos de  $\mathbf{f}_{t-1|t-1}^f$  y  $\mathbf{F}_{t-1|t-1}^f$ .

Sea

$$\begin{aligned} p(\beta_{t+1} | y_{1:t}, z_{t+1}, \theta) &= \int p(\beta_{t+1} | \beta_t, z_{t+1}, \theta) p(\beta_t | y_{1:t}, z_t, \theta) d\beta_t \\ &\propto \int p(\beta_{t+1} | \beta_t, z_{t+1}, \theta) \alpha_t(\beta_t) d\beta_t \end{aligned}$$

Asumiendo  $p(\beta_t | y_{1:t}, z_t, \theta) = N(\mathbf{f}_{t|t}^f, \mathbf{F}_{t|t}^f)$ , la distribución  $p(\beta_{t+1} | y_{1:t}, z_{t+1}, \theta)$  es Normal, y los momentos se pueden encontrar fácilmente. Por simplicidad, las siguientes derivaciones omiten la dependencia de los parámetros dinámicos sobre  $z_t$ . Note que:

$$\begin{aligned}
 p(\boldsymbol{\beta}_{t+1}|\boldsymbol{\beta}_t) &\propto \exp \left\{ -\frac{1}{2}(\boldsymbol{\beta}_{t+1} - A\boldsymbol{\beta}_t)' \Sigma^{-1}(\boldsymbol{\beta}_{t+1} - A\boldsymbol{\beta}_t) \right\} \\
 &\propto \exp \left\{ -\frac{1}{2}(\boldsymbol{\beta}'_{t+1} \Sigma^{-1} - \boldsymbol{\beta}'_t A \Sigma^{-1})(\boldsymbol{\beta}_{t+1} - A\boldsymbol{\beta}_t) \right\} \\
 &\propto \exp \left\{ -\frac{1}{2}(\boldsymbol{\beta}'_{t+1} \Sigma^{-1} \boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}'_t A' \Sigma^{-1} \boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}'_{t+1} \Sigma^{-1} A \boldsymbol{\beta}_t \right. \\
 &\quad \left. + \boldsymbol{\beta}'_t A' \Sigma^{-1} A \boldsymbol{\beta}_t) \right\} \\
 &\propto \exp \left\{ -\frac{1}{2} \begin{bmatrix} \boldsymbol{\beta}_{t+1} \\ \boldsymbol{\beta}_t \end{bmatrix}' \begin{bmatrix} \Sigma^{-1} & -\Sigma^{-1}A \\ -A'\Sigma^{-1} & A'\Sigma^{-1}A \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_{t+1} \\ \boldsymbol{\beta}_t \end{bmatrix} \right\}
 \end{aligned}$$

y

$$\begin{aligned}
 \alpha_t(\boldsymbol{\beta}_t) &\propto \exp \left\{ -\frac{1}{2}(\boldsymbol{\beta}_t - \mathbf{f}_{t|t}^f)' (\mathbf{F}_{t|t}^f)^{-1}(\boldsymbol{\beta}_t - \mathbf{f}_{t|t}^f) \right\} \\
 &= \exp \left\{ -\frac{1}{2} \left[ \boldsymbol{\beta}'_t (\mathbf{F}_{t|t}^f)^{-1} - (\mathbf{f}_{t|t}^f)' (\mathbf{F}_{t|t}^f)^{-1} \right] \begin{bmatrix} \boldsymbol{\beta}_t - \mathbf{f}_{t|t}^f \end{bmatrix} \right\} \\
 &= \exp \left\{ -\frac{1}{2} \left[ \boldsymbol{\beta}'_t (\mathbf{F}_{t|t}^f)^{-1} \boldsymbol{\beta}_t - (\mathbf{f}_{t|t}^f)' (\mathbf{F}_{t|t}^f)^{-1} \boldsymbol{\beta}_t - \boldsymbol{\beta}'_t (\mathbf{F}_{t|t}^f)^{-1} \mathbf{f}_{t|t}^f \right. \right. \\
 &\quad \left. \left. + (\mathbf{f}_{t|t}^f)' (\mathbf{F}_{t|t}^f)^{-1} \mathbf{f}_{t|t}^f \right] \right\} \\
 &\propto \exp \left\{ -\frac{1}{2} \boldsymbol{\beta}'_t (\mathbf{F}_{t|t}^f)^{-1} \boldsymbol{\beta}_t + \boldsymbol{\beta}'_t (\mathbf{F}_{t|t}^f)^{-1} \mathbf{f}_{t|t}^f \right\} \\
 &= \exp \left\{ -\frac{1}{2} \begin{bmatrix} \boldsymbol{\beta}_{t+1} \\ \boldsymbol{\beta}_t \end{bmatrix}' \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (\mathbf{F}_{t|t}^f)^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_{t+1} \\ \boldsymbol{\beta}_t \end{bmatrix} + \begin{bmatrix} \boldsymbol{\beta}_{t+1} \\ \boldsymbol{\beta}_t \end{bmatrix}' \begin{bmatrix} \mathbf{0} \\ (\mathbf{F}_{t|t}^f)^{-1} \mathbf{f}_{t|t}^f \end{bmatrix} \right\}
 \end{aligned}$$

Combinando los términos, el integrando está dado por:

$$\begin{aligned}
 p(\boldsymbol{\beta}_{t+1}|\boldsymbol{\beta}_t) \alpha_t(\boldsymbol{\beta}_t) &\propto \exp \left\{ -\frac{1}{2} \begin{bmatrix} \boldsymbol{\beta}_{t+1} \\ \boldsymbol{\beta}_t \end{bmatrix}' \begin{bmatrix} \Sigma^{-1} & -\Sigma^{-1}A \\ -A'\Sigma^{-1} & A'\Sigma^{-1}A + (\mathbf{F}_{t|t}^f)^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_{t+1} \\ \boldsymbol{\beta}_t \end{bmatrix} \right. \\
 &\quad \left. + \begin{bmatrix} \boldsymbol{\beta}_{t+1} \\ \boldsymbol{\beta}_t \end{bmatrix}' \begin{bmatrix} \mathbf{0} \\ (\mathbf{F}_{t|t}^f)^{-1} \mathbf{f}_{t|t}^f \end{bmatrix} \right\}
 \end{aligned}$$

Marginalizando sobre  $\boldsymbol{\beta}_{t+1}$  se obtiene la integral deseada:

#### 4.4. Regresión con SLDS

---

$$p(\boldsymbol{\beta}_{t+1}|y_{1:t}) \propto \int p(\boldsymbol{\beta}_{t+1}|\boldsymbol{\beta}_t)\alpha_t(\boldsymbol{\beta}_t)d\boldsymbol{\beta}_t \propto N(\mathbf{f}_{t,t+1}, \mathbf{F}_{t,t+1}),$$

donde:

$$\begin{aligned}\mathbf{f}_{t,t+1} &= A\mathbf{f}_{t|t}^f \\ \mathbf{F}_{t,t+1} &= \Sigma + A\mathbf{F}_{t|t}^f A'\end{aligned}$$

Considerando la dependencia de los parámetros dinámicos sobre  $z_t$ , los momentos  $\mathbf{f}_{t,t+1}$  y  $\mathbf{F}_{t,t+1}$  son los utilizados en (4.13a)-(4.13b), respectivamente. Del modelo (4.6), el término  $p(y_t|\boldsymbol{\beta}_t, z_t)$  de la Ec. (4.15) es normal:

$$\begin{aligned}p(y_{t+1}|\boldsymbol{\beta}_{t+1}, z_{t+1}) &\propto \exp\left\{-\frac{1}{2R^{(z_{t+1})}}(y_{t+1} - X_{t+1}\boldsymbol{\beta}_{t+1})'(y_{t+1} - X_{t+1}\boldsymbol{\beta}_{t+1})\right\} \\ &\propto \exp\left\{-\frac{1}{2R^{(z_{t+1})}}\boldsymbol{\beta}'_{t+1}X'_{t+1}X_{t+1}\boldsymbol{\beta}_{t+1} + \frac{1}{R^{(z_{t+1})}}\boldsymbol{\beta}'_{t+1}X'_{t+1}y_{t+1}\right\}.\end{aligned}$$

Por lo que la distribución en (4.15) está dada por:

$$\begin{aligned}p(\boldsymbol{\beta}_{t+1}|y_{1:t+1}, z_{t+1}, \boldsymbol{\theta}) &\propto p(y_{t+1}|\boldsymbol{\beta}_{t+1}, z_{t+1})p(\boldsymbol{\beta}_{t+1}|y_{1:t}, z_{t+1}, \boldsymbol{\theta}) \\ &\propto \exp\left\{-\frac{1}{2R^{(z_{t+1})}}\boldsymbol{\beta}'_{t+1}(X'_{t+1}X_{t+1} + \mathbf{F}_{t,t+1}^{-1})\boldsymbol{\beta}_{t+1} + \right. \\ &\quad \left. \boldsymbol{\beta}'_{t+1}\left(\frac{1}{R^{(z_{t+1})}}X'_{t+1}y_{t+1} + \mathbf{F}_{t,t+1}^{-1}\mathbf{f}_{t,t+1}\right)\right\} \\ &\propto N(\mathbf{f}_{t+1|t+1}^f, \mathbf{F}_{t+1|t+1}^f),\end{aligned}$$

donde:

$$\mathbf{f}_{t+1|t+1}^f = \mathbf{F}_{t+1|t+1}^f \left( \frac{1}{R^{(z_{t+1})}} X'_{t+1} y_{t+1} + \mathbf{F}_{t,t+1}^{-1} \mathbf{f}_{t,t+1} \right) \quad (4.16)$$

$$\mathbf{F}_{t+1|t+1}^f = \left( \frac{1}{R^{(z_{t+1})}} X'_{t+1} X_{t+1} + \mathbf{F}_{t,t+1}^{-1} \right)^{-1}. \quad (4.17)$$

Note que las Ec. (4.16) y (4.17) son equivalentes al vector de medias y la matriz de covarianzas a posteriori, respectivamente, de la actualización de los estados en el algoritmo del filtro de Kalman (ver Zarchan & Musoff, 2009). La forma de estas ecuaciones se conoce en la literatura como *information filtering*.

El siguiente paso es derivar los parámetros para el muestreo *backward* de las observaciones futuras. Defina  $m_{t+1,t}(\boldsymbol{\beta}_t) \propto p(y_{t+1:T}|\boldsymbol{\beta}_t) \propto N(\mathbf{f}_{t+1,t}, \mathbf{F}_{t+1,t})$ . Dada



#### 4.4. Regresión con SLDS

la estructura de independencia condicional del modelo, la distribución conjunta se puede descomponer como:

$$\begin{aligned}
 m_{t,t-1}(\boldsymbol{\beta}_{t-1}) &\propto \int p(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1})p(y_{t:T}|\boldsymbol{\beta}_t)d\boldsymbol{\beta}_t \\
 &\propto \int p(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1})p(y_t|\boldsymbol{\beta}_t)p(y_{t+1:T}|\boldsymbol{\beta}_t)d\boldsymbol{\beta}_t \\
 &\propto \int p(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1})p(y_t|\boldsymbol{\beta}_t)m_{t+1,t}(\boldsymbol{\beta}_t)d\boldsymbol{\beta}_t. \tag{4.18}
 \end{aligned}$$

Y es posible encontrar un mecanismo recursivo *backward in time* para actualizar los parámetros, de manera análoga al usado para encontrar (4.16) y (4.17). Omitiendo la dependencia de  $\{A, \Sigma, R\}$  sobre  $z_t$  por simplicidad, se expresa cada uno de los términos del integrando en (4.18) de la siguiente manera:

$$\begin{aligned}
 p(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1}) &\propto \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}_t - A\boldsymbol{\beta}_{t-1})'\Sigma^{-1}(\boldsymbol{\beta}_t - A\boldsymbol{\beta}_{t-1})\right\} \\
 &\propto \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}'_t\Sigma^{-1}\boldsymbol{\beta}_t - \boldsymbol{\beta}'_{t-1}A'\Sigma^{-1}\boldsymbol{\beta}_t - \boldsymbol{\beta}'_t\Sigma^{-1}A\boldsymbol{\beta}_{t-1} + \boldsymbol{\beta}'_{t-1}A'\Sigma^{-1}A\boldsymbol{\beta}_{t-1})\right\} \\
 &\propto \exp\left\{-\frac{1}{2}\begin{bmatrix} \boldsymbol{\beta}_{t-1} \\ \boldsymbol{\beta}_t \end{bmatrix}' \begin{bmatrix} A'\Sigma^{-1}A & -A'\Sigma^{-1} \\ -\Sigma^{-1}A & \Sigma^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_{t-1} \\ \boldsymbol{\beta}_t \end{bmatrix}\right\};
 \end{aligned}$$

$$\begin{aligned}
 p(y_t|\boldsymbol{\beta}_t) &\propto \exp\left\{-\frac{1}{2}(y_t - X_t\boldsymbol{\beta}_t)'R^{-1}(y_t - X_t\boldsymbol{\beta}_t)\right\} \\
 &\propto \exp\{y'_tR^{-1}y_t - \boldsymbol{\beta}'_tX'_tR^{-1}y_t - y'_tR^{-1}X_t\boldsymbol{\beta}_t + \boldsymbol{\beta}'_tX'_tR^{-1}X_t\boldsymbol{\beta}_t\} \\
 &\propto \exp\left\{-\frac{1}{2}\begin{bmatrix} \boldsymbol{\beta}_{t-1} \\ \boldsymbol{\beta}_t \end{bmatrix}' \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & X'_tR^{-1}X_t \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_{t-1} \\ \boldsymbol{\beta}_t \end{bmatrix} + \begin{bmatrix} \boldsymbol{\beta}_{t-1} \\ \boldsymbol{\beta}_t \end{bmatrix}' \begin{bmatrix} \mathbf{0} \\ X'_tR^{-1}y_t \end{bmatrix}\right\};
 \end{aligned}$$

$$\begin{aligned}
 m_{t+1,t}(\boldsymbol{\beta}_t) &\propto \exp\left\{-\frac{1}{2}\boldsymbol{\beta}'_t\mathbf{F}_{t+1,t}^{-1}\boldsymbol{\beta}_t + \boldsymbol{\beta}'_t\mathbf{F}_{t+1,t}^{-1}\mathbf{f}_{t+1,t}\right\} \\
 &\propto \exp\left\{-\frac{1}{2}\begin{bmatrix} \boldsymbol{\beta}_{t-1} \\ \boldsymbol{\beta}_t \end{bmatrix}' \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{F}_{t+1,t}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_{t-1} \\ \boldsymbol{\beta}_t \end{bmatrix} + \begin{bmatrix} \boldsymbol{\beta}_{t-1} \\ \boldsymbol{\beta}_t \end{bmatrix}' \begin{bmatrix} \mathbf{0} \\ \mathbf{F}_{t+1,t}^{-1}\mathbf{f}_{t+1,t} \end{bmatrix}\right\}.
 \end{aligned}$$

#### 4.4. Regresión con SLDS

Combinando los términos se tiene:

$$\begin{aligned}
& p(\boldsymbol{\beta}_t | \boldsymbol{\beta}_{t-1}) p(y_t | \boldsymbol{\beta}_t) m_{t+1,t}(\boldsymbol{\beta}_t) \\
& \propto \exp \left\{ -\frac{1}{2} \begin{bmatrix} \boldsymbol{\beta}_{t-1} \\ \boldsymbol{\beta}_t \end{bmatrix}' \begin{bmatrix} A'\Sigma^{-1}A & -A'\Sigma^{-1} \\ -\Sigma^{-1}A & \Sigma^{-1} + X_t'R^{-1}X_t + \mathbf{F}_{t+1,t}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_{t-1} \\ \boldsymbol{\beta}_t \end{bmatrix} \right. \\
& \quad \left. + \begin{bmatrix} \boldsymbol{\beta}_{t-1} \\ \boldsymbol{\beta}_t \end{bmatrix}' \begin{bmatrix} \mathbf{0} \\ X_t'R^{-1}y_t + \mathbf{F}_{t+1,t}^{-1}\mathbf{f}_{t+1,t} \end{bmatrix} \right\} \\
& \propto \exp \left\{ -\frac{1}{2} \begin{bmatrix} \boldsymbol{\beta}_{t-1} \\ \boldsymbol{\beta}_t \end{bmatrix}' \begin{bmatrix} A'\Sigma^{-1}A & -A'\Sigma^{-1} \\ -\Sigma^{-1}A & \Sigma^{-1} + \mathbf{F}_{t|t}^b \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_{t-1} \\ \boldsymbol{\beta}_t \end{bmatrix} + \begin{bmatrix} \boldsymbol{\beta}_{t-1} \\ \boldsymbol{\beta}_t \end{bmatrix}' \begin{bmatrix} \mathbf{0} \\ \mathbf{f}_{t|t}^b \end{bmatrix} \right\} \\
& \propto \exp \left\{ -\frac{1}{2} \begin{bmatrix} \boldsymbol{\beta}_{t-1} \\ \boldsymbol{\beta}_t \end{bmatrix}' \begin{bmatrix} A'\Sigma^{-1}A & -A'\Sigma^{-1} \\ -\Sigma^{-1}A & \Sigma^{-1} + \mathbf{F}_{t|t}^b \end{bmatrix} \left( \begin{bmatrix} \boldsymbol{\beta}_{t-1} \\ \boldsymbol{\beta}_t \end{bmatrix} \right. \right. \\
& \quad \left. \left. - 2 \begin{bmatrix} A'\Sigma^{-1}A & -A'\Sigma^{-1} \\ -\Sigma^{-1}A & \Sigma^{-1} + \mathbf{F}_{t|t}^b \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{0} \\ \mathbf{f}_{t|t}^b \end{bmatrix} \right) \right\}, \tag{4.19}
\end{aligned}$$

donde, dada la dependencia sobre la secuencia de *modos*:

$$\mathbf{F}_{t|t}^b = \frac{1}{R^{(z_t)}} X_t' X_t + \mathbf{F}_{t+1,t}^{-1} \tag{4.20}$$

$$\mathbf{f}_{t|t}^b = \frac{1}{R^{(z_t)}} X_t' y_t + \mathbf{F}_{t+1,t}^{-1} \mathbf{f}_{t+1,t} \tag{4.21}$$

Note que el término en (4.19) es proporcional al Kernel de una densidad normal multivariada. Usando entonces los resultados de la normal para integrar sobre  $\boldsymbol{\beta}_t$ , se tiene que:

$$m_{t,t-1}(\boldsymbol{\beta}_{t-1}) \propto N(\mathbf{f}_{t,t-1}, \mathbf{F}_{t,t-1}) \tag{4.22}$$

donde:

$$\begin{aligned}
\mathbf{f}_{t,t-1} &= \mathbf{F}_{t,t-1} (\Sigma^{(z_t)^{-1}} + \mathbf{F}_{t|t}^b)^{-1} A^{(z_t)'} \Sigma^{(z_t)^{-1}} \mathbf{f}_{t|t}^b \\
\mathbf{F}_{t,t-1}^{-1} &= A^{(z_t)'} \Sigma^{(z_t)^{-1}} A^{(z_t)} - A^{(z_t)'} \Sigma^{(z_t)^{-1}} (\Sigma^{(z_t)^{-1}} + \mathbf{F}_{t|t}^b)^{-1} \Sigma^{(z_t)^{-1}} A^{(z_t)}
\end{aligned}$$

Para el modelo (4.6), el mecanismo para muestrear  $\boldsymbol{\beta}_{1:T}$  consiste en calcular primero  $m_{t+1,t}(\boldsymbol{\beta}_t) \propto p(y_{t+1:T} | \boldsymbol{\beta}_t)$  de (4.22), y tomar cada  $\boldsymbol{\beta}_t$  secuencialmente *forward*

#### 4.4. Regresión con SLDS

in time de:

$$\begin{aligned}
p(\boldsymbol{\beta}_t | \boldsymbol{\beta}_{t-1}, y_{1:T}, z_t, \boldsymbol{\theta}) &\propto p(\boldsymbol{\beta}_t | \boldsymbol{\beta}_{t-1}) p(y_t | \boldsymbol{\beta}_t) m_{t+1,t}(\boldsymbol{\beta}_t) \\
&\propto \exp \left\{ -\frac{1}{2} (\boldsymbol{\beta}_t - A\boldsymbol{\beta}_{t-1})' \Sigma^{-1} (\boldsymbol{\beta}_t - A\boldsymbol{\beta}_{t-1}) \right\} \\
&\quad \cdot \exp \left\{ -\frac{1}{2} (y_t - X_t \boldsymbol{\beta}_t)' R^{-1} (y_t - X_t \boldsymbol{\beta}_t) \right\} \\
&\quad \cdot \exp \left\{ -\frac{1}{2} (\boldsymbol{\beta}_t - \mathbf{f}_{t+1,t})' \mathbf{F}_{t+1,t}^{-1} (\boldsymbol{\beta}_t - \mathbf{f}_{t+1,t}) \right\} \\
&\propto \exp \left\{ -\frac{1}{2} \left[ \boldsymbol{\beta}_t' (\mathbf{F}_{t|t}^b + \Sigma^{-1}) \boldsymbol{\beta}_t - 2\boldsymbol{\beta}_t' (\Sigma^{-1} A\boldsymbol{\beta}_{t-1} + \mathbf{f}_{t|t}^b) \right] \right\} \\
&\propto N(\mu_{\boldsymbol{\beta}_t}, \Sigma_{\boldsymbol{\beta}_t}),
\end{aligned}$$

donde:

$$\begin{aligned}
\Sigma_{\boldsymbol{\beta}_t} &= (\mathbf{F}_{t|t}^b + \Sigma^{-1})^{-1} \\
\mu_{\boldsymbol{\beta}_t} &= \Sigma_{\boldsymbol{\beta}_t} \left( \Sigma^{-1} A\boldsymbol{\beta}_{t-1} + \mathbf{f}_{t|t}^b \right).
\end{aligned}$$

Condicionado a la secuencia de modos  $z_{1:T}$  y al conjunto de parámetros del SLDS,  $p(\boldsymbol{\beta}_t | \boldsymbol{\beta}_{t-1}, y_{1:T}, z_t, \boldsymbol{\theta}, z_{1:T}, \boldsymbol{\theta}) = N(\mu_{\boldsymbol{\beta}_t}, \Sigma_{\boldsymbol{\beta}_t})$ , donde:

$$\begin{aligned}
\Sigma_{\boldsymbol{\beta}_t} &= (\mathbf{F}_{t|t}^b + \Sigma^{(z_t)^{-1}})^{-1} \\
\mu_{\boldsymbol{\beta}_t} &= \Sigma_{\boldsymbol{\beta}_t} \left( \Sigma^{(z_t)^{-1}} A^{(z_t)} \boldsymbol{\beta}_{t-1} + \mathbf{f}_{t|t}^b \right).
\end{aligned}$$

#### Muestreo secuencial de $z_t$

Como se mencionó, las distribuciones conjuntas dadas en (4.8) y (4.11) son condicionadas a una muestra de la secuencia de estados  $\boldsymbol{\beta}_{1:T}$ . Para el modelo (4.8), la propuesta de Fox et al. (2011a) incluye un paso previo al muestreo de  $\boldsymbol{\beta}_{1:T}$ , que consiste en muestrear secuencialmente los modos, marginalizando la secuencia de estados, de la distribución (4.12). Es decir, el algoritmo incluye dos pasos de actualización de los modos  $z_{1:T}$ , uno muestrea secuencialmente cada  $z_t$  y el otro muestrea toda la secuencia de la distribución conjunta (muestreo Gibbs por bloques). Si bien el muestreo secuencial es computacionalmente intensivo, intercalándolo periódicamente en el muestreo mejora la convergencia. El Algoritmo 3 resume el muestreo Gibbs para inferencia del modelo (4.11). De manera semejante a Fox et al. (2011a), se incluye también un paso para muestrear secuencialmente los modos (ver Fox et al., 2011a para más detalles sobre el muestreo secuencial).

Algoritmo 4. Muestreo Gibbs para el modelo de regresión HDP-SLDS

Dado un conjunto previo de probabilidades de transición  $\boldsymbol{\pi}$ , la distribución de transición global  $\lambda$ , y los parámetros dinámicos  $\boldsymbol{\theta}$ :

1. Para cada  $t = \{T, \dots, 1\}$ , muestrear secuencialmente  $z_t$  como en [Fox et al. \(2011a\)](#).
2. Inicializando con  $\mathbf{F}_{T|T}^b = \frac{1}{R^{(z_T)}} X_T' X_T$  y  $\mathbf{f}_{T|T}^b = \frac{1}{R^{(z_T)}} X_T' y_T$ , para cada  $t = \{T, \dots, 1\}$  calcular  $\mathbf{F}_{t|t}^b, \mathbf{f}_{t|t}^b$  como sigue:

$$\mathbf{F}_{t|t}^b = \frac{1}{R^{(z_t)}} X_t' X_t + \mathbf{F}_{t+1,t}^{-1}$$

$$\mathbf{f}_{t|t}^b = \frac{1}{R^{(z_t)}} X_t' y_t + \mathbf{F}_{t+1,t}^{-1} \mathbf{f}_{t+1,t}$$

donde:

$$\mathbf{f}_{t,t-1} = \mathbf{F}_{t,t-1} (\Sigma^{(z_t)^{-1}} + \mathbf{F}_{t|t}^b)^{-1} A^{(z_t)'} \Sigma^{(z_t)^{-1}} \mathbf{f}_{t|t}^b$$

$$\mathbf{F}_{t,t-1}^{-1} = A^{(z_t)'} \Sigma^{(z_t)^{-1}} A^{(z_t)} - A^{(z_t)'} \Sigma^{(z_t)^{-1}} (\Sigma^{(z_t)^{-1}} + \mathbf{F}_{t|t}^b)^{-1} \Sigma^{(z_t)^{-1}} A^{(z_t)}$$

3. Para cada  $t = \{1, \dots, T\}$ , muestrear  $\boldsymbol{\beta}_t$ :

$$\boldsymbol{\beta}_t \sim N(\boldsymbol{\mu}_{\boldsymbol{\beta}_t}, \Sigma_{\boldsymbol{\beta}_t})$$

$$\Sigma_{\boldsymbol{\beta}_t} = (\mathbf{F}_{t|t}^b + \Sigma^{(z_t)^{-1}})^{-1}$$

$$\boldsymbol{\mu}_{\boldsymbol{\beta}_t} = \Sigma_{\boldsymbol{\beta}_t} \left( \Sigma^{(z_t)^{-1}} A^{(z_t)} \boldsymbol{\beta}_{t-1} + \mathbf{f}_{t|t}^b \right).$$

4. Actualizar la secuencia de modos  $z_{1:T}$ , las distribuciones  $\lambda$  y  $\pi_k$ , y los hiperparámetros del modelo de regresión [HDP-SLDS](#) con el Algoritmo 3.
5. Para cada  $k = \{1, \dots, K\}$ , muestrear los parámetros dinámicos  $\{A^{(k)}, \Sigma^{(k)}, R^{(k)}\}$  con el Algoritmo 1.

#### 4.4.2. Estudio de simulación

Para evaluar el desempeño en ajuste del modelo de regresión [HDP-SLDS](#) se examinó un conjunto de datos simulados, generados de un [HMM](#) con dos modos, vector de estados [LDS](#) de tamaño  $p = 3$ , y alta probabilidad de persistencia en un mismo modo. Específicamente, se simularon  $T = 500$  observaciones de la siguiente

#### 4.4. Regresión con SLDS

---

manera:

1. Para cada  $k = \{1, 2\}$ :

a) Los elementos fuera de la diagonal de la matriz de evolución  $A^{(k)}$  se obtienen de la a priori sobre las columnas de  $A^{(k)}$ :

$$N(\mathbf{a}_j^{*(k)}; \mathbf{0}, 1/\alpha_j^{(k)} I_{p-1})$$

$$\alpha_j^{(k)} \sim \text{Gam}(a, b)$$

donde:  $\mathbf{a}_j^{*(k)} = \{a_{1j}^{(k)}, a_{2j}^{(k)}, \dots, a_{j-1,j}^{(k)}, a_{j+1,j}^{(k)}, \dots, a_{pj}^{(k)}\}$ ,  $a$  : parámetro de forma y  $b$  : parámetro de escala. Los hiperparámetros de la distribución de  $\alpha_j^{(k)}$  fueron elegidos de manera que  $\alpha_j^{(k)}$  sea grande ( $a = 10$ ,  $b = 100$ ), implicando varianzas pequeñas en la distribución de  $\mathbf{a}_j^{*(k)}$ . Los elementos de la diagonal de  $A^{(1)}$  se fijaron en 0.5, y la diagonal de  $A^{(2)}$  en 1.0. Esta elección de la matriz  $A^{(k)}$  hace que los estados  $\beta_{tj}$  dependan principalmente de  $\beta_{t-1,j}$ , y en menor medida de cualquier otro  $\beta_{t-1,i}$ ,  $i \neq j$ .

b) La matriz de varianza del error evolución  $\Sigma^{(k)}$  se toma de la a priori:  $\Sigma^{(k)} \sim \text{IW}(n_0, S_0^{(k)})$ , donde  $n_0 = p + 2$  grados de libertad,  $S_0^{(1)} = 0.1I_p$  y  $S_0^{(2)} = 0.5I_p$ .

c) El error de medición se asume indexado por los modos; la precisión se toma de la a priori:  $(1/R^{(k)}) \sim \text{Gam}(a_R, b_R)$ , donde  $a_R = 1$  (parámetro de forma) y  $b_R = 2$  (parámetro de escala).

2. Generar la secuencia de modos  $z_{1:T}$  con probabilidad 0.95 de persistencia en un mismo modo.

3. Para cada  $t = 1, \dots, T$  y  $j = 2, 3, \dots, p$ , generar el valor  $x_{tj}$  de la matriz diseño:

$$x_{tj} = x_{t-1,j} + h_{tj}, \quad h_{tj} \sim N(0, 1)$$

$$x_{0j} = 0 \quad \forall j$$

Esta elección de las covariables como observaciones dependientes entre puntos adyacentes en el tiempo obedece a que, por naturaleza, las series de tiempo exhiben algún tipo de dependencia o correlación. Adicionalmente, en el contexto de regresión de series de tiempo es común que las covariables, o variables independientes, sean también series de tiempo.

#### 4.4. Regresión con SLDS

---

4. Para cada  $t = 1, \dots, T$ , generar el vector de estados  $\boldsymbol{\beta}_t$  y las observaciones  $y_t$ :

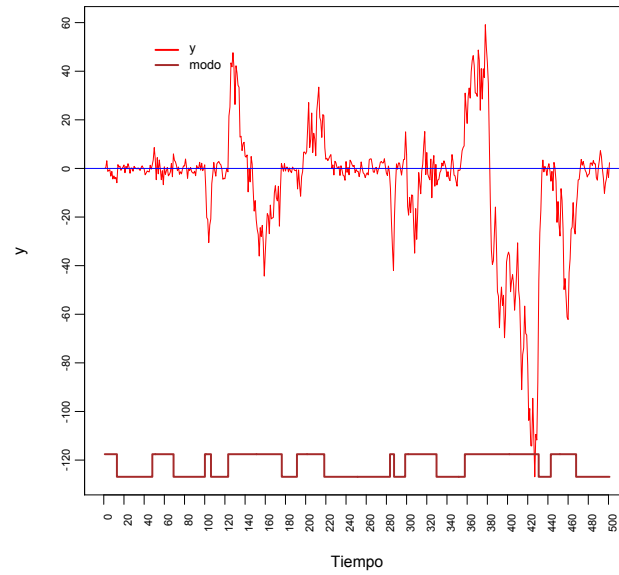
$$\begin{aligned}\boldsymbol{\beta}_t &= A^{(z_t)} \boldsymbol{\beta}_{t-1} + e_t, & e_t &\sim N(\mathbf{0}, \Sigma^{(z_t)}) \\ y_t &= \mathbf{X}'_t \boldsymbol{\beta}_t + w_t, & w_t &\sim N(0, R^{(z_t)})\end{aligned}$$

donde  $\mathbf{X}'_t = \{1, x_{t2}, x_{t3}, \dots, x_{t,p-1}\}$  y  $\boldsymbol{\beta}_0 = \mathbf{0}$ . A diferencia de un modelo estático de regresión en el que los coeficientes de pendiente  $\beta_j$  son fijos para toda  $t$ , en un modelo dinámico de regresión los estados  $\beta_{tj}$  evolucionan con el tiempo. La configuración de  $\{A^{(k)}, \Sigma^{(k)}\}$  establece dos casos para generar la dinámica de los estados: (1)  $A^{(1)}$  con 0.5 en la diagonal principal y valores pequeños fuera de la diagonal es un proceso autoregresivo de orden 1 de *media reversible*, de modo que si  $\boldsymbol{\beta}_t$  es alto con respecto a su media, se espera que disminuya en el siguiente periodo; (2)  $A^{(2)}$  con 1.0 en la diagonal principal y valores pequeños fuera de la diagonal es (como) una caminata aleatoria, por lo que supone que los coeficientes de pendiente  $\boldsymbol{\beta}_t$  se mueven aleatoriamente y no siguen una tendencia. La elección hecha para  $A^{(1)}$  supone que los coeficientes de pendiente en el modelo de regresión retornan lentamente a su valor medio, mientras que la elección para  $A^{(2)}$  como una caminata aleatoria es una manera de modelar transiciones suaves, en el sentido de que el sistema de estados no tiene desplazamientos abruptos.

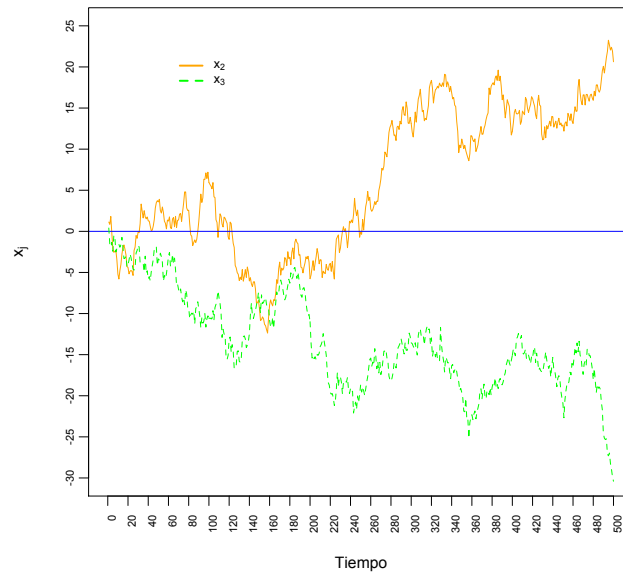
La elección de  $A^{(k)}$  se justifica porque, en ciertas aplicaciones, como en economía, no existe teoría que conduzca directamente a considerar coeficientes aleatorios. Una manera de relajar el supuesto demasiado fuerte de coeficientes de regresión constantes es tener coeficientes que evolucionen lentamente en el tiempo. Esta propiedad deseada de coeficientes de regresión suave, y quizás con tendencia, se puede capturar bien con caminatas aleatorias (Moryson, 1998).

Las Figura 4.6-(a) muestra las observaciones  $y_t$  y la Figura 4.6-(b) las covariables  $\mathbf{x}_j = \{x_{1j}, \dots, x_{Tj}\}$ ,  $j = 2, 3$ . Siguiendo el Algoritmo 3 de la sección anterior, se usó una a priori ARD sobre las columnas de la matriz dinámica  $A^{(k)}$ , una IW sobre  $\Sigma^{(k)}$  e Inversa-Gamma (IG) sobre la varianza del error de medición; el límite de truncamiento para el HDP se fijó como  $K = 30$ . El algoritmo se iteró 10,000 veces. El número de iteraciones se decidió con base en dos aspectos: (1) garantizar consistencia empírica de los estimadores; (2) el tiempo de ejecución del algoritmo.

#### 4.4. Regresión con SLDS



(a)



(b)

**Figura 4.6:** Datos simulados para el modelo de regresión HDP-SLDS. (a):  $y_t$ , (b):  $\mathbf{x}_j$ ,  $j = 2, 3$

##### Ajuste

La Figura 4.7 muestra el ajuste de la serie simulada y la secuencia de modos generada en la última iteración, donde  $\hat{y}_t$  es tal que:

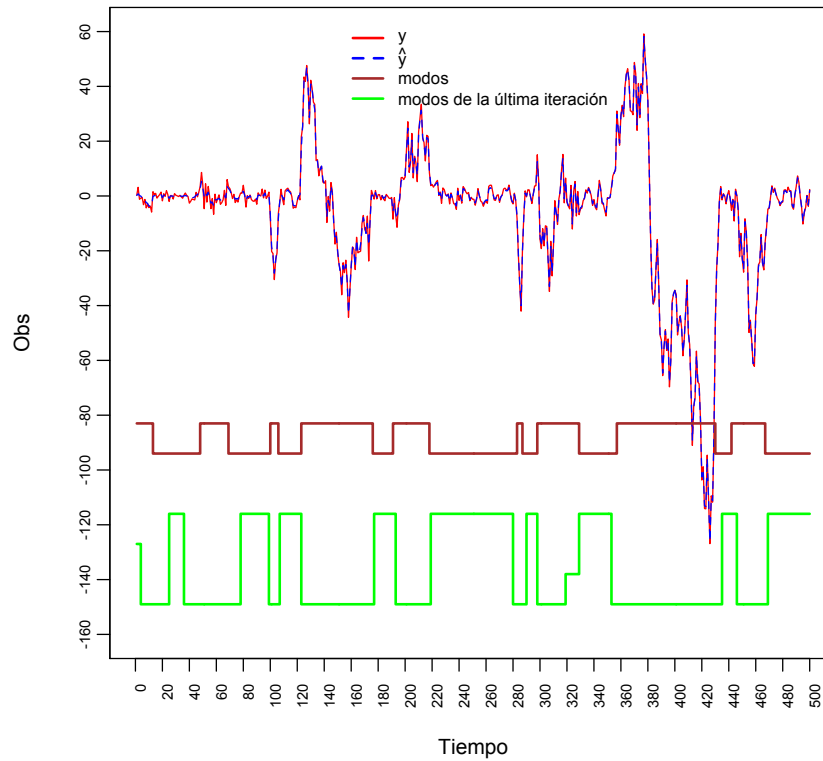
$$\hat{y}_t = \mathbf{X}'_t \hat{\boldsymbol{\beta}}_t, \quad t = 1, \dots, T$$

y  $\hat{\boldsymbol{\beta}}_t$  se obtuvo como el promedio de las iteraciones, después de descartar las primeras 4,000 como el periodo en el que el vector de *estados* alcanza convergencia, estimado con base en los valores de las 10,000 iteraciones. Debido a que hay un vector  $\boldsymbol{\beta}$  para cada  $t$ ,  $t = 1, \dots, 500$ , no se muestran gráficos.

Las observaciones se simularon con alta persistencia en los modos con el objetivo de caracterizar una serie con cambios poco frecuentes pero significativos. La secuencia de modos de la última iteración es muy semejante a la simulada, con algunas diferencias en donde es más difícil discriminar. El algoritmo identifica entre 2 y 3 modos principalmente (ver Figura 4.8); sin embargo, cuando las observaciones se agrupan en 3 o más modos sólo 2 contienen casi todos los datos. Por ejemplo, la iteración 7,400 identificó 4 modos, con número de observaciones 247, 3, 239, y 11 (ver Figura 4.9).



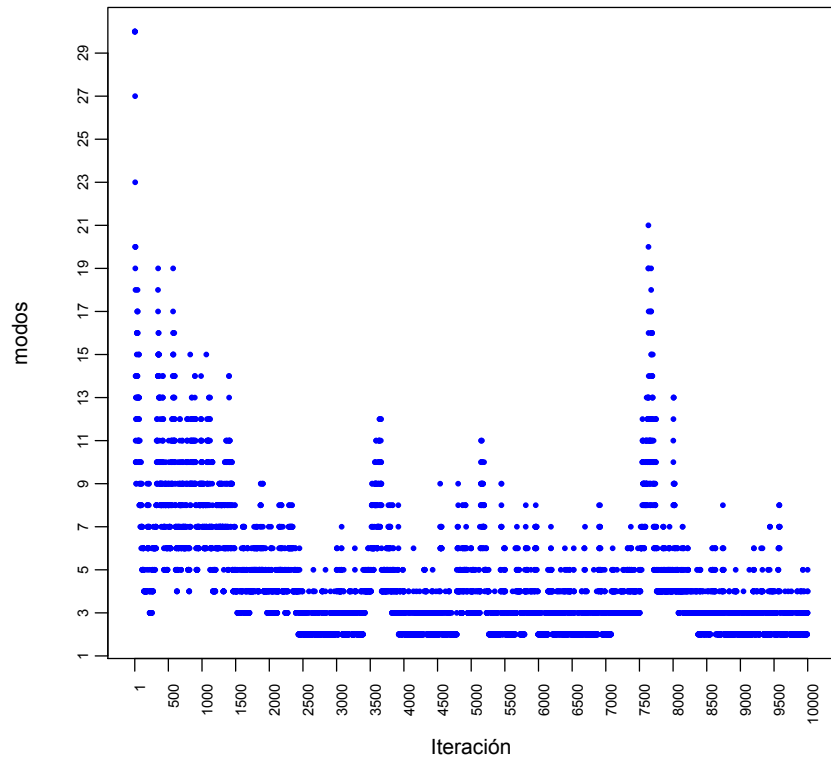
#### 4.4. Regresión con SLDS



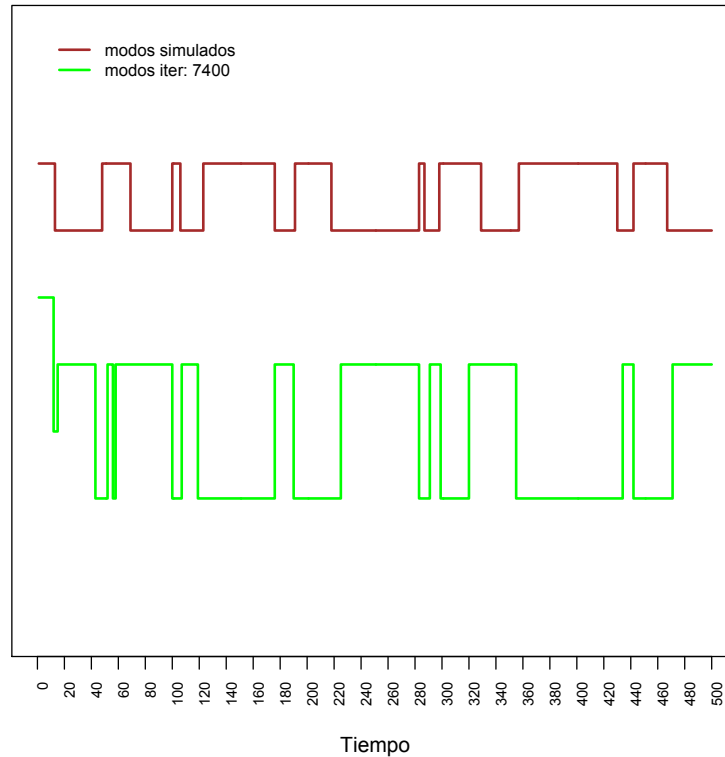
**Figura 4.7:** Ajuste para el modelo de regresión HDP-SLDS sobre 10,000 iteraciones Gibbs del Algoritmo 3, usando límite de truncamiento  $K = 30$ :  $\hat{y}_t = \mathbf{X}'_t \hat{\beta}_t$ , donde  $\hat{\beta}_t$  es el promedio de las iteraciones, después de descartar las primeras 4,000.

#### 4.4. Regresión con SLDS

---

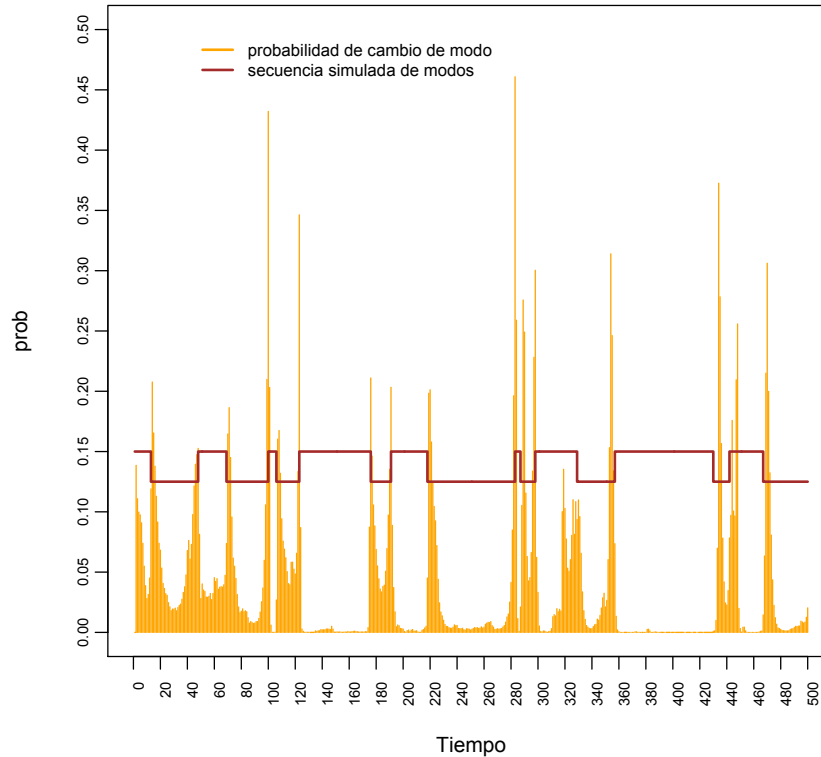


**Figura 4.8:** Número de modos en cada iteración para los datos simulados.



**Figura 4.9:** Secuencia de modos en la iteración 7400.

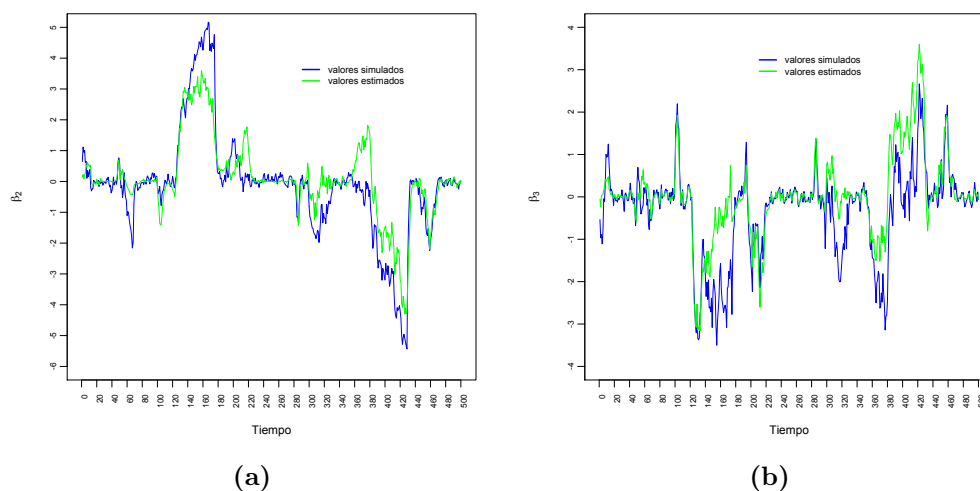
En el análisis Bayesiano de modelos de mezclas finito la estimación de parámetros y la agrupación estimada suelen ser menos sencillas de lo que se podría esperar (Stephens, 2000), esto debido al llamado *label switching problem* originado por la simetría en la verosimilitud de los parámetros del modelo. Jasra et al. (2005) hacen una revisión de varias soluciones a este problema no abordadas en esta Tesis. La Figura 4.7 mostró la secuencia de modos generada en la última iteración, sin embargo, la etiqueta asignada a estos modos podría ser distinta en cualquier otra iteración; es decir, el *label switching problem* puede ocasionar confusión al tratar de identificar un modo en particular en dos iteraciones distintas. Entonces, una manera de resumir la información es mediante la probabilidad estimada en cada  $t$  de un cambio de modo. Debido a que las observaciones simuladas pertenecen sólo a 2 modos, la Figura 4.10 resume adecuadamente las secuencias de todas las iteraciones. Note que las probabilidades más altas son muy cercanas o coincidentes con los tiempo  $t$  en donde se presenta un cambio de modo.



**Figura 4.10:** Probabilidad estimada en cada  $t$  de un cambio de modo para los datos simulados.

Por último, es de especial interés la estimación del vector de estados. La Figura 4.11 muestra la trayectoria simulada (línea azul) y la estimada (línea verde) de cada componente del vector de estados. En aplicaciones de análisis de regresión los coeficientes de regresión tienen cierto significado teórico, y cuantifican el efecto que la covariable asociada tiene sobre las observaciones. En el modelo de regresión dinámica se parte del supuesto de que los coeficientes de regresión varían con el tiempo, pero también es deseable, con fines de interpretación, conocer qué provoca tal variación y de qué manera (por ejemplo, el signo esperado de los coeficientes). Probar si el modelo de regresión dinámica es el modelo apropiado para el proceso que genera los datos observados, o si un modelo de regresión con coeficientes constantes es suficiente, no es tema de esta Tesis. Un caso particular para probar la hipótesis nula de que los coeficientes del modelo de regresión son constantes en el tiempo vs. la alternativa de que algunos coeficientes siguen una caminata aleatoria puede consultarse en [Moryson \(1998\)](#).

#### 4.4. Regresión con SLDS



**Figura 4.11:** Estimación de las pendientes del vector de estados para el modelo de regresión HDP-SLDS: (a)  $\beta_{t2}$ , (b)  $\beta_{t3}$ .

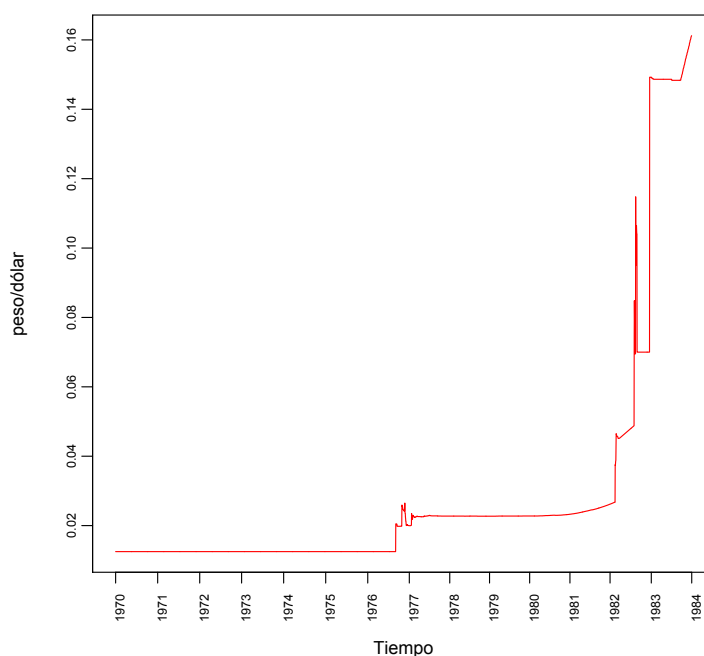
#### 4.4.3. Caso 1: Dinámica del tipo de cambio en México, 1970-2016

La ruptura de los acuerdos de Bretton Woods a inicios de los años 70 fue seguida por una nueva etapa en la evolución del sistema monetario internacional, marcada por una fuerte integración de las economías emergentes en los mercados financieros internacionales. El proceso de integración obligó a los países a llevar a cabo reformas estructurales que permitieran adoptar una economía de mercado y abrir su economía a los flujos de capitales internacionales. En el caso de México, las reformas y desregulación se profundizaron en la segunda mitad de los años 80. Como parte de la ronda de Uruguay, las tarifas de importación se redujeron rápidamente; adicionalmente, se adoptaron políticas para desregular muchas industrias, iniciando la privatización de numerosas empresas propiedad del gobierno, y relajando las restricciones a la inversión extranjera. Las reformas económicas y liberalización del comercio internacional continuaron a inicios de la década de los 90, destacando la negociación del *North American Free Trade Agreement* (NAFTA) con Estados Unidos y Canadá. Después de 1993, la cuenta de capital de México se liberalizó más, y el gobierno permitió al mercado de valores la entrada de capital extranjero (para una extensa revisión de las reformas estructurales ocurridas en México durante 1950-1994 ver [Cárdenas, 1996](#)).

Un importante componente de la estrategia de reforma del gobierno mexicano fue

#### 4.4. Regresión con SLDS

mantener fijo el valor de la moneda con respecto al dólar estadounidense. Esta política sirvió al menos para tres propósitos (Musacchio, 2012): (1) ofrecía a los inversores extranjeros seguridad de que sus inversiones no perderían valor en circunstancias normales; (2) permitió a las compañías mexicanas adquirir préstamos en los mercados internacionales para financiar la expansión necesaria para poder competir ante la apertura del libre comercio en enero de 1994 y; (3) ayudó a las autoridades mexicanas a combatir la inflación interna, forzando a la política monetaria a fluctuar de acuerdo con la balanza de pagos. Sin embargo, un tipo de cambio fijo debe ser sostenido por la capacidad del banco central de dar confianza en la moneda, y por lo tanto acumular reservas. Estas condiciones no se pudieron mantener, y el gobierno mexicano se vio presionando a devaluar el peso en varios momentos (ver Figura 4.12), destacando las devaluaciones de 1976 (Córdoba & Ortíz, 1979) y 1982 (Angeles et al., 1982), y finalmente a abandonar el tipo de cambio fijo en 1994 (Musacchio, 2012).



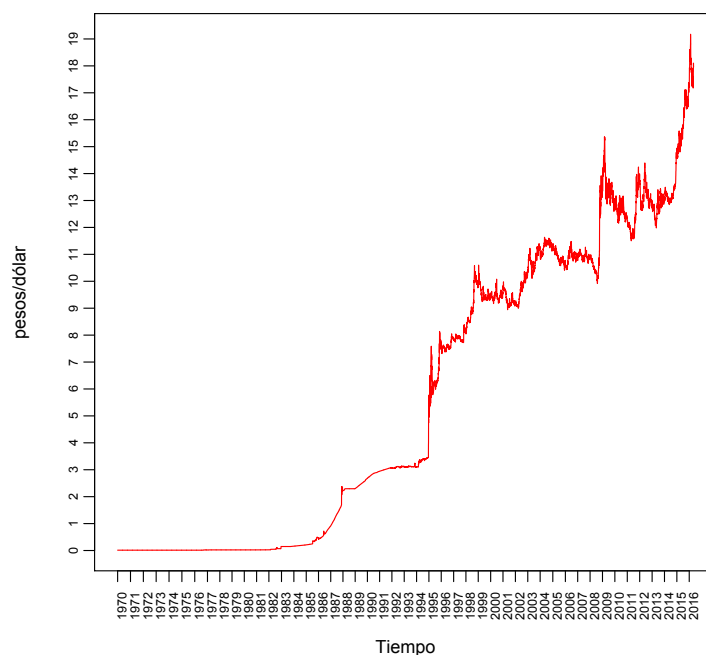
**Figura 4.12:** Tipo de cambio peso-dólar (nuevos pesos), 01/01/1970-30/12/1983. Cifras diarias.

A partir del establecimiento del régimen de libre flotación, el precio del peso con respecto al dólar ha sufrido varias depreciaciones<sup>3</sup> importantes. La Figura 4.13

<sup>3</sup>En términos estrictos, el término *depreciación* se usa en un esquema de flotación libre del

#### 4.4. Regresión con SLDS

muestra el tipo de cambio diario peso-dólar del 01/01/1970 al 11/05/2016. En la gráfica son evidentes las depreciaciones conocidas de 1986-1987, 1994-1995, 1998, 2008-2009 y 2015-2016<sup>4</sup>. Hasta antes de la última depreciación, el comportamiento se había caracterizado por alcanzar un máximo para después ajustarse a la baja. En el periodo más reciente la conducta parece distinta, se está observando la tendencia de mayor duración en la depreciación del peso en la historia económica del país.



**Figura 4.13:** Tipo de cambio peso-dólar, 01/01/1970-11/05/2016. Cifras diarias.

En cualquier economía de libre mercado el tipo de cambio es un precio de gran importancia; la conversión de la moneda local en otra facilita el comercio internacional de bienes y servicios, y de transferencia de fondos entre países. Generalmente se usa la volatilidad<sup>5</sup> del tipo de cambio como una medida de riesgo de la actividad comercial, sin embargo, es difícil tener una conclusión firme sobre la relación entre ésta y el comercio. Algunos estudios han mostrado que un incremento en la

---

tipo de cambio; el término *devaluación* se usa en un esquema de tipo de cambio fijo.

<sup>4</sup>Las devaluaciones de 1976 y 1982 son menos claras debido a que las unidades de la información son *nuevos pesos*, vigentes a partir del 1o de enero de 1993

<sup>5</sup>La *volatilidad* se define como la desviación estándar de las variaciones diarias del tipo de cambio.

#### 4.4. Regresión con SLDS

---

volatilidad del tipo de cambio tiene efectos adversos en el volumen del comercio internacional, mientras que otros dan evidencia de que tiene un impacto positivo (Baum et al., 2004). Por otra parte, varias economías emergentes han usado al tipo de cambio para mantener bajos niveles de inflación, término que se conoce en la literatura como *exchange-rate-based stabilization*. Desafortunadamente estas medidas han terminado en crisis, como las ocurridas en México (1994-1995), Asia (1997-1998), Brasil (1999), y Argentina (2001) (ver Fiess & Shankar, 2009).

La importancia que tiene el tipo de cambio en el comercio internacional y en la estabilidad económica de un país ha motivado el enorme desarrollo de literatura sobre temas concernientes a su estudio con diversos objetivos. Por ejemplo, algunos trabajos se han concentrado en medir los impactos de las intervenciones del banco central en la volatilidad del tipo de cambio (Hung, 1997; Huang, 2007; Mondal, 2013), y otros en investigar la interacción de la volatilidad con otras variables macroeconómicas (Baum et al., 2004; Ruiz & Pozo, 2008; Zhang & Buongiorno, 2010; Ding & Vo, 2012). Más extensa literatura se encuentra cuando el objetivo es modelar la volatilidad. Las propuestas se pueden agrupar en dos corrientes principales que reconocen que la dinámica del tipo de cambio varía en el tiempo: modelos del tipo *autoregressive conditional heteroscedasticity* (ARCH) (Engle, 1982; Bollerslev, 1986), y modelos *stochastic volatility* (SV) (Taylor, 1986). Entre los primeros están los trabajos de McKenzie & Mitchell (2002); Chan (2003); Edmonds Jr. & So (2004); Li et al. (2010); modelos SV se proponen en Tims & Mahieu (2006) y Jouchi (2013). Propuestas más generales explican la persistencia en la volatilidad del tipo de cambio basados en el *Markov switching model* (MSM) de Hamilton (1989, 1990), un modelo del tipo HMM también conocido como *regime switching model* y ampliamente utilizado en la literatura de series de tiempo para caracterizar comportamientos con diferentes *regímenes*. Ver por ejemplo: Bazdresch & Werner (2002); Fiess & Shankar (2009); Jouchi (2013); Nikolsko Rzhevskyy & Prodan (2012). Sin embargo, estos trabajos utilizan un enfoque simplificador que asume que el tipo de cambio transita solo entre dos modos, uno de alta volatilidad y otro de baja volatilidad.

Particularmente para el caso de México, Bazdresch & Werner (2002) explican los movimientos diarios del tipo de cambio durante el periodo 1996-2001 utilizando un modelo del tipo MSM. Los autores encuentran evidencia significativa sobre la existencia de dos regímenes, uno sin tendencia y con poca volatilidad, y el otro con depreciaciones positivas y alta volatilidad. Aunque el modelo permite transitar de un régimen a otro, el número de regímenes se fija inicialmente. Herrera et al. (2011) utilizan también un tipo MSM para modelar la tasa de depreciación del peso frente al dólar, considerando un modelo simple con sólo dos estados, alta y baja volatilidad, con el argumento de que la autoridad monetaria instrumenta medidas de política tendientes a mantener una baja volatilidad, es decir, sin movimientos



#### 4.4. Regresión con SLDS

---

bruscos en la paridad cambiaria, y que cualquier otro estado que indique una mayor volatilidad, significa inestabilidad en el comportamiento cambiario a pesar de las medidas del Banco Central. No se encuentran en la literatura estudios más recientes que modelen cambios de régimen del tipo de cambio en México. En esta sección se emplea un SLDS para modelar el tipo de cambio en el periodo 1970-2016. Se plantean dos objetivos: (1) encontrar coincidencias de los cambios de régimen ocurridos durante el periodo con los eventos de la Tabla 4.1. Los eventos están asociados a periodos de crisis. (2) observar la tendencia de depreciación más reciente.

**Tabla 4.1:** Eventos asociados a los cambios de régimen: México, 1970-2016

Fecha	Evento
18/11/1987	Banco de México cesa su intervención en el mercado cambiario. Se decreta una devaluación de 55 %.
20/12/1994	Devaluación del peso. Evento conocido como <i>error de diciembre</i> .
31/08/1998	Baja en Wall Street. Efecto de la crisis financiera asiática.
15/09/2008	El banco de inversión Lehman Brothers se declara oficialmente en bancarrota. Efecto de la crisis financiera internacional (colapso de la burbuja inmobiliaria en EU).
10/10/2008	<i>Crash</i> generalizado de las bolsas de valores.
16/12/2015	La Reserva Federal (Fed) anuncia aumento en la tasa de interés por primera vez desde 2008.

#### Metodología y datos

Los datos consisten de 11,687 observaciones del tipo de cambio nominal con frecuencia diaria, del 01 de enero de 1970 al 11 de Mayo de 2016, publicados por el Banco de México en la serie histórica del tipo de cambio. La serie se construye de la siguiente manera:

- (a) 1 al 31 de agosto de 1976: tipo de cambio fijo de 12.50 pesos por dólar;
- (b) 1 de septiembre de 1976 al 5 de agosto de 1982: tipo de cambio para operaciones en billete (promedio entre compra y venta);
- (c) 6 al 31 de agosto de 1982: tipo de cambio general;
- (d) 1 de septiembre al 19 de diciembre de 1982: tipo de cambio ordinario;

#### 4.4. Regresión con SLDS

- (e) 20 de diciembre de 1982 al 4 de agosto de 1985: tipo de cambio libre;
- (f) 5 de agosto de 1985 al 10 de noviembre de 1991: tipo de cambio libre;
- (g) 11 de noviembre de 1991 a la fecha: tipo de cambio Fix<sup>6</sup> (Fecha de determinación).

Hay muchas variantes del modelo SLDS general de las Ec. (4.4a)-(4.4b). Fox et al. (2011a) usan una adaptación del SLDS al modelo *Markov switching stochastic volatility* (MSSV) propuesto por Carvalho & Lopes (2007) para estudiar los retornos diarios del índice BOVESPA de la bolsa de valores de Sao Paulo, Brasil, en el periodo 01/03/1997-01/16/2001. Sin embargo, los autores encuentran superioridad del HDP-SLDS (4.4a)-(4.4b) sobre la adaptación al MSSV para encontrar cambios de modo coincidentes con una serie de eventos mundiales ocurridos entre 1997 y 1999. Aunque se utilizan distintas representaciones de los datos en los dos modelos, los cambios de modo capturados por el HDP-SLDS (4.4a)-(4.4b) se alinean mejor con los eventos mundiales. Siguiendo estos resultados, se examinó el tipo de cambio en México mediante el siguiente modelo:

$$\begin{aligned}
 z_t | z_{t-1}, \{\pi_k\}_{k=1}^{\infty} &\sim \pi_{z_{t-1}} \\
 \beta_t &= a^{(z_t)} \beta_{t-1} + e_t^{(z_t)} \\
 y_t &= X_t \beta_t + w_t^{(z_t)}
 \end{aligned} \tag{4.23}$$

donde:

- $y_t$ : tipo de cambio (pesos/dólar) nominal en el tiempo  $t$ .
- $z_t$ : modo en el tiempo  $t$ ,  $z_t = \{1, 2, \dots\} \forall t$ .
- $\pi_j$ : distribución de transición,  $j = 1, 2, \dots, K$ , determinada por un HDP tal que:

$$\begin{aligned}
 G_0 &= \sum_{k=1}^{\infty} \lambda_k \delta_{\theta_k} & \lambda | \gamma &\sim \text{GEM}(\gamma) \\
 G_j &= \sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta_k} & \pi_j | \alpha, \kappa, \lambda &\sim \text{DP}\left(\alpha + \kappa, \frac{\alpha \lambda + \kappa \delta_j}{\alpha + \kappa}\right) \\
 & & \theta_k | H &\sim H
 \end{aligned}$$

---

<sup>6</sup>El tipo de cambio FIX es determinado por el Banco de México los días hábiles bancarios con base en un promedio de las cotizaciones del mercado de cambios al mayoreo para operaciones liquidables el segundo día hábil bancario siguiente. Dichas cotizaciones se obtienen de plataformas de transacción cambiaria y otros medios electrónicos con representatividad en el mercado de cambios. El Banco de México da a conocer el FIX a partir de las 12:00 horas de todos los días hábiles bancarios.

#### 4.4. Regresión con SLDS

---

donde los parámetros son definidos como en (4.4a).

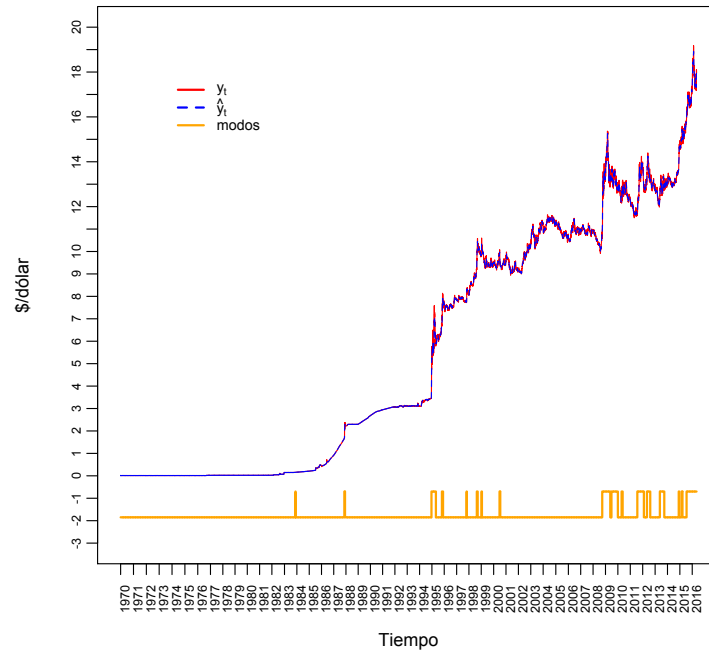
- $X_t = 1$ .
- $a^{(k)} \sim N(0, 1)$
- $w_t^{(k)} \sim N(0, R^{(k)})$ ,  $R^{(k)} \sim \text{IG}(c, d)$
- $\beta_t$ : estado en el tiempo  $t$ .
- $e_t^{(k)} \sim N(0, \sigma^{2(k)})$ ,  $\sigma^{2(k)} \sim \text{IG}(n_0/2, S_0/2)$

Note que al asumir unidimensionalidad en el vector de estados el modelo (4.23) resulta ser el mismo de Fox et al. (2011a) con  $C = [1 \ 0]$ , y un modelo de la clase AR(1). La distribución a priori asumida para los parámetros dinámicos  $\{a^{(k)}, \sigma^{2(k)}\}$  es un caso particular de la MNIG con hiperparámetros:  $M = 0$ ,  $K = V = S_0 = 1$  y  $n_0 = 2$  (grados de libertad); la precisión del ruido de medición se asume  $\text{Gam}(c, 1/d)$ ,  $c = 1$ ,  $d = 0.1$ . Para el HDP se fijó el truncamiento en  $K = 20$ ;  $\lambda$  y  $\pi_j$  se actualizaron como en Fox et al. (2011b). El algoritmo inicia con parámetros tomados de las distribuciones a priori. Se ejecutaron 2000 iteraciones; los resultados se configuraron descartando las primeras 500 iteraciones como periodo de *burn-in*.

#### Resultados

La Figura 4.14 muestra el ajuste de la serie. En la gráfica, la línea naranja representa los cambios de modo que se distinguen, por ejemplo: el primer modo comienza con el inicio de la serie hasta el 16/11/1987; después de esa fecha se distingue otro modo que dura hasta el 02/12/1987; entonces, la serie regresa al modo 1 hasta el 19/12/1994, y así sucesivamente. Estos modos son interpretados como los regímenes de baja y alta volatilidad en términos de Bazdresch & Werner (2002) y Herrera et al. (2011). Después de 2008 la serie tiene un comportamiento más inestable, con frecuentes periodos de alta volatilidad. Para la última parte de la serie (segundo semestre de 2015 hasta el final del periodo) se mantiene el régimen de alta volatilidad. Esta secuencia de modos es tomada de la última iteración del algoritmo, y se justifica por los resultados presentados en la Figura 4.15, que dan evidencia de dos modos predominantes en la serie de tiempo.

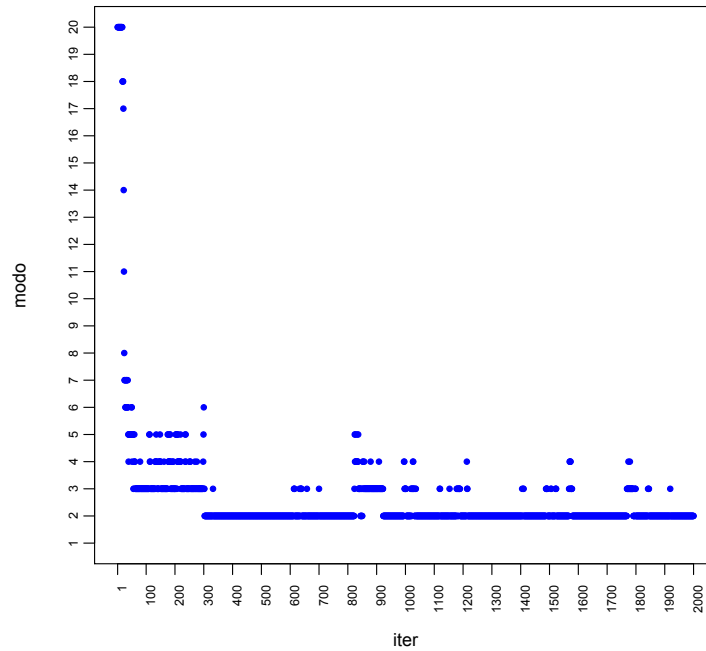
#### 4.4. Regresión con SLDS



**Figura 4.14:** Ajuste de la serie: tipo de cambio peso-dólar, 1970-2016, usando un modelo HDP-SLDS. Línea roja: serie original; línea azul: ajuste; línea naranja: serie de modos dinámicos de la última iteración.

#### 4.4. Regresión con SLDS

---



**Figura 4.15:** Número de modos en cada iteración para los datos de tipo de cambio.

El objetivo de examinar los datos del tipo de cambio en México con el modelo (4.23) es encontrar coincidencias en los cambios de modo con los eventos dados en la Tabla 4.1. Las coincidencias se establecen calculando la probabilidad estimada diaria de un cambio de modo al muestrear la secuencia de modos con el Algoritmo 3. La Fig 4.16 presenta la serie (rojo), el ajuste (azul), la probabilidad de un cambio de modo (naranja), y los eventos de la Tabla 4.1 (verde). Altas probabilidades de cambio de modo (o régimen) son cercanamente coincidentes con los eventos de interés. Adicionalmente, se observa que, con alta probabilidad, la serie presenta cambios de régimen más frecuentes después de 2008 y hasta el primer semestre de 2015. A partir del segundo semestre de 2015 la persistencia en un modo (el de alta volatilidad) se mantiene.

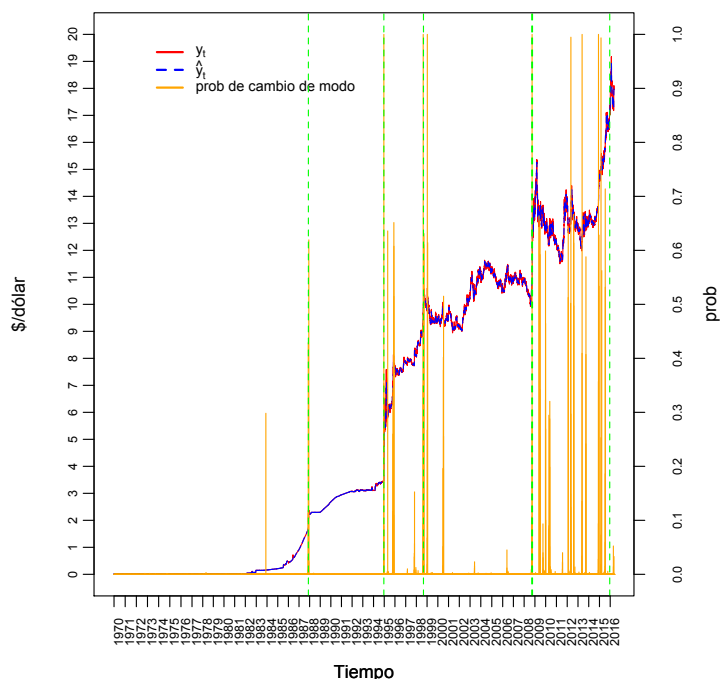


Figura 4.16: Probabilidad estimada diaria de un cambio de modo.

## Conclusiones

Se detectaron cambios de modo coincidentes con una serie de eventos económicos asociados a periodos de crisis. Esto da evidencia del riesgo cambiario como consecuencia de la disminución de la actividad económica.

Por otra parte, después de 2008 el tipo de cambio presenta periodos más frecuentes de alta volatilidad. A partir del segundo semestre de 2015 la persistencia en el régimen de alta volatilidad se mantiene. Esto es, quizás, el hallazgo más importante del modelo. El constante incremento reciente del tipo cambio, y sus niveles máximos históricos, de nuevo dan evidencia del actual desequilibrio económico y la vulnerabilidad de la moneda a ataques especulativos.

### 4.4.4. Caso 2: Niveles de ozono de la ciudad de México

El hecho de que la Tierra ha experimentado un aumento significativo en la temperatura durante el último siglo es indiscutible. Los datos del Goddard Institute for Space Studies (GISS) de la NASA reportan un aumento en la temperatura global de 1.4°F desde 1880. Los 10 años más calurosos de los últimos 136 años han ocurrido después de 2000, y el 2005 se posiciona como el más caluroso (NASA, 2016b). Las consecuencias de este calentamiento observado se reflejan en la contaminación del aire, la temperatura de los océanos, el deshielo de los polos, el aumento del nivel del mar y de la frecuencia e intensidad de fenómenos meteorológicos, que han mostrado cambios sin precedentes en las últimas décadas (Crimmins et al., 2016).

La principal causa de la actual tendencia del calentamiento global es la expansión humana del *efecto invernadero*<sup>7</sup>. Ciertos gases en la atmósfera, como el vapor de agua (H<sub>2</sub>O), el dióxido de carbono (CO<sub>2</sub>), el óxido de nitrógeno (N<sub>2</sub>O), y el metano (CH<sub>4</sub>) bloquean el escape de calor y contribuyen al efecto invernadero (NASA, 2016a). La principal emisión doméstica de estos gases resulta de la generación de electricidad y transporte. Otro gas de efecto invernadero que contribuye significativamente al calentamiento global es el ozono (troposférico o ambiental). Las concentraciones de ozono localizadas en la parte superior de la tropósfera (*global background ozone*) están determinadas por las emisiones mundiales de CH<sub>4</sub>, monóxido de carbono (CO), óxidos de nitrógeno (NO<sub>x</sub>) y compuestos orgánicos volátiles (COV) derivados de fuentes como la quema de combustibles. Altas concentraciones de ozono tienen efectos adversos en la salud como: enfermedades pulmonares (asma, enfisema, bronquitis crónica, cáncer), aumento de riesgo de accidente cerebrovascular, enfermedades del corazón, y nacimientos y muertes prematuros (WHO, 2016). Aunque niños, adultos mayores y personas con enfermedades pulmonares son particularmente vulnerables, todas las personas se encuentran en riesgo ante elevados niveles de ozono.

Expertos en medio ambiente, autoridades gubernamentales y ONG's tienen especial interés en el ozono por su efecto en daños a la salud humana, la vegetación y su contribución al aumento del calentamiento global. Diversos estudios se han concentrado en medir la relación entre la contaminación del aire y la temperatura, por ejemplo Roberts (2004); Huerta et al. (2004); Puza & Roberts (2013); Romero Lankao et al. (2013). Particularmente, Huerta et al. (2004) estudian los niveles de ozono en la Cd. de México considerando un modelo de regresión espaciotemporal con la temperatura como covariable. Los datos provienen de la Red

---

<sup>7</sup>Calentamiento que resulta cuando la atmósfera atrapa el calor que se irradia desde la Tierra hacia el espacio.

#### 4.4. Regresión con SLDS

Automática de Monitoreo Atmosférico (RAMA) de la Cd. de México. Se usa la información de 19 estaciones en las que se miden los niveles de ozono, en partes por mil millones (ppb). Adicionalmente, en 10 de esas estaciones se tienen mediciones de la temperatura en grados centígrados. El objetivo es pronosticar e interpolar los niveles de ozono. El modelo espacio-temporal propuesto por los autores es justificado por los resultados obtenidos de un análisis univariado inicial en el que se relaciona al ozono y la temperatura mediante un DLM para cada una de las 10 estaciones en donde se dispone de mediciones de temperatura. El DLM univariado empleado es el siguiente:

$$\begin{aligned}
 Y_t &= S'_t \boldsymbol{\alpha}_t + Z_t \gamma_t + \epsilon_t, & \epsilon_t &\sim N(0, V) \\
 \boldsymbol{\alpha}_t &= \boldsymbol{\alpha}_{t-1} + \mathbf{w}_{1t}, & \mathbf{w}_{1t} &\sim N(0, \mathbf{W}_{1t}) \\
 \gamma_t &= \gamma_{t-1} + w_{2t}, & w_{2t} &\sim N(0, W_{2t}) \\
 t &= 1, \dots, T
 \end{aligned}$$

donde  $S'_t = (\cos(\pi t/12), \sin(\pi t/12), \cos(\pi t/6), \sin(\pi t/6))$ ,  $\boldsymbol{\alpha}'_t = (\alpha_{1t}, \alpha_{2t}, \alpha_{3t}, \alpha_{4t})$ ,  $\gamma_t$  es un escalar, los términos de error  $\epsilon_t$ ,  $\mathbf{w}_{1t}$  y  $w_{2t}$  se asumen independientes uno del otro, la variable respuesta  $Y_t$  denota la raíz cuadrada de la concentración de ozono y  $Z_t$  la temperatura. La varianza de  $\epsilon_t$  es desconocida pero igual para toda  $t$ , y las varianzas de  $\mathbf{w}_{1t}$  y  $w_{2t}$  son modeladas con un factor de descuento de 0.97 (ver [West & Harrison, 1997](#)). Los términos senos y cosenos con periodicidades  $2\pi/24$  y  $2\pi/12$  se determinaron mediante un periodograma Bayesiano.

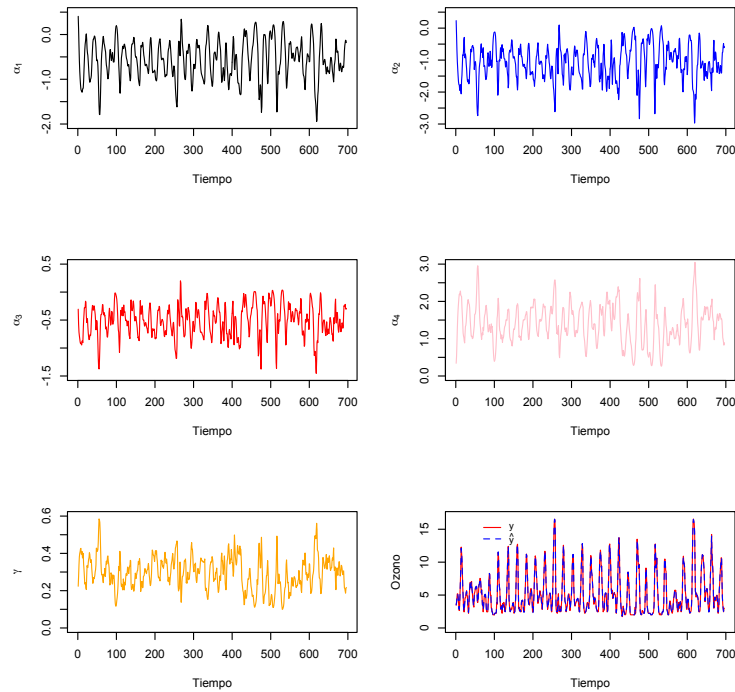
Siguiendo la propuesta univariada de [Huerta et al. \(2004\)](#), se ajustó el modelo (4.6) con el Algoritmo 3 para 2 estaciones: Plateros y Tlalnepantla, con información del 2 al 30 de septiembre de 1997<sup>8</sup>. En este caso, la matriz diseño  $X_t$  (sin intercepto) contiene a los elementos de  $S'$  y la temperatura  $Z_t$ . El error de observación, el error del proceso y las columnas de la matriz de evolución  $A^{(k)}$  se asumen normales con media cero y varianzas  $R^{(k)}$ ,  $\Sigma^{(k)}$  y  $1/\alpha_j^{(k)} I_5$ , respectivamente; sin embargo, se espera que el modelo identifique un sólo grupo para todas las observaciones, ya que los coeficientes de pendiente capturan el comportamiento cíclico de las observaciones. Los hiperparámetros usados para las distribuciones a priori fueron los siguientes:  $n_0 = p + 2 = 7$  grados de libertad y  $S_0 = 0.01 I_5$  para la varianza del error del proceso;  $r_a = 1$  (parámetro de forma) y  $r_b = 2$  (parámetro de escala) de una distribución  $\text{Ga}(r_a, r_b)$  para la precisión del error de observación;  $a = 1$  (parámetro de forma) y  $b = 100$  (parámetro de escala) de una distribución  $\text{Ga}(a, b)$  para el parámetro de precisión  $\alpha_j^{(k)}$ . El algoritmo se iteró 6000 veces; los resultados que se muestran son el promedio descartando las primeras 4000, después de las que se alcanza convergencia de los coeficientes.

<sup>8</sup>Consultada el 05 de julio de 2015 de la base de datos de Red Automática de Monitoreo Atmosférico (RAMA) de la Cd. de México.



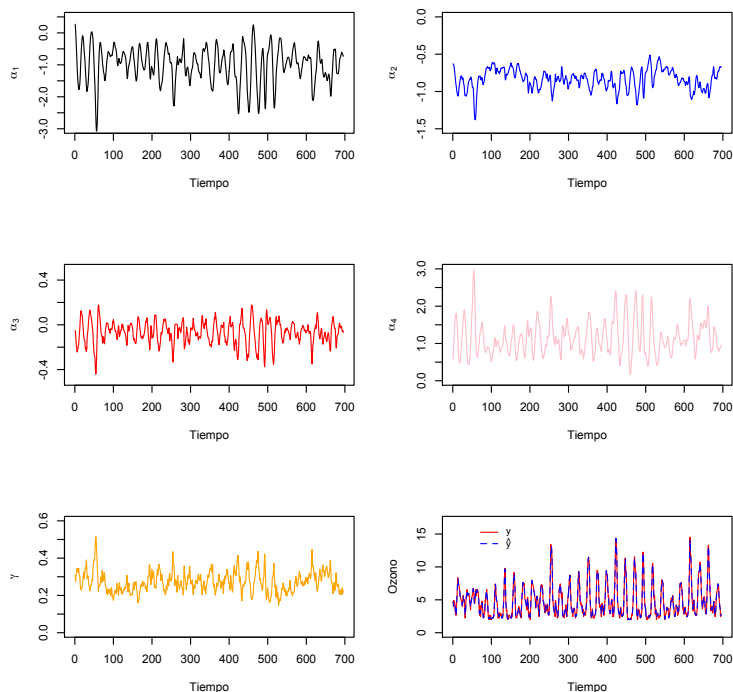
#### 4.4. Regresión con SLDS

Los coeficientes estimados se muestran en las Figura 4.17 y 4.18. Aunque el modelo (4.6) se ajustó sólo para 2 estaciones, los resultados concuerdan con los obtenidos por Huerta et al. (2004). Las estimaciones para  $\gamma_t$  son muy variables a través del tiempo pero relativamente menos variables entre estaciones, es decir, la temperatura tiene el mismo efecto en todas las estaciones, pero el efecto cambia con el tiempo. En cambio, los parámetros  $\alpha_t$  son bastante variables entre estaciones y en el tiempo: la fase de la serie es constante entre estaciones mientras que la amplitud varía sustancialmente, lo que significa que hay zonas más expuestas a contaminación ambiental. En lo que respecta al ajuste, Huerta et al. (2004) no presentan resultados para el caso univariado. La gráfica inferior derecha de las Figura 4.17 y 4.18 muestra la serie y el buen ajuste del modelo de regresión HDP-SLDS para los datos de los niveles de ozono. De manera similar al trabajo de Huerta et al. (2004), el modelo de regresión HDP-SLDS puede ser adaptado como un modelo espacio-temporal. A priori, se espera que un modelo de este tipo agrupe las observaciones e identifique parámetros dinámicos distintos para cada grupo.



**Figura 4.17:** Parámetros estimados y ajuste para la estación Plateros de monitoreo atmosférico.

## 4.5. Regresión espuria y cointegración



**Figura 4.18:** Parámetros estimados y ajuste para la estación Tlalneptla de monitoreo atmosférico.

## 4.5. Regresión espuria y cointegración

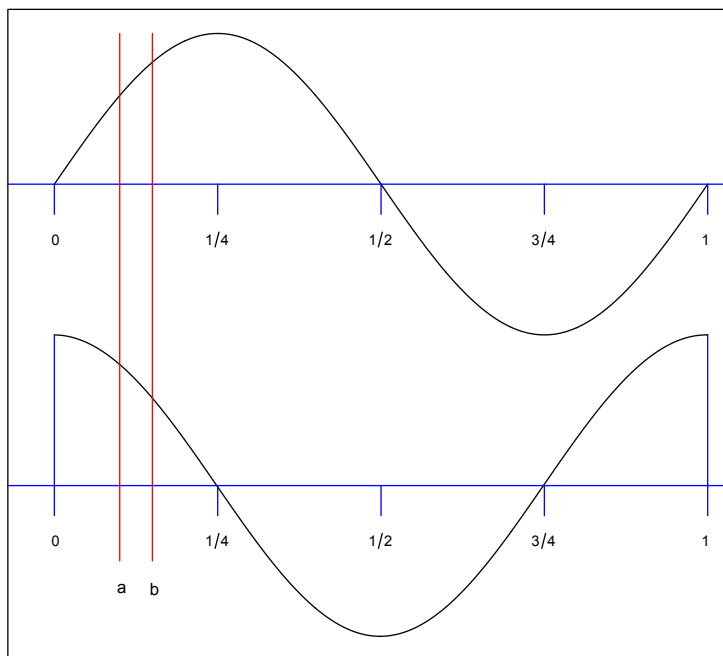
El término *relación espuria* hace referencia a una relación matemática en la cual dos o más variables no están causalmente relacionadas entre sí, es decir, son independientes, y sin embargo se podría inferir *erróneamente* una relación debida a la casualidad o a algún factor no observado. El término *erróneamente* significa que se obtienen correlaciones significativas entre variables cuya relación no tiene sentido o interpretación. Esta situación fue observada primeramente por [Yule \(1926\)](#) en un contexto de series de tiempo. El argumento es que lo que se observa es solamente una parte infinitesimal de un periodo muy grande de tiempo, por lo que tales correlaciones son una fluctuación del muestreo; si se tuviera una muestra sobre un periodo mucho más grande, no se encontraría ninguna correlación entre ellas.

De acuerdo con [Yule \(1926\)](#), dos series que presentan una correlación significativa cuando se estudian en un periodo corto de tiempo, pero que resultan no correlacio-

## 4.5. Regresión espuria y cointegración

---

nadas al considerar el periodo completo, son funciones armónicas simples, ambas en el mismo periodo, pero que difieren en su *fase*, como las que se muestran en la Figura 4.19.



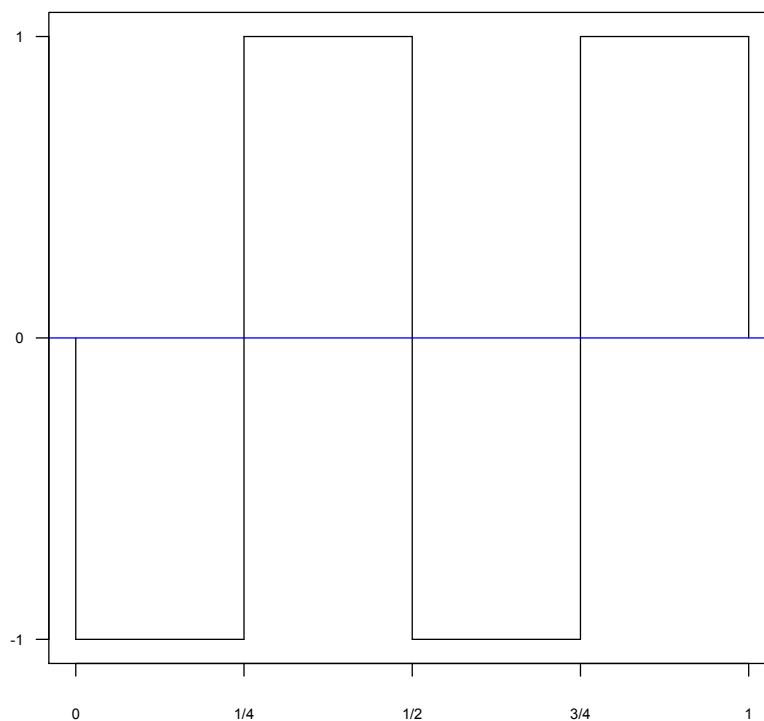
**Figura 4.19:** Curvas senoidales con distinta *fase*.

Suponga que las curvas de la Figura 4.19 representan las fluctuaciones de dos variables sobre un periodo de tiempo tan grande que no es posible tener toda la información, por ejemplo muchos siglos. Las curvas difieren en su *fase* por un cuarto de periodo. Es evidente que la correlación entre las dos variables, considerando todo el intervalo de tiempo, es cero: desviaciones positivas en una variable corresponden con igual frecuencia a desviaciones positivas y negativas en la otra variable. Ahora, si se considera un intervalo de tiempo tan corto como el delimitado entre las líneas verticales  $a, b$ , entonces los segmentos de las dos curvas son casi líneas rectas, la superior con pendiente positiva y la inferior con pendiente negativa: la correlación entre las observaciones en este segmento será muy cercana a  $-1$ . Si el intervalo delimitado se hiciera infinitesimalmente pequeño, tal que los segmentos de las dos curvas puedan considerarse como lineales, y se trazan los

## 4.5. Regresión espuria y cointegración

---

cambios en el coeficiente de correlación conforme el centro del intervalo se mueve en la figura de izquierda a derecha, entonces el coeficiente de correlación variaría como se muestra en la Figura 4.20. Yule (1926) muestra que la distribución de frecuencias de las correlaciones se concentra cada vez más cerca de cero conforme aumenta la proporción del intervalo de tiempo en que se toman las muestras, sin embargo, la distribución mantiene siempre una forma de U, y valores distintos de cero son siempre los más frecuentes. En consecuencia, sugiere que se obtienen correlaciones sin sentido (*espurias*) entre series de tiempo análogas a las series armónicas de la Figura 4.19 cuando se dispone de muestras muy pequeñas, comparadas con la medida total del intervalo.



**Figura 4.20:** Variación de la correlación entre dos elementos infinitesimales de las curvas armónicas de la Figura 4.19, cuando el centro del intervalo se mueve de izquierda a derecha.

Estudiar series de tiempo con modelos de regresión clásicos requiere especial atención; por su naturaleza, los datos de series de tiempo pueden incidir en las propiedades de los estimadores de regresión y de los métodos de inferencia. En general,

## 4.5. Regresión espuria y cointegración

---

las técnicas de regresión clásicas son apropiadas para el análisis de series estacionarias.

Un proceso estacionario es un proceso estocástico cuya distribución de probabilidad conjunta no cambia cuando se desplaza en el tiempo. En consecuencia, los parámetros, tales como la media y la varianza, tampoco cambian en el tiempo ni siguen alguna tendencia. Es decir, una serie de tiempo  $y_{1:T}$  es estacionaria si (Zivot & Wang, 2006):

$$\begin{aligned} F_Y(y_{t_1+s}, \dots, y_{t_k+s}) &= F_Y(y_{t_1}, \dots, y_{t_k}) \\ E[y_t] &= \mu \quad \forall t \\ \text{Cov}(y_t, y_{t-1}) &= E[(y_t - \mu)(y_{t-j} - \mu)] = \gamma_j \quad \forall t \text{ y cualquier } j \end{aligned}$$

Cuando se tiene una serie no estacionaria es práctica común calcular las diferencias entre observaciones consecutivas para tener una serie estacionaria. Este proceso se conoce como diferenciación, y se dice que una serie es integrada de orden  $d$ , denotada como  $I(d)$ , si  $d$  es el mínimo número de diferencias que se requieren para obtener una serie estacionaria en covarianza (Engle & Granger, 1987). Como se mencionó, el análisis de regresión clásico con series de tiempo requiere que las variables sean estacionarias, es decir  $I(0)$ , para que se mantengan los resultados del modelo. Considere la siguiente regresión lineal de series de tiempo:

$$y_t = \beta_0 + \beta_1 x_{1t} + \dots + \beta_k x_{kt} + \epsilon_t = \mathbf{x}_t^T \boldsymbol{\beta} + \epsilon_t, \quad t = 1, \dots, T \quad (4.24)$$

donde  $\mathbf{x}_t = (1, x_{1t}, \dots, x_{p-1,t})^T$  es un vector de covariables,  $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_{p-1})$  es un vector de coeficientes, y  $\epsilon_t$  es un término de error aleatorio. Los modelos clásicos de regresión de series de tiempo funcionan bajo los siguientes supuestos (Hayashi, 2000):

1. El modelo lineal (4.24) es correctamente especificado.
2.  $\{y_t, \mathbf{x}_t\}$  es conjuntamente estacionario y ergódico.
3. Los regresores  $\mathbf{x}_t$  están predeterminados:  $E[x_{is}\epsilon_t] = 0 \quad \forall s \neq t \text{ y } i = 1, \dots, k$ .
4.  $E[\mathbf{x}_t \mathbf{x}_t^T] = \sum_{XX}$  es de rango completo  $p$ .
5.  $\{\mathbf{x}_t \epsilon_t\}$  es un proceso no correlacionado con matriz de covarianzas finita  $E[\epsilon_t^2 \mathbf{x}_t \mathbf{x}_t^T] = \sigma^2 \sum_{XX}$ .

Bajo estos supuestos, los estimadores de mínimos cuadrados ordinarios (MCO)  $\hat{\boldsymbol{\beta}}$  son consistentes y asintóticamente normalmente distribuidos. Un caso en el cual

## 4.5. Regresión espuria y cointegración

las propiedades de los estimadores MCO no se mantienen es el de *regresión espuria*, producida cuando los regresores (covariables) son  $I(1)$  y son no *cointegrados*. Los componentes de un vector  $\mathbf{z}_t$  son cointegrados de orden  $(d, b)$ , denotado como  $\mathbf{z}_t \sim CI(d, b)$ , si (Granger, 1981; Engle & Granger, 1987):

- todos los componentes de  $\mathbf{z}_t$  son  $I(d)$ , y
- existe un vector  $\alpha (\neq 0)$  tal que  $\mathbf{v}_t = \alpha^T \mathbf{z}_t \sim I(d - b)$ ,  $b > 0$ .

El vector  $\alpha$  se conoce como vector de cointegración. El siguiente ejemplo ilustra el problema de *regresión espuria*.

**Ejemplo:** Considere dos procesos independientes y no cointegrados  $\{y_{1t}\}$  y  $\{y_{2t}\}$  tales que:

$$y_{it} = y_{it-1} + e_{it}, \quad e_{it} \sim N(0, 1), \quad i = 1, 2, \quad t = 1, \dots, T$$

Se simularon  $T = 500$  observaciones para cada serie y se graficaron en la Figura 4.21. La gráfica de los datos sugiere que las dos series están positivamente correlacionadas. El estimador de MCO para la pendiente de la regresión de  $y_{1t}$  sobre  $y_{2t}$  refuerza esta observación:

Coefficientes:

	Estimate	Std. Error	<i>t</i> value	Pr(>   <i>t</i>  )
(Intercept)	5.61618	0.21009	26.732	< 2e-16
$\mathbf{y}_2$	0.39299	0.04229	9.293	< 2e-16

Res std error: 4.696 on 498 df

$R^2$ : 0.1478      Adjusted  $R^2$ : 0.1461

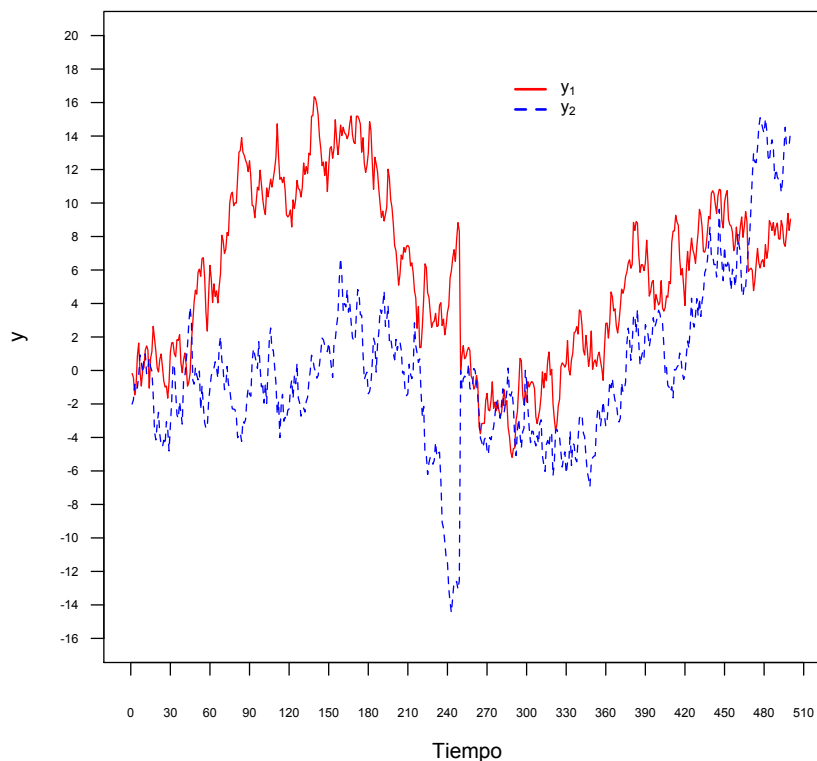
DW stat: 0.064965      *p*-value: < 2.2e-16

F-stat: 86.36 on 1 and 498 df      *p*-value: < 2.2e-16

donde:  $\mathbf{y}_i = \{y_{i1}, \dots, y_{iT}\}$ . El coeficiente de pendiente estimado es 0.39299 con valor grande del estadístico *t* y una  $R^2$  moderada. Sin embargo, el estadístico Durbin-Watson (DW stat) sugiere que hay fuerte correlación positiva entre los residuales. Estos resultados están comúnmente asociados con una *regresión espuria*.

## 4.5. Regresión espuria y cointegración

---



**Figura 4.21:** Datos simulados de dos procesos  $I(1)$  independientes.

Considere el modelo de regresión (4.24) y suponga que las series  $\mathbf{Y}_t = (y_t, \mathbf{x}_t^T)$  son  $I(d)$  no cointegradas. Entonces (4.24) es una *regresión espuria* y el verdadero valor de  $\beta$  es cero. Los estimadores MCO tienen las siguientes implicaciones (Phillips, 1986):

- $\hat{\beta}$  no converge en probabilidad a cero, si no que converge en distribución a una variable aleatoria no-normal, no necesariamente centrada en cero. Este es el fenómeno de *regresión espuria*.
- El estadístico  $t$  para probar que los elementos de  $\beta$  son cero diverge a  $\pm\infty$  cuando  $T \rightarrow \infty$ .
- La  $R^2$  de la regresión converge a uno cuando  $T \rightarrow \infty$ .
- La regresión con datos  $I(1)$  tiene sentido solo cuando los datos son cointegrados.

## 4.5. Regresión espuria y cointegración

Si las series  $\mathbf{Y}_t = (y_t, \mathbf{x}_t^T)$  del modelo (4.24) son  $I(1)$  cointegradas, entonces existe una combinación lineal de ellas que es estacionaria, o  $I(0)$ . En economía, la combinación lineal  $\boldsymbol{\beta}^T \mathbf{Y}_t$  se conoce como relación de *equilibrio de largo plazo* (Engle & Granger, 1987). La justificación es que las series  $I(1)$  con una relación de equilibrio de largo plazo no pueden apartarse demasiado de éste porque las *fuerzas económicas* actúan para restablecer el equilibrio. Debido a que el vector  $\boldsymbol{\beta}$  no es único ( $c\boldsymbol{\beta}^T \mathbf{Y}_t = \boldsymbol{\beta}^{*T} \mathbf{Y}_t \sim I(0)$ ), el vector es normalizado tal que la relación de cointegración se puede expresar como:

$$y_t = \beta_1 x_{1t} + \cdots + \beta_k x_{kt} + \epsilon_t$$

donde  $\epsilon_t \sim I(0)$  es conocido como el *error de desequilibrio*. En el largo plazo  $\epsilon_t$  es cero, y la relación de equilibrio de largo plazo es

$$y_t = \beta_1 x_{1t} + \cdots + \beta_k x_{kt}$$

Las desviaciones del equilibrio de largo plazo ejercen influencia en la dinámica de corto plazo, descrita en la siguiente sección. Un modelo que permite la estimación de los efectos de corto y largo plazo se conoce en la literatura como modelo de corrección de error.

### 4.5.1. Modelo de corrección de error

Considere un vector bivariado  $\mathbf{Y}_t = (y_t, x_t)^T$ , y asuma que  $\mathbf{Y}_t$  es cointegrado con vector de cointegración  $\boldsymbol{\beta} = (1 - \beta_1)^T$ , tal que  $\boldsymbol{\beta}^T \mathbf{Y}_t = y_t - \beta_1 x_t$  es  $I(0)$ . Engle & Granger (1987) mostraron que la cointegración implica la existencia de un modelo de corrección de error (ECM) de la forma:

$$\Delta y_t = \beta_{01} + \alpha_1 (y_{t-1} - \beta_1 x_{t-1}) + \sum_j \psi_{11}^j \Delta y_{t-j} + \sum_j \psi_{12}^j \Delta x_{t-j} + \epsilon_{1t} \quad (4.25)$$

$$\Delta x_t = \beta_{02} + \alpha_2 (y_{t-1} - \beta_1 x_{t-1}) + \sum_j \psi_{21}^j \Delta y_{t-j} + \sum_j \psi_{22}^j \Delta x_{t-j} + \epsilon_{2t} \quad (4.26)$$

que describe el comportamiento de  $y_t$  y  $x_t$ , donde  $\Delta$  denota la primera diferencia. El ECM vincula la relación de equilibrio de largo plazo implicada por la cointegración (segundo término del lado derecho) con el mecanismo de ajuste de corto plazo que describe cómo las variables reaccionan cuando se alejan del equilibrio (tercero y cuarto términos del lado derecho). El coeficiente  $\alpha_i$ ,  $i = 1, 2$ , captura la velocidad de ajuste hacia el equilibrio, y el término  $\epsilon_{it}$ ,  $i = 1, 2$ , representa *shocks* aleatorios que el sistema recibe.



## 4.5. Regresión espuria y cointegración

---

Se encuentran en la literatura varios métodos para probar cointegración de las series de tiempo y examinar relaciones entre variables de la forma ECM, por ejemplo [Engle & Granger \(1987\)](#); [Johansen \(1991\)](#); [Pesaran et al. \(2001\)](#). En un contexto Bayesiano, recientes propuestas se pueden consultar en [Kleibergen & Paap \(2002\)](#); [Paap & van Dijk \(2003\)](#); [Sugita \(2008\)](#); [Diniz et al. \(2012\)](#). Aplicaciones de los modelos ECM se encuentran principalmente en economía y finanzas. La literatura al respecto es muy basta, pero varios ejemplos pueden revisarse en [Alexander \(2001\)](#).

Desde los primeros trabajos sobre cointegración la atención se había centrado en las relaciones de equilibrio en las que se esperaba que algunas series estuvieran cointegradas con otras. Sin embargo, en algunos casos no es posible rechazar la hipótesis nula de no cointegración. [Brenner & Kroner \(1995\)](#) dan fundamentos teóricos a situaciones en las que no se ha encontrado suficiente evidencia de cointegración. Alternativamente, varios autores han propuesto modelos donde la cointegración ocurre temporalmente. En una aplicación, [Siklos & Granger \(1996\)](#) muestran que ocurre cointegración en la paridad de las tasas de interés US-Canadá sólo bajo el régimen de control de inflación en Canadá. Los autores seleccionan un punto que divide la serie en dos grupos. El método propuesto requiere información sobre la localización exacta del cambio de régimen. [Balke & Fomby \(1997\)](#) estudian la cointegración como un modelo de umbral, en el que las series son cointegradas si se apartan demasiado del equilibrio, y son no cointegradas mientras están cerca del equilibrio. El argumento es que hay un mecanismo de ajuste hacia el equilibrio de largo plazo, y sólo cuando las desviaciones del equilibrio exceden cierto umbral, los beneficios del ajuste exceden los costos; por tanto, los agentes económicos actúan para regresar el sistema al equilibrio. [Sugita \(2008\)](#) introduce una aproximación Bayesiana a un modelo de cointegración *Markov switching* que permite que la relación de cointegración esté o no presente dependiendo del *regimen*. En el modelo, la variable de estado (*regimen*) denota la presencia o ausencia de cointegración, y el vector de cointegración se estima de su distribución condicional, dado el conjunto estimado de los estados.

En el modelo de regresión [HDP-SLDS](#) de la sección 4.3 no se establece explícitamente algún supuesto sobre  $\{y, \mathbf{x}\}$ , sin embargo, se asume que hay una relación significativa (no *espuria*) entre las variables, y es de interés para el investigador cuantificar tal relación. Adicionalmente, la naturaleza dinámica del modelo de regresión [HDP-SLDS](#) permite que el grado de dependencia de  $y$  sobre las covariables cambie en el tiempo, pero donde observaciones continuas comparten un modelo dinámico determinado por un *modo* específico. Intuitivamente, si para un modo particular se estima un coeficiente de pendiente igual a cero, es decir, si el grado de dependencia de  $y$  sobre una covariable es cero, entonces se espera también que no se encuentre evidencia de cointegración en ese modo. En cambio, cuando el

## 4.5. Regresión espuria y cointegración

---

coeficiente de pendiente es distinto de cero, puede ser de interés validar la significancia de la relación mediante cointegración. Sin embargo, en cualquier caso la evaluación de la cointegración no es objetivo de este trabajo. Ahora, si se parte del conocimiento de que un conjunto de variables son cointegradas, la ventaja del modelo de regresión HDP-SLDS frente a un modelo estático, por ejemplo del tipo ECM<sup>9</sup>, es que permite tener estimaciones para cada punto en el tiempo; es decir, a diferencia de las propuestas de Siklos & Granger (1996); Balke & Fomby (1997) y; Sugita (2008), en las que se asumen dos estados del sistema (presencia o ausencia de cointegración), el modelo permite tener cualquier partición (finita) de la muestra y un modelo para cada una, o bien, modelos compartidos entre grupos de ellas.

### 4.5.2. Caso 3: Crecimiento, crédito bancario e inflación en México: un modelo de regresión lineal dinámica

Es común encontrar que las series de tiempo económicas sean no estacionarias; en estos casos, las técnicas clásicas de análisis de regresión darán estimaciones espurias. El tratamiento habitual es hacer las series estacionarias. Si las series no son integradas al mismo nivel, la aproximación de Pesaran et al. (2001) para probar cointegración y estimar las relaciones de la forma ECM permite a las variables tener distinto orden de integración entre ellas.

Utilizando el modelo de cointegración autoregresivo de rezagos distribuidos (ARDL, por sus siglas en inglés) propuesto por Pesaran et al. (2001), Shah et al. (2012) analizan la cointegración y las relaciones de corto y largo plazo entre el índice de mercado Karachi de Pakistan y un conjunto de variables macroeconómicas (inflación, tipo de cambio y tasa de interés) de 2003 a 2009. En un trabajo similar, Tinoco Zermeño et al. (2014) exploran los efectos de la inflación, el crédito privado y el desarrollo financiero en el crecimiento económico de México, medido por el producto interno bruto, durante 1969-2011. De manera concreta, uno de los objetivos del trabajo de Tinoco Zermeño et al. (2014) es investigar el efecto de la inflación en las bajas tasas de crédito bancario al sector privado, y su relación de largo plazo con el crecimiento económico. La hipótesis de los autores es que el crédito bancario al sector privado es un factor que explica las bajas tasas de inversión del país y, por tanto, de crecimiento económico. La estimación del

---

<sup>9</sup>Aunque es común en la literatura hacer referencia a los modelos ECM como modelos *dinámicos*, el término está asociado a que el efecto de las covariables difiere en el corto y largo plazo. Sin embargo, los coeficientes de pendiente son globales en el sentido de que no son dependientes del tiempo. En cambio, en un modelo dinámico del tipo LDS los estados evolucionan con el tiempo.

#### 4.5. Regresión espuria y cointegración

modelo ARDL se realiza en dos etapas: (1) verificar el número óptimo de rezagos para la primera diferencia de las variables; (2) probar la existencia de cointegración. Se estiman cuatro modelos para investigar cualquier efecto negativo de la inflación sobre el crecimiento. Todos los modelos incorporan como covariables: el gasto en consumo del gobierno, la formación bruta de capital fijo, las exportaciones de bienes y servicios, la inflación, y una variable dummy de liberalización financiera. Adicionalmente, cada uno incluye una covariable distinta (préstamos al sector privado, activos bancarios, pasivos líquidos, e índice de desarrollo financiero, construido por los autores usando análisis de componentes principales) y su interacción con la inflación. La Ec. (4.27) reproduce el modelo que incluye los préstamos al sector privado y su interacción con la inflación.

$$\begin{aligned}
 \Delta PIB_t = & \alpha_0 + \sum_{j=0}^p \beta_j \Delta PIB_{t-j} + \sum_{j=0}^p \gamma_j \Delta GOB_{t-j} + \sum_{j=0}^p \phi_j \Delta XP_{t-j} \\
 & + \sum_{j=0}^p \zeta_j \Delta INV_{t-j} + \sum_{j=0}^p \eta_j \Delta INF_{t-j} + \sum_{j=0}^p \varphi_j \Delta CP_{t-j} \\
 & + \sum_{j=0}^p \psi_j \Delta INF \times CP_{t-j} + \sigma_1 PIB_{t-1} + \sigma_2 GOB_{t-1} \\
 & + \sigma_3 XP_{t-1} + \sigma_4 INV_{t-1} + \sigma_5 INF_{t-1} + \sigma_6 CP_{t-1} \\
 & + \sigma_7 INF \times CP_{t-1} + u_t
 \end{aligned} \tag{4.27}$$

donde:

*PIB*: logaritmo del producto interno bruto.

*GOB*: logaritmo del gasto en consumo del gobierno.

*INV*: logaritmo de formación bruta de capital fijo.

*XP*: logaritmo de las exportaciones de bienes y servicios.

*INF*: incremento anual promedio del índice de precios al consumidor.

*CP*: logaritmo del credito bancario al sector privado.

*LIB*: variable dummy que captura el proceso de liberalización.

La variable *LIB* captura el proceso de liberalización, de manera que  $LIB_t = 1$  si  $t \leq 1987$ , y  $LIB_t = 0$  si  $t > 1987$ ; el resto de las series de datos tienen como año base 2005.  $p$  es la medida óptima del rezago,  $\Delta$  indica primera diferencia de

## 4.5. Regresión espuria y cointegración

---

las variables, y  $u_t$  es el término de error. Los coeficientes  $\beta_j$ ,  $\gamma_j$ ,  $\phi_j$ ,  $\zeta_j$ ,  $\eta_j$ ,  $\varphi_j$ , y  $\psi_j$  representan la dinámica de corto plazo de las variables, mientras que los coeficientes  $\sigma_i$ ,  $i = 1, \dots, 7$ , representan la dinámica de largo plazo.

Con el objetivo de comparar el modelo de regresión [HDP-SLDS](#) frente al modelo ECM, considere la siguiente ecuación:

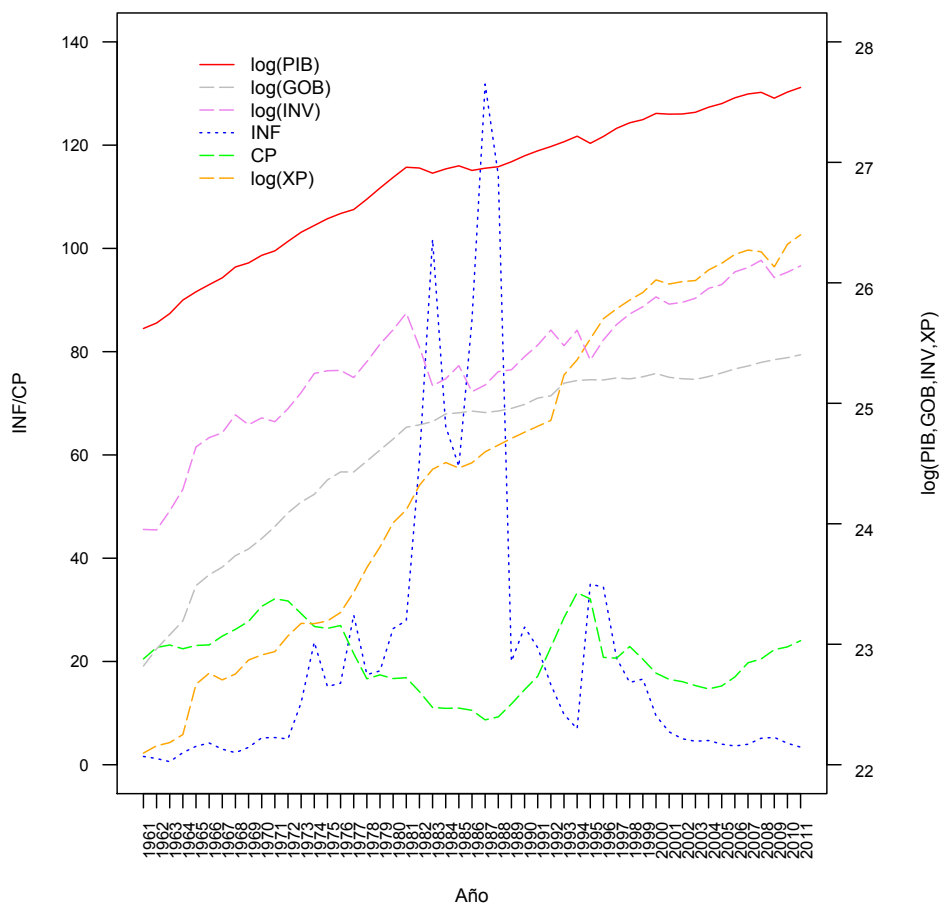
$$\begin{aligned} PIB_t = & \beta_{0t} + \beta_{1t}XP_t + \beta_{2t}GOB_t + \beta_{3t}INV_t + \beta_{4t}CP_t \\ & + \beta_{5t}INF_t + \beta_{6t}INF \times CP_t + \beta_{7t}LIB_t + u_t \end{aligned} \quad (4.28)$$

donde el significado de las variables involucradas es el mismo dado para la Ec. (4.27). Note que el modelo (4.28) es más parsimonioso que el (4.27), no obstante que los coeficientes de pendiente están indexados por  $t$ .

En la propuesta de [Tinoco Zermeño et al. \(2014\)](#), la dinámica de los coeficientes de pendiente se captura mediante las estimaciones de corto y largo plazo de un modelo ECM (ver sección 4.4.1). La dinámica de los coeficientes de la Ec. (4.28) se modela como en (4.6), de manera que la evolución se captura teniendo una estimación para cada punto  $t$  del tiempo. Los valores de los hiperparámetros asumidos para las distribuciones a priori del modelo fueron como sigue:  $K = 30$  como el límite de truncamiento de la distribución de transición;  $n_0 = 9$  grados de libertad y  $S_0 = I_9$  (matriz identidad de tamaño  $9 \times 9$ ) para la varianza del error del proceso;  $r_a = r_b = 10$  (parámetros de forma y escala) de una distribución  $\text{Ga}(r_a, r_b)$  para la precisión del error de observación  $y$ ;  $a = 1$  (parámetro de forma) y  $b = 100$  (parámetro de escala) de una distribución  $\text{Ga}(a, b)$  para el parámetro de precisión  $\alpha_j^{(k)}$ .

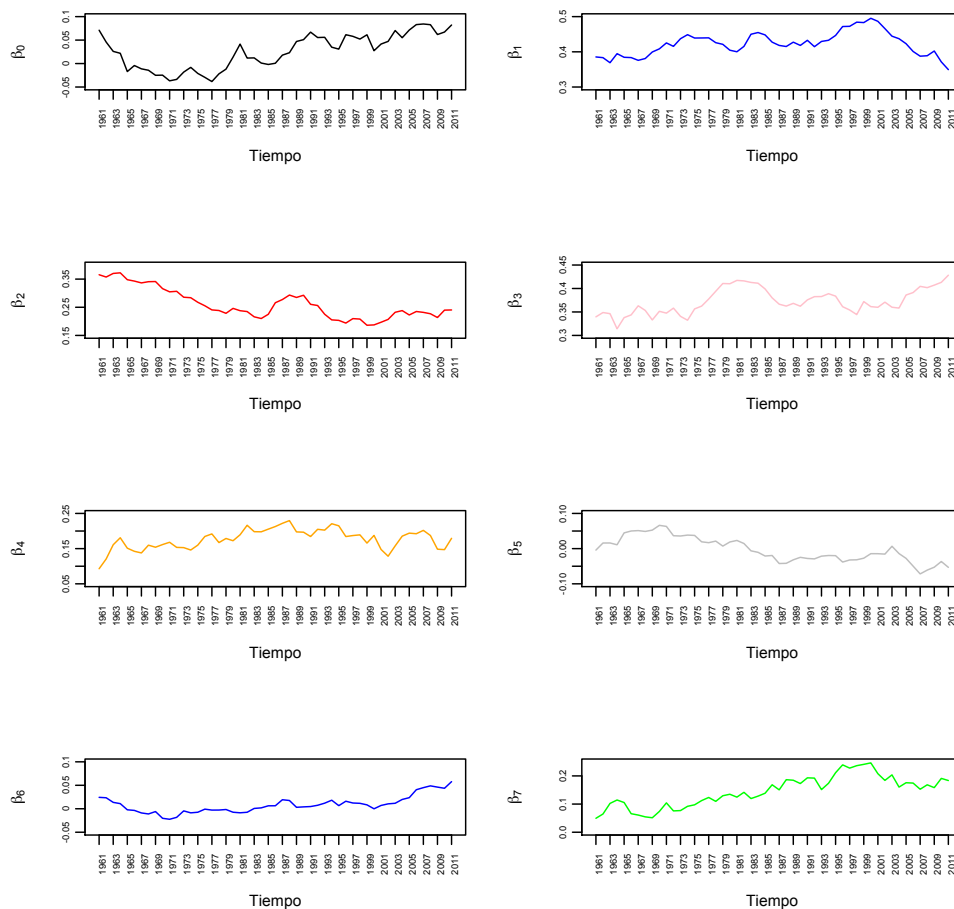
El Algoritmo 3 (sección 4.4) se iteró 30,000 veces. Los resultados que se reportan se obtuvieron como el promedio de 4,001 muestras obtenidas tomando 1 de cada 5 iteraciones de la 10,000 a la 30,000. La información usada fue obtenida de la base de datos *World Development Indicators* del Banco Mundial. La diferencia con la información usada por [Tinoco Zermeño et al. \(2014\)](#) es que se incluyen 8 años más a la serie (ver Figura 4.22): 1961-2011 vs. 1969-2011. La razón de esto es que hay disponibilidad de la información. La Figura 4.23 muestra la evolución de los coeficientes de pendiente estimados; los valores se comparan en seguida de la figura con los coeficientes de largo plazo estimados por [Tinoco Zermeño et al. \(2014\)](#), que se muestran en paréntesis y son llamados nivel de referencia.

## 4.5. Regresión espuria y cointegración



**Figura 4.22:**  $\log(PIB, GOB, INV, XP), INF, CP$ : México, 1961-2011.

## 4.5. Regresión espuria y cointegración



**Figura 4.23:** Estimaciones dinámicas de los coeficientes de pendiente  $\beta_{jt}$ .

### Exportaciones

(0.203). El coeficiente de pendiente estimado asociado a las exportaciones se encuentra por encima del nivel de referencia. Las principales características de la dinámica del coeficiente es la tendencia decreciente a partir de los últimos años de la década de los 80, coincidente con el declive en los precios del petróleo, y un breve crecimiento después la entrada en vigor del TLCAN.

## 4.5. Regresión espuria y cointegración

---

### Gasto en consumo del gobierno

(0.352). El coeficiente de pendiente estimado asociado al gasto público oscila entre el nivel de referencia; sin embargo, el promedio es superior hasta la primera mitad de la década de los 80, y a partir de entonces comienza a descender. Esto se explica por la función que el gobierno, a través del gasto público, ha tenido en la economía mexicana en distintos periodos, como mecanismo de ajuste de la actividad económica. Durante la mayor parte de la década de los 80, el gasto público tuvo una recomposición, producto de la crisis económica de 1982, reduciendo la proporción del gasto programable (gasto corriente y de capital) y aumentando el gasto no programable (intereses, amortizaciones y participación en la deuda pública), por lo que los efectos del gasto público en la economía se constituyeron como no productivos. Adicionalmente, como parte del programa de austeridad fiscal introducido en 1986 con el fin de estabilizar la economía, las autoridades gubernamentales redujeron el gasto, particularmente en el periodo 1990-1999, de manera que la participación del gasto público total en el producto interno bruto pasó de un promedio anual de 36.58 % en la década de los ochenta a una participación promedio anual de 21.93 % en la década de los noventa, y de 22.26 % en la primera década del siglo XXI ([Hernández Mota, 2011](#)).

### Inversión

(-0.009). Una de las principales diferencias en las estimaciones de los coeficientes de pendiente en los dos modelos es la asociada a la inversión. En el trabajo de [Tinoco Zermeño et al. \(2014\)](#) el coeficiente resultó no significativo y con signo contrario al esperado. En cambio, las estimaciones con el modelo de regresión [HDP-SLDS](#) oscilan entre 0.35 y 0.5. La formación bruta de capital fijo (FBCF) es un componente fundamental de la inversión productiva, y está vinculada positivamente al crecimiento económico de un país. Los países con menor FBCF respecto al PIB tienen menores tasas de crecimiento (ver [Góngora Pérez, 2012](#)).

### Crédito bancario al sector privado

(0.260). El coeficiente de pendiente estimado asociado al crédito privado se encuentra por debajo del nivel de referencia. La dinámica mantiene un nivel más o menos constante, pero destaca la disminución en el periodo 1977-1985. Sin embargo, como señala [Tinoco Zermeño et al. \(2014\)](#), los resultados sugieren que la disponibilidad de crédito bancario al sector privado en la economía ejerce un

## 4.5. Regresión espuria y cointegración

---

impacto positivo sobre el producto interno bruto real.

### Inflación

(-0.099). La dinámica del coeficiente de pendiente estimado asociado a la medida de la inflación se puede dividir en dos periodos: hasta antes de 1980 el valor del coeficiente es positivo, y a partir de ese año el coeficiente es negativo. No hay consenso en la literatura sobre el efecto del incremento en los precios sobre el PIB de un país. Hay argumentos a favor de que la inflación es necesaria para el crecimiento económico (ver por ejemplo [Mallik & Chowdhury, 2001](#)), mientras que los opositores prueban una relación negativa directa o indirecta (por ejemplo, [Fisher & Modigliani, 1978](#); [Clark, 1982](#); [Smyth, 1995](#)). Otros estudios sugieren que el efecto de la inflación sobre el producto es negativo pero sólo en el corto plazo, mientras que en el largo plazo no tiene un efecto real (ver [Faria & Carneiro, 2001](#)), o bien, que el efecto de la inflación sobre el crecimiento es negativo y estadísticamente significativo sólo si el nivel de inflación está por encima de un umbral (el umbral encontrado por [Risso & Carrera, 2009](#) para el caso de la economía mexicana en el periodo 1970-2007 es de 9%). Por otro lado, [Ghosh & Phillips \(1998\)](#); [Paul et al. \(1997\)](#) discuten que no hay relación causal entre la inflación y el crecimiento económico.

La Figura 4.23 muestra el coeficiente estimado entre -0.05 y 0.05, es decir, un aumento de 1% en la inflación altera el *PIB* en  $\pm 0.05\%$ . El cambio de signo coincide con el señalamiento de [Kilic & Arica \(2014\)](#) de que la dirección de la relación inflación-crecimiento económico es sensible al periodo de tiempo estudiado; hasta la década de 1970, cuando la inflación no era considerada un problema muy serio, los estudios empíricos indican un efecto positivo de la inflación sobre el crecimiento económico, pero a partir de 1980 se encuentran efectos negativos de esa relación. En el caso de México, después de la crisis de deuda de 1982 la economía mostró tasas crecientes de inflación, estancamiento y severa devaluación. En 1987 la inflación superó 130%. Ante esta situación se adoptó un programa de estabilización que logró reducir la inflación gradualmente (ver [Cárdenas, 1996](#)). A partir de entonces, y particularmente en la última década del periodo de análisis, uno de los objetivos de política monetaria ha sido mantener bajas tasas de inflación<sup>10</sup>.

---

<sup>10</sup>En 1999 el banco central anunció como objetivo mantener niveles de inflación por debajo de 2 dígitos, y a partir de 2002 alcanzar una inflación anual de 3%, con un intervalo de  $\pm 1\%$  (<http://banxico.org.mx>).



## 4.5. Regresión espuria y cointegración

---

### Inflación y crédito privado

(-0.07 %) El valor del coeficiente estimado para el término de interacción entre la inflación y el crédito privado es semejante al asociado para la inflación: hasta 1982 el valor es positivo, y a partir de ese año es negativo. Como señalan [Tinoco Zermeño et al. \(2014\)](#), un incremento en la inflación está asociado con una disminución en el producto a través de su efecto en el crédito bancario al sector privado. Sin embargo, la evidencia empírica sugiere que el efecto es negativo cuando se presentan periodos de alta inflación y se adoptan políticas de control de inflación.

### Liberalización financiera

(0.178). El coeficiente asociado a la variable dummy que captura el proceso de liberalización financiera oscila entre (0.05, 0.25), intervalo que cubre el valor estimado por [Tinoco Zermeño et al. \(2014\)](#). La categoría de referencia usada para esta variable es la eliminación de represión financiera, es decir, liberalización. Por tanto, el coeficiente sugiere que la represión está positivamente asociada al crecimiento, contrario a la interpretación que dan [Tinoco Zermeño et al. \(2014\)](#), y al signo que se esperaría para este coeficiente. Más aún, el incremento medio del *PIB* durante el periodo de represión financiera ha sido mayor al incremento medio después de la liberalización, lo que sugiere que las políticas de liberalización financiera por sí solas no contribuyen a estimular el crecimiento económico.

### Ajuste

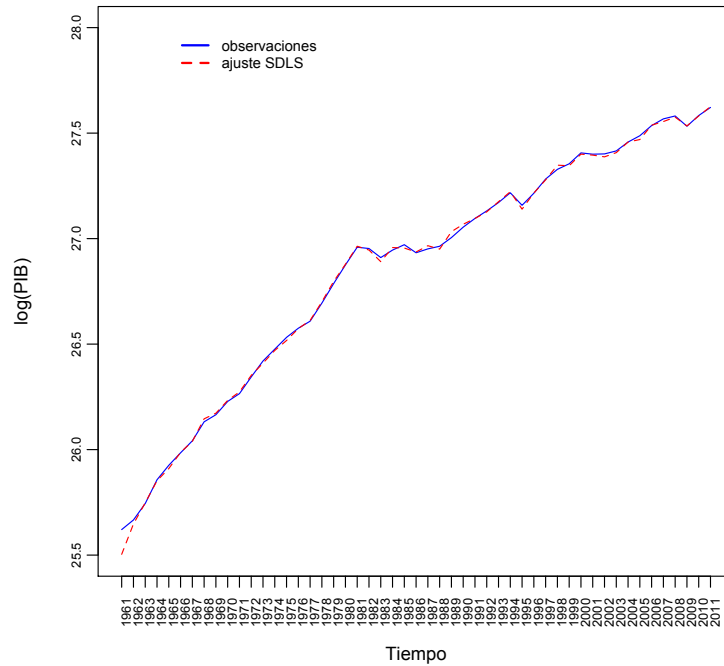
La Figura 4.24 muestra la serie del log  $PIB_t$  (línea continua azul) y el ajuste (línea punteada roja), obtenido como:

$$\begin{aligned} P\hat{I}B_t = & \hat{\beta}_{0t} + \hat{\beta}_{1t}XP_t + \hat{\beta}_{2t}GOB_t + \hat{\beta}_{3t}INV_t + \hat{\beta}_{4t}CP_t \\ & + \hat{\beta}_{5t}INF_t + \hat{\beta}_{6t}INF \times CP_t + \hat{\beta}_{7t}LIB_t \end{aligned} \quad (4.29)$$

donde:  $\hat{\beta}_{jt}$ ,  $j = 0, 1, \dots, 7$ ;  $t = 1961, \dots, 2011$ , son los valores de la Figura 4.23. [Tinoco Zermeño et al. \(2014\)](#) reportan el coeficiente de determinación  $R^2(\text{ajustada}) = 0.791$  para el modelo dado en (4.27). Calculando el mismo coeficiente de una regresión con los valores observados vs. los valores ajustados por (4.29) se obtiene:  $R^2(\text{ajustada}) = 0.9989$ . Aunque el modelo (4.28) propuesto para los datos del *PIB* no identificó cambios de *modo*, las estimaciones obtenidas proporcionan buen ajuste para la serie.

## 4.6. Conclusiones

---



**Figura 4.24:** Ajuste para los datos del logaritmo del producto interno bruto (PIB), 1961-2011.

## 4.6. Conclusiones

El capítulo desarrolla una propuesta de modelos de regresión dinámica para series de tiempo que identifica cambios de *modo* en una serie con alta probabilidad de persistencia en un mismo modo. Con datos simulados, hay evidencia de que el modelo identifica aceptablemente los modos; sin embargo, para una evaluación más precisa se debe buscar una alternativa de solución al inconveniente generado por el *label switching problem*. La estimación del vector de estados, es decir, de los coeficientes de pendiente en análisis de regresión que cuantifican el efecto que las covariables asociadas tienen sobre las observaciones, tiene una trayectoria similar a los valores usados para simulación, y aunque hay discrepancia entre las dos series, debida a que las diagonales de las matrices de evolución se forzaron a tener valores fijos, el ajuste de las observaciones es casi perfecto.

Adicional a la evaluación con datos simulados, la propuesta fue ilustrada con tres casos prácticos. El primero trata un tema que actualmente ha sido objeto

## 4.6. Conclusiones

---

de especial atención en el ámbito económico por la reciente evolución que ha mostrado: la dinámica diaria del tipo de cambio en México. La principal diferencia con estudios similares hasta ahora encontrados en la literatura, como los *Markov switching models*, es que no se asume que el tipo de cambio transita sólo entre dos modos. El periodo en estudio fue 1970-2016. El Algoritmo identificó dos modos con cambios coincidentes con una serie de eventos económicos asociados a periodos de crisis, lo que sugiere que el tipo de cambio es vulnerable a la disminución de la actividad económica. Adicionalmente, identificó periodos de alta volatilidad más frecuentes después de 2008 y persistencia en el régimen de alta volatilidad a partir del segundo semestre de 2015, acompañada de una tendencia creciente en el nivel. El Algoritmo mostró muy buen ajuste para la serie.

El segundo caso investiga la relación entre niveles de ozono y la temperatura. Se usa información de dos estaciones de monitoreo de la Cd. de México. Las estimaciones de los coeficientes de pendiente coinciden con los resultados obtenidos por [Huerta et al. \(2004\)](#): la exposición a la contaminación ambiental cambia entre estaciones, y el efecto de la temperatura cambia con el tiempo pero no entre estaciones. El Algoritmo identificó sólo un modo para las series, tal como se esperaba para este ejemplo.

El último caso de estudio compara los resultados obtenidos de un modelo de corrección de error model propuesto por [Tinoco Zermeño et al. \(2014\)](#) que relaciona el crecimiento en México, medido por el producto interno bruto, con un conjunto de variables macroeconómicas (inflación, crédito bancario al sector privado, gasto en consumo del gobierno, formación bruta de capital fijo, exportaciones y una variable dummy que denota la liberalización financiera). Los coeficientes estimados con el modelo de regresión *HDP-SLDS* capturan la dinámica de evolución de las covariables en cada punto del tiempo. La dinámica es el resultado del efecto que la política económica tiene sobre las covariables y, por ende, sobre las observaciones. Esto contribuye a evaluar los alcances de esas políticas y a la toma de decisiones.

# Capítulo 5

## Selección de variables en **SLDS**

### 5.1. Introducción

El objetivo de la de variables es elegir el *mejor* subconjunto de un conjunto de posibles predictores. En términos de regresión, implica identificar las variables explicativas más informativas para obtener un modelo más parsimonioso e interpretable.

Cuando se trabaja con series de tiempo económicas, por ejemplo, es común asociar una variable objetivo a un subconjunto de variables explicativas dependiendo del interés del investigador, de manera que, si un modelo incluye todos los posibles subconjuntos, el tamaño del vector de predictores sería grande. Entonces, es importante determinar el modelo más simple, en términos de las variables más relevantes, que mejor explique a la variable objetivo. La hipótesis planteada en esta tesis es que una variable puede resultar relevante en un periodo, dada su interacción con algún evento no anticipado o que no está presente a lo largo de todo el tiempo en estudio, pero no en otros periodos. Esta presunción se basa en el carácter dinámico de las series de tiempo; comúnmente, las series de tiempo económicas están disponibles por grandes periodos que cubren diferentes eventos, como crisis, cambios políticos, y cambios en los mercados financieros, por mencionar algunos. Entonces, es razonable suponer que la dinámica de los parámetros de regresión obedece también a las consecuencias que esos eventos tienen sobre las variables explicativas.

Como ya se ha mencionado en los capítulos previos, los **LDS** son flexibles para incorporar información relevante para predecir el vector de *estados* conforme el

## 5.2. El modelo

---

tiempo evoluciona, y son usados para modelar series de tiempo con comportamientos irregulares. Los **SLDS** permiten explicar fenómenos aun más complejos, permitiendo *cambios* en el modelo a través del tiempo. En el contexto de regresión lineal, dichos *cambios* han sido expresados hasta ahora por distintos valores de los coeficientes de regresión.

El objetivo en este capítulo es incorporar la selección de variables para identificar las relevantes en cada uno de los modos. Esto implica que la ecuación de observaciones es específica a cada *modo* a través del error de medición (extensión del **LDS** desarrollada en el Capítulo 4) y de la matriz diseño. La selección de variables es entonces un elemento adicional para distinguir entre *modos*. El modelo resultante permite expresar la relación entre la variable respuesta y los predictores de una forma más simple y adecuada a cada periodo de tiempo.

El capítulo está organizado de la siguiente manera: la sección 5.2 describe la metodología propuesta para la selección de variables en un **SLDS**. En la sección 5.3 se evalúa el desempeño en ajuste del modelo propuesto mediante un estudio de simulación considerando tres escenarios, cada uno con distinta configuración de parámetros. Por último, en la sección 5.4 se resumen los resultados del capítulo.

## 5.2. El modelo

La propuesta se basa en el algoritmo de **Kuo & Mallick (1998)**, que introduce variables indicadoras en la ecuación de regresión para incorporar los  $2^p$  submodelos posibles (asumiendo que siempre se incluye un término de intercepto), donde  $p$  es el número de covariables en el modelo. Para desarrollar la extensión a los **SLDS**, se asume que  $\gamma_j^{(z_t)}$  es la variable indicadora con soporte en  $\{0, 1\}$ ; entonces, la ecuación para  $y_t$  está dada por:

$$y_t = \beta_0 + \sum_{j=1}^p \beta_{tj} \gamma_j^{(z_t)} x_{tj} + w_t^{(z_t)} \quad (5.1)$$

La modelación de las variables indicadoras  $\gamma_j^{(k)}$  se realiza con información de las observaciones  $y_t$  tales que  $z_t = k$ . El superíndice significa que la inclusión o no de una covariable es particular a cada *modo*. Los detalles se presentan más adelante. Como antes, se asume que  $w_t^{(z_t)} \sim N(0, R^{(z_t)})$ .

Para cada  $k$ , las  $\gamma_j^{(k)}$  se eligen independientemente una de otra con distribución

## 5.2. El modelo

Bernoulli,  $Be(p_j)$ ,  $j = 1, \dots, p$ . El parámetro  $p_j$  es la probabilidad de incluir la  $j$ -ésima covariable en el modelo. Cuando  $\gamma_j^{(k)} = 1$ , las  $x_{tj}$  tales que  $z_t = k$  se incluye en el modelo de regresión; si  $\gamma_j^{(k)} = 0$ , entonces esos valores se omiten del modelo de regresión. En ausencia de información a priori, se asume  $p_j = \frac{1}{2}$  para toda  $j$ , que implica igual probabilidad para todos los  $2^p$  posibles submodelos en cada *modo*.

En forma matricial, la Ec. (5.1) se puede escribir de la siguiente manera:

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

donde:  $X_t^{(z_t)'} = \{1 \ \gamma_1^{(z_t)} x_{t1} \ \dots \ \gamma_p^{(z_t)} x_{tp}\}$  y  $\boldsymbol{\beta}_t = \{\beta_{t0} \ \beta_{t1} \ \dots \ \beta_{tp}\}'$ . El modelo completo es el siguiente:

$$\begin{aligned} z_t | z_{t-1} &\sim \pi_{z_{t-1}} \\ \gamma_j^{(z_t)} &\sim Be(1, p_j^{(z_t)}) \\ \boldsymbol{\beta}_t &= A^{(z_t)} \boldsymbol{\beta}_{t-1} + \mathbf{e}_t^{(z_t)} \\ y_t &= X_t^{(z_t)'} \boldsymbol{\beta}_t + w_t^{(z_t)}. \end{aligned} \tag{5.2}$$

Como en el Capítulo 4, la variable latente  $z_t$ , que sigue un proceso de Markov discreto de primer orden, con distribución de transición  $\pi_{z_{t-1}}$ , denota el *modo* oculto al tiempo  $t$ ;  $\boldsymbol{\beta}_t$  es el vector de *estados* del LDS al tiempo  $t$ , y  $\mathbf{e}_t \sim N(0, \Sigma^{(z_t)})$ . Dado  $\boldsymbol{\gamma}^{(z_t)} = \{\gamma_1^{(z_t)} \ \dots \ \gamma_p^{(z_t)}\}$ , la actualización de  $z_t$  y  $\boldsymbol{\beta}_t$  se realiza como se especifica en el Algoritmo 3, solamente reemplazando el vector  $X_t$  por  $X_t^{(z_t)}$ . Para cada  $k$ ,  $k = 1, \dots, K$ , donde  $K$  es el límite de truncamiento del HDP, hay  $2^p$  distintos vectores posibles  $\boldsymbol{\gamma}^{(k)}$ ; sin embargo,  $K$  y  $2^p$  no necesariamente son iguales. En consecuencia, el proceso de agrupación de las observaciones distingue los siguientes casos:

1. Si  $K > 2^p$  y el modelo distingue  $M$  *modos*, tal que  $M > 2^p$ , algunos de ellos compartirán vectores  $\boldsymbol{\gamma}^{(k)}$ , y la distinción entre estos estará dada solamente por los parámetros dinámicos del LDS.
2. Independientemente de la magnitud de  $K$  con respecto a  $2^p$ , si el modelo distingue  $M$  *modos*, tal que  $M \leq 2^p$ , entonces cada *modo* puede estar diferenciado por las covariables que determinan las observaciones, adicional a la distinción determinada por el SLDS. Sin embargo, puede ocurrir que dos o más *modos* compartan las mismas covariables; en este caso, la distinción entre ellos estará dada también sólo por los parámetros dinámicos del LDS.

## 5.2. El modelo

Para determinar la actualización de  $\gamma^{(k)}$  considere lo siguiente:

$$\begin{aligned} \vartheta_{tj}^{(z_t)} &= \beta_{tj} \gamma_j^{(z_t)}, \quad j = 0, 1, \dots, p \\ \gamma_0^{(k)} &= 1 \quad \text{para toda } k \in \{1, \dots, K\} \\ \boldsymbol{\vartheta}_t^{(z_t)} &= (\vartheta_{t0}^{(z_t)}, \vartheta_{t1}^{(z_t)}, \dots, \vartheta_{tp}^{(z_t)})' = (\beta_{t0}, \beta_{t1} \gamma_1^{(z_t)}, \dots, \beta_{tp} \gamma_p^{(z_t)})', \end{aligned}$$

tal que:

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

La distribución a priori para  $\vartheta_{tj}^{(z_t)}$  es una mezcla de una densidad normal con media  $A^{(z_t)} \beta_{t-1}$  y varianza  $\Sigma^{(z_t)}$  con probabilidad  $p_j$ , y masa en 0 con probabilidad  $(1-p_j)$ . Esta elección de la distribución a priori permite, con probabilidad positiva, hacer 0 los coeficientes de regresión (Kuo & Mallick, 1998). El método de selección de variables es entonces un proceso discreto, donde las covariables son retenidas o retiradas del modelo, y la actualización se toma de:

$$\gamma_j^{(k)} | \gamma_{-j}^{(k)}, \boldsymbol{\beta}^{(k)}, R^{(k)}, y_{1:T} \sim Be(1, \tilde{p}_j^{(k)}), \quad j = 1, \dots, p \quad (5.3)$$

donde  $\gamma_{-j}^{(k)} = (\gamma_1^{(k)}, \dots, \gamma_{j-1}^{(k)}, \gamma_{j+1}^{(k)}, \dots, \gamma_p^{(k)})$ ,  $\boldsymbol{\beta}^{(k)}$  contiene los coeficientes  $\beta_t$  tales que  $z_t = k$ , y:

$$\tilde{p}_j^{(k)} = c_j^{(k)} / (c_j^{(k)} + d_j^{(k)})$$

con:

$$c_j^{(k)} = p_j^{(k)} \exp \left\{ -\frac{1}{2} R^{(k)-1} \sum_{\{t: z_t=k\}} (y_t - X_t' \boldsymbol{\vartheta}_t^{(z_t)*})^2 \right\} \quad (5.4)$$

$$d_j^{(k)} = (1 - p_j^{(k)}) \exp \left\{ -\frac{1}{2} R^{(k)-1} \sum_{\{t: z_t=k\}} (y_t - X_t' \boldsymbol{\vartheta}_t^{(z_t)**})^2 \right\}. \quad (5.5)$$

En (5.4), el vector columna  $\boldsymbol{\vartheta}_t^{(z_t)*}$  es el vector  $\boldsymbol{\vartheta}_t^{(z_t)}$  con el  $j$ -ésimo elemento reemplazado por  $\beta_{tj}$ ; en (5.5), el vector  $\boldsymbol{\vartheta}_t^{(z_t)**}$  se obtiene de  $\boldsymbol{\vartheta}_t^{(z_t)}$  reemplazando el  $j$ -ésimo elemento por 0. Note que para la actualización de  $\gamma_j^{(k)}$ , el vector de covariables es  $X_t$ ; para la actualización de  $\beta_t$ , el vector de variables indicadoras  $\gamma^{(z_t)}$  es absorbido en el vector de covariables  $X_t^{(z_t)}$ . Cuando un *modo*  $k$  no tiene observaciones, entonces la actualización de cada  $\gamma_j^{(k)}$  correspondiente se toma de la a priori. El algoritmo 4 resume el muestreo Gibbs para inferencia del modelo (5.2).

### 5.3. Estudio de simulación

Algoritmo 4. Muestreo Gibbs para el modelo de regresión HDP-SLDS con selección de variables

1. Seguir los pasos del Algoritmo 3 de la Sección 4.4, solamente reemplazando  $X_t$  por  $X_t^{(z_t)}$  en donde corresponda.
2. Para cada  $k \in \{1, \dots, K\}$  y  $j = 1, \dots, p$ : Si  $\{t : z_t = k\} = \{\emptyset\}$ , entonces muestrear  $\gamma_j^{(k)}$  de  $B(p_j)$ ; de otro modo, actualizar  $\gamma_j^{(k)}$  de  $B(\tilde{p}_j^{(k)})$ , con  $\tilde{p}_j^{(k)} = c_j^{(k)} / (c_j^{(k)} + d_j^{(k)})$ , donde  $c_j^{(k)}$  y  $d_j^{(k)}$  se calculan como en (5.4) y (5.5), respectivamente.

### 5.3. Estudio de simulación

Para evaluar el desempeño en ajuste del modelo (5.2) se consideran 3 escenarios:

1.  $T = 500$  observaciones igualmente distribuidas en 2 modos; cada modo se observa en 2 intervalos de tiempo no adyacentes de igual longitud.
2.  $T = 500$  observaciones distribuidas aleatoriamente en 2 modos, con alta probabilidad de permanencia en un mismo modo.
3.  $T = 600$  observaciones igualmente distribuidas en 3 modos; cada modo se observa una vez (en un intervalo de tiempo).

Los resultados obtenidos con estos escenarios dan evidencia de la habilidad del modelo para aprender acerca del número de modos en la distribución y su localización, para estimar los coeficientes de regresión, y para identificar las variables importantes en cada modo.

En cada uno de los tres escenarios, los modos están definidos por el conjunto  $\{A^{(k)}, \Sigma^{(k)}, R^{(k)}, \gamma^{(k)}\}$ . El proceso de simulación de las observaciones en los escenarios 1 y 2 se realizó bajo la siguiente configuración de parámetros dinámicos:

1. Las matrices dinámicas  $A^{(k)}$ ,  $k = 1, 2$ , se definieron como

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



### 5.3. Estudio de simulación

---

2. Las matrices de covarianzas del error de evolución  $\Sigma^{(k)}$  se tomaron de la distribución a priori:  $\Sigma^{(k)} \sim \text{IW}(n_0, S_0^{(k)})$ , donde  $n_0 = p + 2$  grados de libertad,  $S_0^{(1)} = \mathbf{I}_p$  y  $S_0^{(2)} = 2\mathbf{I}_p$ .
3. La precisión del error observacional se seleccionó como:  $1/R^{(k)} \sim \text{Gam}(a_r, b_r)$ , donde  $a_r = 1$  (parámetro de forma) y  $b_r = 2$  (parámetro de escala).
4. La secuencia de modos  $z_{1:T}$  para el escenario 1 se estableció como se muestra en la Figura 5.1: las observaciones en los intervalos  $t \in [1, 125]$  y  $t \in [251, 375]$  pertenecen al modo 1, y las observaciones en  $t \in [126, 250]$  y  $t \in [376, 500]$  pertenecen al modo 2. La secuencia de modos para el escenario 2 se muestra en la Figura 5.5a.
5. El vector de variables indicadoras  $\gamma^{(k)}$ ,  $k = 1, 2$ , usado en cada modo en ambos escenarios, fue el siguiente:

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

donde el primer elemento de cada vector está asociado al intercepto, y los demás elementos están asociados a covariables.

Dados los parámetros dinámicos, la simulación de las observaciones en los tres escenarios se realiza de la siguiente manera:

#### Pasos para simular las observaciones

1. Para cada  $t = 1, \dots, T$  y  $j = 1, \dots, p$ , generar el valor  $x_{tj}$  de la matriz diseño como sigue:

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

donde  $v = 1$  en el escenario 1, y  $v = 0.9$  en los escenarios 2 y 3.

2. Para cada  $t = 1, \dots, T$ ,

a) Generar el vector de estados  $\beta_t$  de tamaño  $p$ :

$$\beta_t = A^{(z_t)}\beta_{t-1} + e_t, \quad e_t \sim \text{N}(\mathbf{0}, \Sigma^{(z_t)}), \quad \beta_0 = \mathbf{0}.$$

### 5.3. Estudio de simulación

---

b) Generar las observaciones  $y_t$  mediante

$$y_t = \sum_{j=0}^p \beta_{tj} \gamma_j^{(z_t)} x_{tj} + w_t, \quad w_t \sim N(0, R^{(z_t)}), \quad x_{t0} = 1.$$

Siguiendo el Algoritmo 4, se realizaron 20,000 iteraciones de cada escenario definiendo  $K = 30$  como nivel de truncación del [HDP](#), y usando ARD como distribución a priori para las columnas de la matriz  $A^{(k)}$ , IW para  $\Sigma^{(k)}$ , e IG para la varianza del error de medición, con la siguiente configuración inicial de hiperparámetros:

#### Valores iniciales de los hiperparámetros

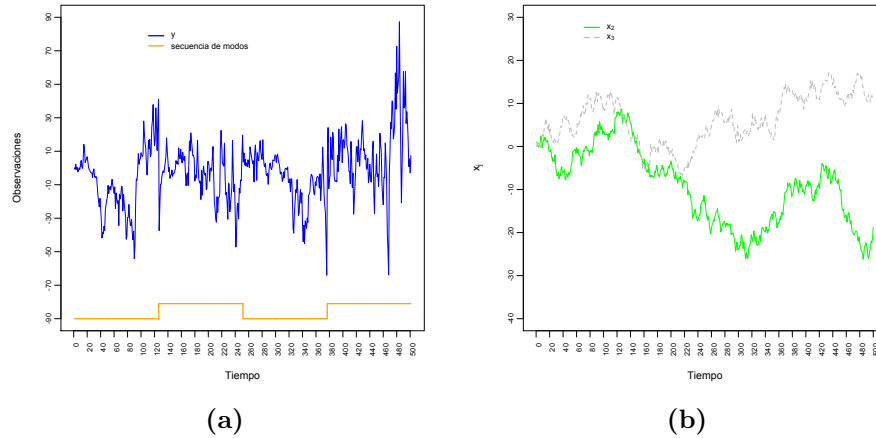
1. Se asume que todas las covariables están presentes en todos los modos, es decir,  $\gamma_j^{(k)} = 1$ , para toda  $j$  y toda  $k$ .
2. La transición del modo  $k$ ,  $k = 1, \dots, K$ , a cualquier otro modo, se asume igualmente probable; es decir, cada elemento de la matriz de probabilidades de transición es  $1/K$ :  $\pi_k = [1/K, \dots, 1/K] \forall k$ .
3. Cada elemento de  $z_{1:T}$  se toma independiente de una distribución categórica con probabilidad  $1/K$  para cada una de las  $K$  categorías.
4. Para cada  $k = 1, \dots, K$ :
  - a) Las columnas de  $A^{(k)}$  se seleccionan independientes de  $N(0, 1/\alpha_j^{(k)} I_p)$ , donde  $\alpha_j^{(k)} \sim \text{Gam}(1, 1/100)$ , tal que  $E(\alpha_j^{(k)}) = 100$ . Esta elección implica que inicialmente se asume una pequeña contribución de cada componente del vector de estados al sistema.
  - b)  $\Sigma^{(k)} \sim \text{IW}(p + 2, I_p)$ .
  - c) La precisión del error de medición  $1/R^{(k)}$  se toma de  $\text{Gam}(1, 1/2)$ , tal que  $E(1/R^{(k)}) = 2$ .
5. Para el *sticky* [HDP-HMM](#) implicado en la actualización del vector de modos  $z_{1:T}$ , se usa la misma configuración de hiperparámetros que en la Sección 4.4.3.

Los resultados que se presentan se calcularon descartando las primeras 5,000 iteraciones como *burn-in*.

### 5.3. Estudio de simulación

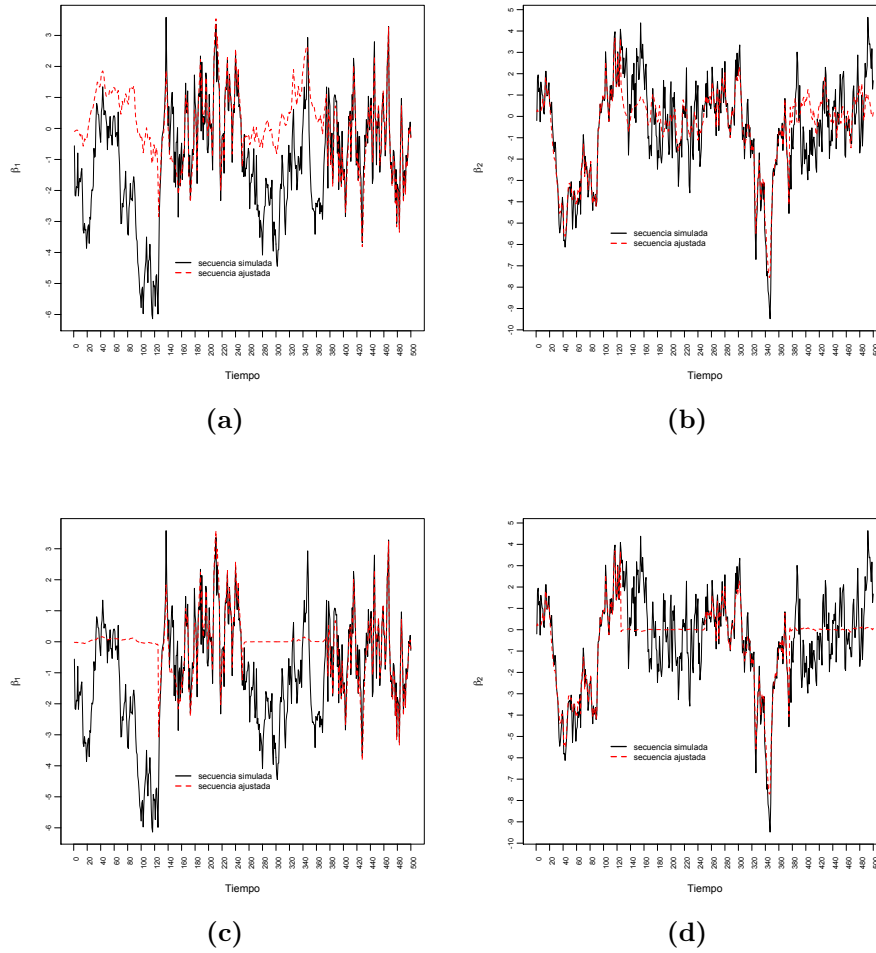
#### Escenario 1.

La Figura 5.1 muestra las observaciones, secuencia de modos y covariables simuladas. En la Figura 5.2 se presentan las secuencias simuladas y estimadas de los componentes del vector de estados: las Figuras 5.2a-5.2b corresponden a las gráficas de la secuencia de los componentes de  $\beta_t$  y  $\hat{\beta}_t$ , mientras que las Figuras 5.2c-5.2d corresponden a las gráficas de la secuencia de los componentes de  $\beta_t$  y  $\hat{\vartheta}_t^{(z_t)}$ . En las dos gráficas de la fila superior, las series estimadas siguen muy bien a las simuladas en los intervalos en los que la covariable asociada a cada componente está presente. Por ejemplo, en el modo 1 la covariable presente es  $x_{t2}$ , y la secuencia correspondiente a  $\hat{\beta}_{t2}$  ajusta bien en los intervalos  $t \in [1, 125]$  y  $t \in [251, 375]$ , no así en el resto de la serie. Las gráficas de la segunda fila dan evidencia de la capacidad del modelo para seleccionar las covariables importantes en cada modo. Como se esperaba, en los intervalos en los que la covariable asociada a cada componente está ausente,  $\hat{\vartheta}_{tj}^{(k)} = \hat{\beta}_{tj} \hat{\gamma}_j^{(k)}$  es cercano a cero.



**Figura 5.1:** Escenario 1. (a): observaciones y secuencia de modos; (b): covariables.

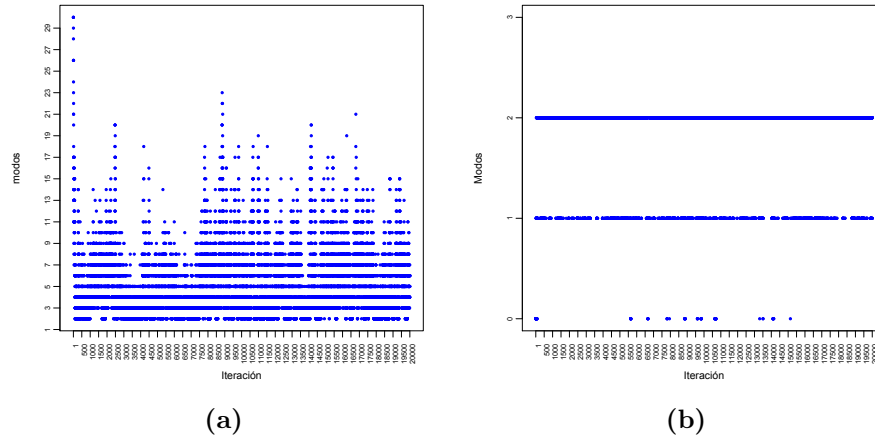
### 5.3. Estudio de simulación



**Figura 5.2:** Escenario 1. (a)  $\beta_{t1}, \hat{\beta}_{t1}$ ; (b)  $\beta_{t2}, \hat{\beta}_{t2}$ ; (c)  $\beta_{t1}, \hat{\vartheta}_{t1}^{(z_t)}$ ; (d)  $\beta_{t2}, \hat{\vartheta}_{t2}^{(z_t)}$ .

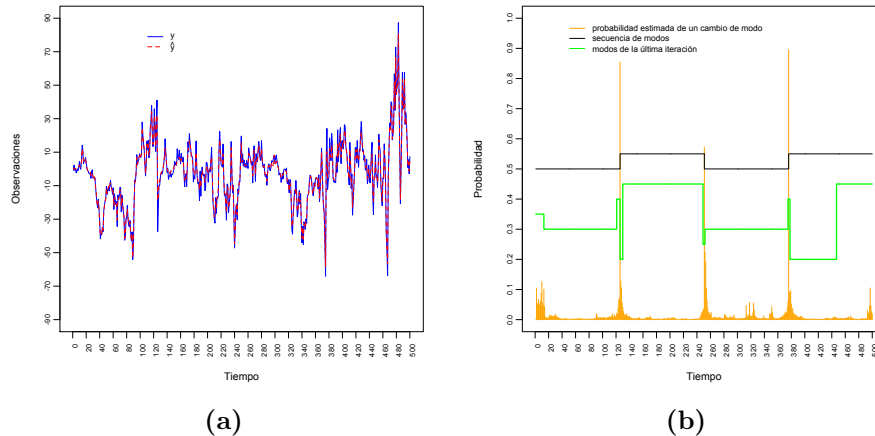
En la Figura 5.3a se grafica el número de modos en los que se agrupa a las observaciones en cada iteración. La figura no sugiere claramente que las observaciones se estén agrupando en algún número de modos. Un acercamiento mayor a los resultados se muestra en la Figura 5.3b; la gráfica corresponde al número de modos que concentra al menos 30% de las observaciones en cada iteración. Esta figura revela que los datos se agrupan principalmente en 2 modos.

### 5.3. Estudio de simulación



**Figura 5.3:** Escenario 1. (a): número de modos en cada iteración; (b): modos significativos.

Por último, en la gráfica 5.4a se muestra el ajuste de la serie de observaciones,  $\hat{y}_t = X_t' \hat{\vartheta}_t^{(z_t)}$ , y la gráfica 5.4b compara la secuencia de modos simulada con la obtenida de la última iteración. Esta gráfica confirma lo que se mencionó antes, el algoritmo identifica esencialmente dos modos. Se incluye además una estimación de la probabilidad de cambio de modo. Las probabilidades mayores coinciden con los cambios de la secuencia.

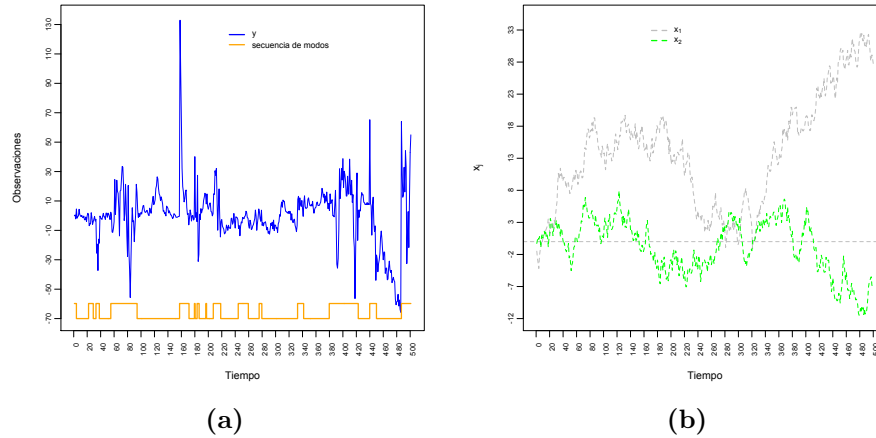


**Figura 5.4:** Escenario 1. Ajuste de observaciones, secuencia de modos y probabilidad de punto de cambio.

### 5.3. Estudio de simulación

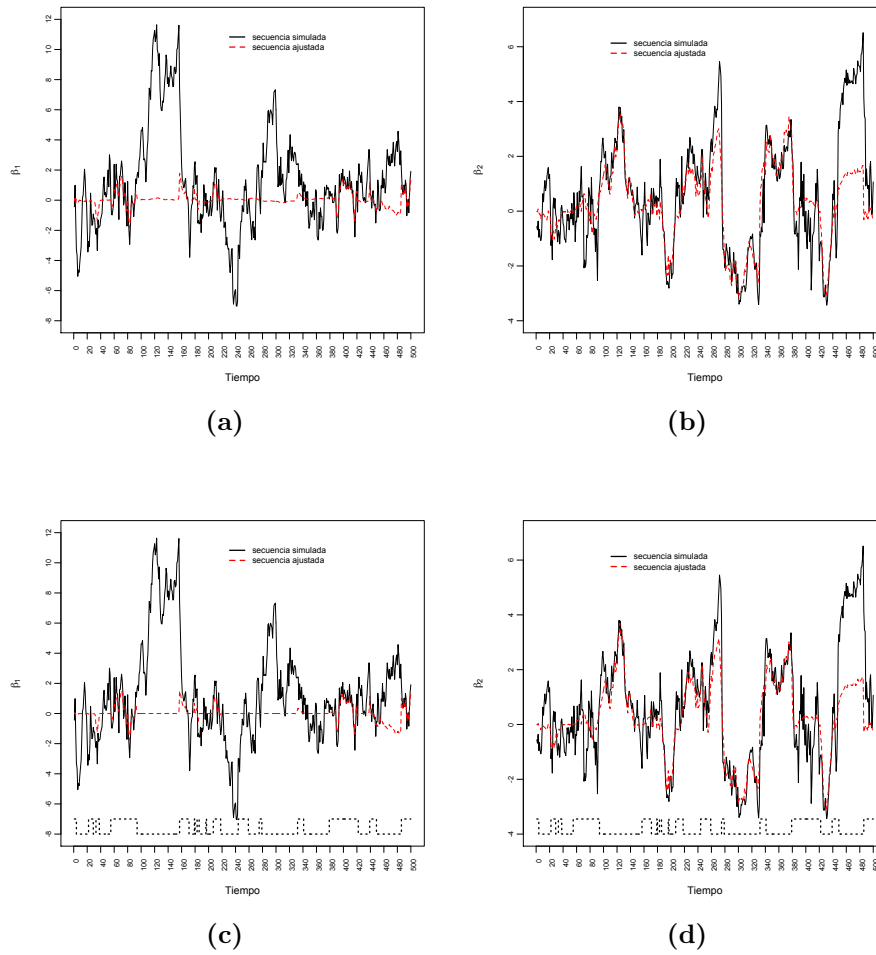
#### Escenario 2.

La Figura 5.5 presenta la serie de datos, secuencia de modos y covariables simuladas. Las gráficas de la Figura 5.6 muestran que el modelo es capaz de identificar las variables relevantes en cada modo, aún cuando alguno tenga pocas observaciones. Para ver esto más claramente, las gráficas 5.6c y 5.6d contienen la secuencia de modos simulada (línea negra punteada). Se puede apreciar que la secuencia ajustada es cero en los intervalos en los que la covariable asociada se simuló como no importante para el modo. Como consecuencia, las probabilidades más grades de un cambio de modo coinciden con los cambios *verdaderos* de la secuencia de modos, no obstante que hay subintervalos con baja persistencia temporal (Figura 5.8b); adicionalmente, las observaciones son clasificadas principalmente en dos modos (Figura 5.7). El resultado final es el buen ajuste de la serie (Figura 5.8a).



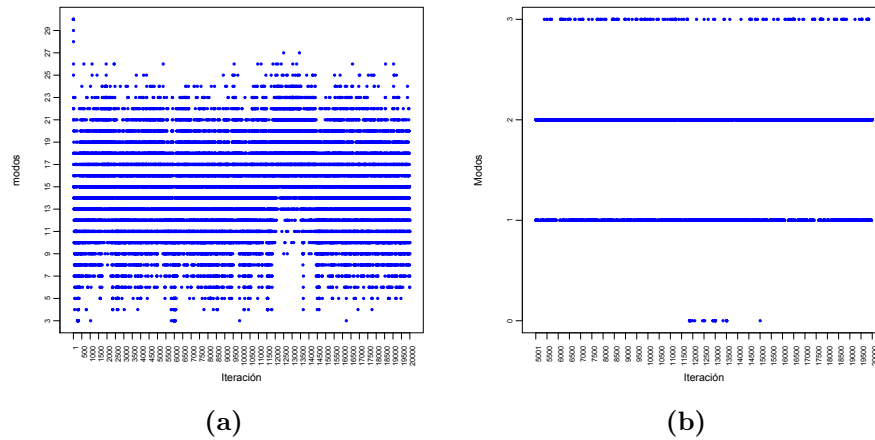
**Figura 5.5:** Escenario 2. (a): observaciones y secuencia de modos; (b): covariables.

### 5.3. Estudio de simulación

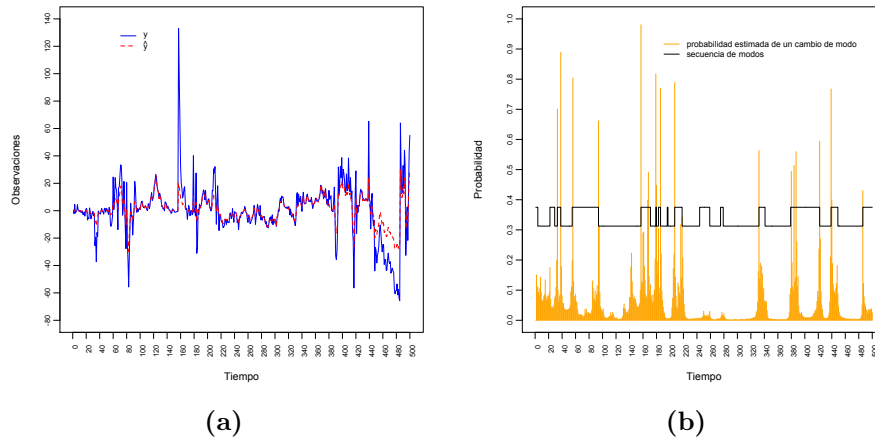


**Figura 5.6:** Escenario 2. (a)  $\beta_{t1}, \hat{\beta}_{t1}$ ; (b)  $\beta_{t2}, \hat{\beta}_{t2}$ ; (c)  $\beta_{t1}, \hat{v}_{t1}^{(z_t)}$ ; (d)  $\beta_{t2}, \hat{v}_{t2}^{(z_t)}$ .

### 5.3. Estudio de simulación



**Figura 5.7:** Escenario 2. (a): número de modos en cada iteración; (b): modos significativos.65.8 %



**Figura 5.8:** Escenario 2. Ajuste de observaciones, secuencia de modos y probabilidad de punto de cambio.

### Escenario 3.

Con el objetivo de investigar el comportamiento del modelo ante la presencia de más de 2 modos, se simuló una secuencia de 600 datos agrupados en 3 modos, con la siguiente configuración de parámetros dinámicos:



### 5.3. Estudio de simulación

---

1. Las matrices dinámicas  $A^{(k)}$ ,  $k = 1, 2, 3$ , se definieron como

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

2. Las matrices de covarianzas del error de evolución  $\Sigma^{(k)}$  se tomaron de la distribución a priori:  $\Sigma^{(k)} \sim \text{IW}(n_0, S_0^{(k)})$ , donde  $n_0 = p + 2$  grados de libertad,  $S_0^{(1)} = S_0^{(3)} = 2\mathbf{I}_p$ ,  $S_0^{(2)} = \mathbf{I}_p$ .
3. La varianza del error observacional se eligió aleatoriamente en cada modo como  $1/R^{(k)} \sim \text{Gam}(1, 0.5)$ , obteniendo:  $R^{(1)} = 4.2$ ,  $R^{(2)} = 2.2$  y  $R^{(3)} = 0.4$ .
4. La secuencia  $z_{1:T}$  se estableció con igual número de observaciones en cada modo. Las primeras 200 observaciones corresponden al modo 1, las siguientes 200 observaciones al modo 2, y las últimas al modo 3.
5. El vector de variables indicadoras  $\gamma^{(k)}$ ,  $k = 1, 2, 3$ , se eligió aleatoriamente, tal que  $\gamma_j^{(k)} \sim \text{Be}(1, p)$ , para  $j = 1, 2, 3$  independientes, con  $p \sim U(0, 1)$ . Los vectores resultaron como sigue

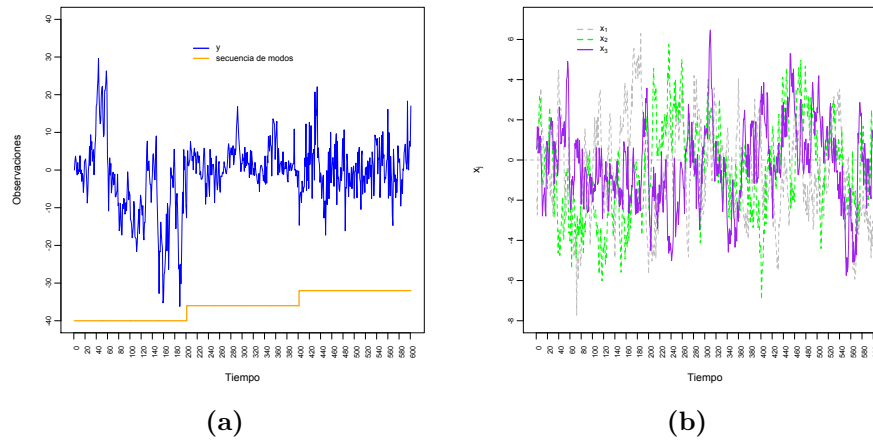
$$\begin{aligned} \gamma^{(1)} &= (1, 0, 1, 1) \\ \gamma^{(2)} &= (1, 0, 0, 1) \\ \gamma^{(3)} &= (1, 1, 1, 0), \end{aligned}$$

donde el primer elemento de cada vector indica el intercepto, y los demás elementos indican las covariables.

La Figura 5.9 muestra las observaciones, secuencia de modos y covariables simuladas. Note que sin información previa acerca de los modos a los que pertenecen las observaciones, no es fácil distinguir entre los modos 2 y 3. Las gráficas en la columna izquierda de la Figura 5.10 corresponden a las secuencias de los componentes de las pendientes en  $\beta_t$  y  $\hat{\beta}_t$ ; de forma similar, las gráficas de la columna derecha corresponden a las secuencias  $\beta_t$  y  $\hat{\vartheta}_t^{(z_t)}$ . En todos los casos, las secuencias tienen el ajuste esperado, los valores estimados siguen a los valores verdaderos en los intervalos en los que la covariable asociada al componente es significativa.

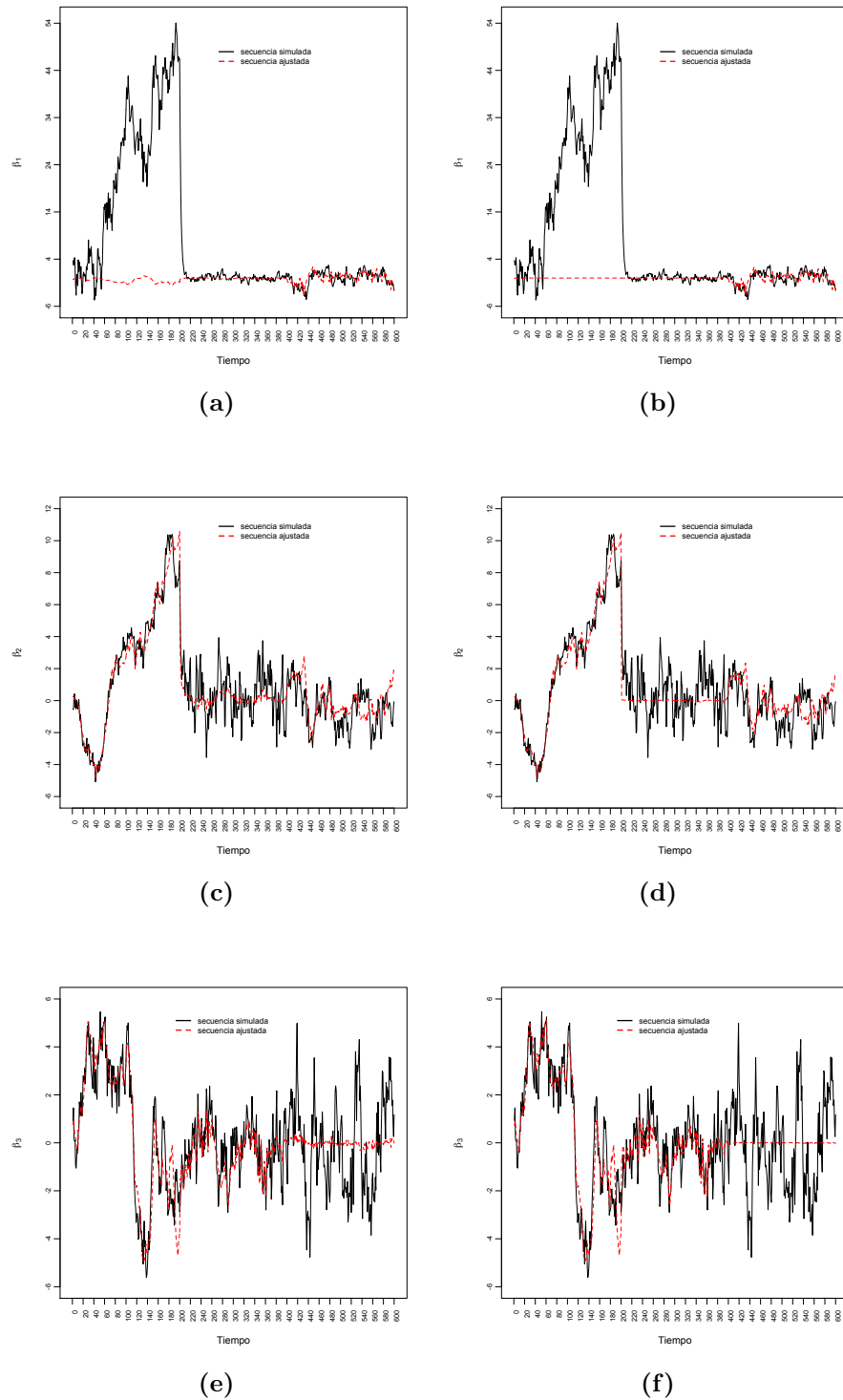
### 5.3. Estudio de simulación

---



**Figura 5.9:** Escenario 3. (a): observaciones y secuencia de modos; (b): covariables.

### 5.3. Estudio de simulación



**Figura 5.10:** Escenario 3. (a)  $\beta_{t1}, \hat{\beta}_{t1}$ ; (b)  $\beta_{t2}, \hat{\beta}_{t2}$ ; (c)  $\beta_{t3}, \hat{\beta}_{t3}$ ; (d)  $\beta_{t1}, \hat{v}_{t1}^{(z_t)}$ ; (e)  $\beta_{t2}, \hat{v}_{t2}^{(z_t)}$ ;  $\beta_{t3}, \hat{v}_{t3}^{(z_t)}$ .

### 5.3. Estudio de simulación

---

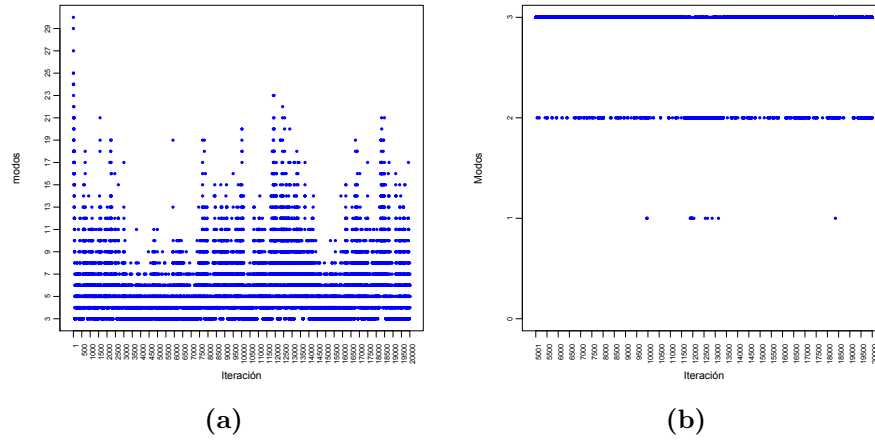
Como en los escenarios 1 y 2, la Figura 5.11a muestra el número de modos en los que se agrupan las observaciones en cada iteración, y la Figura 5.11b corresponde al número de modos significativos, con al menos 25% de las observaciones en cada uno. La gráfica 5.12b muestra la probabilidad estimada de cambio de modo para cada  $t$ . Se distinguen 2 puntos  $t$  con alta probabilidad que dividen la secuencia de datos en 3 intervalos. En una situación real en la que no se tiene conocimiento del número de modos y dónde se originan los cambios, no es posible conocer con esta gráfica qué intervalos corresponden a cada modo. El investigador debe entonces establecer sus conclusiones con base en la toda la información resultante. Los siguientes puntos son una guía para hacer inferencia sobre los objetivos del modelo: estimación de estados, número de modos en los que se agrupan los datos, variables significativas en cada modo, y qué intervalos corresponden a cada modo. Esos puntos resumen los resultados presentados para los tres escenarios.

- El número de modos en los que se agrupan los datos se infiere con la información usada para calcular el número de modos significativos. Para conocer qué intervalos corresponden a cada uno, se usa la estimación de las probabilidades de cambio de modo junto con la secuencia de modos de la última iteración. Las probabilidades estimadas sugieren en dónde se originan los cambios, y se espera que la última secuencia de modos generada corresponda a la agrupación encontrada después de convergencia.
- La estimación de la secuencia de estados, o de los coeficientes de regresión en el contexto de modelos de regresión, se infiere con la información de  $\hat{\boldsymbol{\vartheta}}_t^{(z_t)}$ .
- Las variables significativas en cada modo se infieren con los dos puntos anteriores. Específicamente, la estimación de  $\boldsymbol{\gamma}^{(k)}$  para cada  $k$  significativo, proporciona información sobre qué variables son relevantes en cada modo.

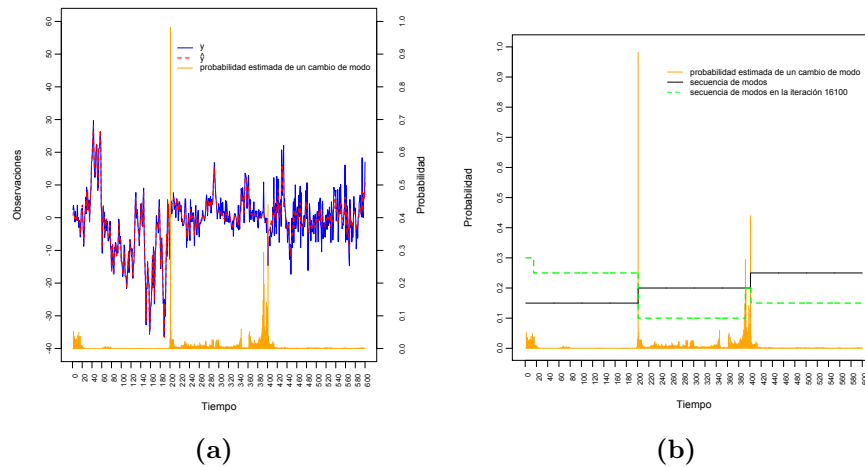
Aunque los puntos antes mencionados proporcionan una dirección para hacer conclusiones con el modelo (5.2), una parte importante en el reporte de resultados y toma de decisiones derivadas de éste, es la experiencia e información adicional que tenga el investigador.

Por último, la Figura 5.12a muestra el ajuste de la serie, calculada como en los escenarios anteriores:  $\hat{y}_t = X_t' \hat{\boldsymbol{\vartheta}}_t^{(z_t)}$ .

## 5.4. Conclusiones



**Figura 5.11:** Escenario 3. (a): número de modos en cada iteración; (b): modos significativos.



**Figura 5.12:** Escenario 3. Ajuste de observaciones, secuencia de modos y probabilidad de punto de cambio.

## 5.4. Conclusiones

El capítulo extiende el modelo [SLDS](#) de [Fox et al. \(2011a\)](#) para problemas de regresión con selección de variables. El modelo propuesto tiene buen desempeño para estimar el vector de estados o coeficientes de regresión, identificar el número

## 5.4. Conclusiones

---

de modos presentes en los datos, estimar los puntos en los que ocurren cambios de modo, e identificar las variables relevantes en cada modo. La selección de variables en los modelos dinámicos con cambios de régimen permiten describir relaciones entre variables más apropiadas conforme el tiempo evoluciona. Para evaluar el desempeño del modelo propuesto se utilizaron tres conjuntos de datos simulados con distinta configuración de los parámetros dinámicos y distinta persistencia en los modos. El modelo muestra buen desempeño aún cuando hay poca persistencia temporal en los modos.

Por otra parte, aunque en los resultados de la simulación presentados no se muestra evidencia del problema de *label switching*, es importante señalar que al introducir selección de variables en el [SLDS](#), este problema se incrementa considerablemente.

# Capítulo 6

## Conclusiones y trabajo futuro

La motivación principal en esta tesis es la evidencia de que muchos fenómenos complejos encontrados en la práctica no se pueden representar adecuadamente por una única distribución o modelo. Es decir, el supuesto de que se observan datos idénticamente distribuidos de un fenómeno no se sostiene con frecuencia. Una manera simple de tratar la presencia de datos heterogéneos en un conjunto de observaciones es mediante un modelo de mezclas. En un contexto clásico, los modelos de mezclas representan la heterogeneidad en un número finito y conocido de componentes en la mezcla; en un contexto Bayesiano, los procesos Dirichlet para modelos de mezclas (DPMM) permiten encontrar el número de componentes y los parámetros que las definen. En el Capítulo 2 se ilustran diversos métodos MCMC para muestrear de la distribución a posteriori de un DPMM; el muestreo Gibbs por bloques de Ishwaran & James (2001), que se basa en la representación *stick-breaking* truncada de un proceso Dirichlet (DP), tiene mejores propiedades para la mezcla.

La violación al supuesto de observaciones idénticamente distribuidas se hace más natural cuando se tienen series de tiempo. Intrínsecamente, los datos de series de tiempo son *dinámicos*, es decir, evolucionan en el tiempo. La metodología tradicional de análisis oculta la *dinámica* mediante transformaciones que producen una serie *estacionaria*, en el sentido de media y varianzas constantes en el tiempo. En cambio, los *modelos dinámicos* proporcionan flexibilidad para representar la dinámica que genera los datos observados, en términos de transiciones en el tiempo de variables latentes, o no observables, llamadas *estados*. En el Capítulo 3 se explican las principales características de los modelos de Markov ocultos, en los que la dinámica se describe en términos de transiciones de una variable aleatoria discreta, y de los sistemas dinámicos lineales (LDS), en los que la variable aleatoria

## 6. Conclusiones y trabajo futuro

---

que describe la dinámica es Gaussiana.

En muchas aplicaciones es de interés describir un fenómeno en términos de su relación con otras variables; en este contexto, los modelos de regresión representan una importante herramienta para el análisis. Típicamente, el investigador obtiene una medida global del efecto que tienen las variables explicativas, o covariables, sobre la variable respuesta. Sin embargo, el carácter dinámico de las series de tiempo demanda flexibilidad para permitir que dicho efecto evolucione también. La representación de los LDS como modelos de *espacio-estado* es consistente con los modelos de regresión dinámica en los que los coeficientes de regresión varían con el tiempo. Un LDS en su forma *espacio-estado* relaciona linealmente a las observaciones con un vector de parámetros de espacio continuo, los *estados*, que evoluciona como una cadena de Markov de primer orden. Los LDS son útiles para pronósticos de corto plazo, pero especial atención se debe dedicar a la magnitud de la varianza de observación con respecto a la varianza de los *estados*, a fin de controlar la incertidumbre en las predicciones. Una sección del Capítulo 3 se dedica a describir e ilustrar una sugerencia práctica, usando un *factor de descuento*, para determinar la magnitud de la varianza de los *estados*.

Series de tiempo más complejas, en las que la *dinámica* se relaciona con eventos que originan cambios estructurales en el tiempo, no se pueden describir adecuadamente por un solo LDS, pero se pueden aproximar mediante una secuencia de modelos de un conjunto de sistemas lineales. Cada modelo del conjunto está asociado a una configuración de parámetros dinámicos, y una variable latente, de espacio discreto, que indica cuál modelo es más apropiado para cada tiempo  $t$ . Cuando la variable latente es un proceso de Markov de tiempo discreto, el modelo se conoce como sistema dinámico lineal de cambio de régimen (SLDS). Debido al extenso interés por el uso de modelos que pretenden describir el comportamiento de una serie de tiempo mediante un conjunto de variables, en el Capítulo 4 se desarrolla una extensión de los SLDS para modelos de regresión que involucran variables explicativas. La propuesta se ilustra con tres estudios de caso.

Si bien la inclusión de covariables en un modelo de regresión puede contribuir a una mejor comprensión del sistema, no es deseable un modelo con demasiadas, o con redundantes, variables explicativas. Adicionalmente, el carácter dinámico de las series de tiempo puede originar que un modelo sea válido sólo en determinado periodo, según la relación que guarden las covariables con algún evento no presente a lo largo de todo el tiempo en estudio. El Capítulo 5 incorpora la selección de variables a los *modos* dinámicos como un criterio más para distinguir entre ellos; es decir, el subconjunto de covariables presentes en el modelo es particular a cada *modo*.



## 6.1. Trabajo futuro

---

Las propuestas desarrolladas en los capítulos 4 y 5, que constituyen las principales contribuciones de esta tesis, son todavía perfectibles. Las siguientes secciones puntualizan tres importantes tópicos pendientes para futura investigación: sobreajuste, convergencia y predicción.

## 6.1. Trabajo futuro

### 6.1.1. Sobreajuste

Uno de los intereses más comunes en la estadística práctica consiste en modelar un conjunto de datos de *entrenamiento*, de manera que se puedan hacer predicciones confiables sobre datos no observados. Si el modelo propuesto sobreajusta a los datos, lo más probable es que el modelo se adapta a las peculiaridades de la muestra y al ruido aleatorio en lugar de reflejar a la población en general. El sobreajuste se produce cuando un modelo es excesivamente complejo, por ejemplo, en un contexto clásico, cuando se tienen demasiados parámetros relativos al número de observaciones. Entonces, se busca encontrar el modelo más parsimonioso, esto es, el modelo más simple que describa satisfactoriamente a los datos. Pero un modelo nunca debe ser adoptado sólo porque es simple si no describe la asociación de interés adecuadamente.

Es bien conocido que un modelo que presenta sobreajuste tiene un pobre rendimiento predictivo, ya que reacciona exageradamente a pequeñas fluctuaciones en los datos de *entrenamiento*. En los [LDS](#) del capítulo 3, el sobreajuste de una serie de tiempo está relacionado con el valor relativamente grande de la varianza de la ecuación de evolución respecto a la varianza observacional; una varianza relativamente grande de los *estados* da mucho peso a las observaciones en la predicción. Para disminuir este efecto, una sugerencia práctica es especificar la varianza de los *estados* mediante un factor de descuento, un valor entre cero y uno que aumenta la varianza en la predicción en una proporción igual al inverso del factor de descuento, por lo que, entre mayor sea ese valor la incertidumbre del modelo es menor.

En los [SLDS](#) para problemas de regresión aún queda la tarea de tratar el sobreajuste, sobretodo si se desea hacer predicciones con el modelo. La dirección a este problema puede ser una extensión del factor de descuento en los [LDS](#). Sin embargo, la presencia de la variable latente que denota a los *modos* agrega complejidad al problema. Para ver esto, considere la varianza de la distribución a posteriori

## 6.1. Trabajo futuro

---

del vector de *estados*:

$$V(\beta_t|y_{1:t}) = \mathbf{F}_{t|t}^f = (\mathbf{X}_t'(R^{(z_t)})^{-1}\mathbf{X}_t + \mathbf{F}_{t-1,t}^{-1})^{-1},$$

donde  $\mathbf{X}_t$  es el vector de covariables al tiempo  $t$ ,  $R^{(z_t)}$  es la varianza observacional indexada por el *modo* al tiempo  $t$ , y

$$\mathbf{F}_{t-1,t} = \Sigma^{(z_t)} + A^{(z_t)}\mathbf{F}_{t-1|t-1}^f A^{(z_t)'}$$

es la varianza de la distribución predictiva de los *estados* dadas las observaciones, que depende de los parámetros dinámicos  $(\Sigma^{(z_t)}, A^{(z_t)})$  y de  $\mathbf{F}_{t-1|t-1}^f$ . Entonces, especificar  $\Sigma^{(z_t)}$  mediante un factor de descuento  $\delta$  implica considerar a este como otro parámetro dinámico indexándolo por los *modos*; o bien, establecer el factor de descuento como único para todo  $z_t$ . De cualquier manera, será necesaria la distribución conjunta de los parámetros dinámicos y  $\delta$ ; fijar un valor para el factor de descuento de manera equivalente a los [LDS](#) conlleva a no distinguir  $\Sigma$  por cada *modo*, pues la varianza de la distribución predictiva de los *estados* no es específica por *modo*.

### 6.1.2. Convergencia

Un elemento no considerado formalmente en esta tesis es un análisis de convergencia. En los capítulos 4 y 5 se llevaron a cabo estudios de simulación para demostrar que los [SLDS](#) tienen capacidad para inferir sobre el número de *modos* y los *estados* dinámicos del modelo en un contexto de regresión que involucra covariables. Sin embargo, en modelos de alta dimensionalidad como los desarrollados en esta tesis, el periodo de *burn-in* es difícil de cuantificar y evaluar.

### 6.1.3. Predicción

En muchas aplicaciones se tienen observaciones de una serie de tiempo para un cierto periodo, y se desea estudiar retrospectivamente el comportamiento del sistema que genera las observaciones. En términos de un [SLDS](#), esto consiste en estimar la secuencia de *estados* y *modos* para cada  $t$ ,  $t = 1, \dots, T$ , dada la secuencia de observaciones  $y_{1:T}$ , que es lo que se ha desarrollado en esta tesis. Sin embargo, con frecuencia el principal interés del análisis de series de tiempo es hacer pronósticos del futuro. En general, los pronósticos están basados en tres tipos de predictores: (1) predictores puntuales; (2) intervalos predictivos; (3) distribuciones predictivas. Las distribuciones predictivas son más informativas que los predictores puntuales

## 6.1. Trabajo futuro

---

o los intervalos predictivos, pues toda la información sobre el futuro está contenida en la distribución.

En el caso de modelos dinámicos lineales, la predicción de las observaciones futuras involucra la estimación del vector de *estados*; para el pronóstico de un punto adelante en el tiempo, primero se estima el siguiente valor del vector de *estados*, y entonces, con base en esta estimación, se calcula el pronóstico para la observación  $y_{t+1}$ . En general, cuando el interés es en el pronóstico de un punto  $k$ ,  $k \geq 1$ , más adelante en el tiempo, se estima el vector de estados en  $t + k$  con base en la densidad predictiva  $p(z_{t+k}|y_{1:t})$ , y se obtiene la densidad predictiva  $p(y_{t+k}|y_{1:t})$  para la observación futura. Pero la estimación de los *estados* es sólo un paso para predecir el valor de las observaciones futuras. Del capítulo 3,  $p(z_{t+k}|y_{1:t}) = N(f_{t,k-1}, S_{t+k})$  y  $p(y_{t+k}|y_{1:t}) = N(a_{t+k}, Q_{t+k})$ , donde:

$$\begin{aligned} f_{t,k-1} &= A_{t+k} \cdot A_{t+k-1} \cdots A_{t+1} f_t \\ S_{t+k} &= A_{t+k} F_{t,k-1} A'_{t+k} + \Sigma_{t+k} \\ a_{t+k} &= C_{t+k} f_{t,k-1} \\ Q_{t+k} &= R_t + C'_{t+k} S_{t+k} C_{t+k} \end{aligned}$$

Note que la predicción para  $y_{t+k}$  involucra también el conocimiento de la matriz diseño  $C$  en el tiempo  $t + k$ . Pero, en un contexto de regresión como el de interés en esta tesis, la matriz diseño  $C_{t+k}$  contiene valores de las covariables que no han sido observados, por lo que será necesario estimarlos.

Para un [SLDS](#), además de la matriz diseño  $C_{t+k}$ , la predicción del vector de *estados*  $\beta_{t+k}$  y de la observación  $y_{t+k}$  involucran la predicción del *modo*  $z_{t+k}$ , que a su vez involucra el conocimiento de *estados* y observaciones. Adicionalmente, para la predicción de un *modo* con  $k$  pequeño, es razonable restringir a que su valor sea uno de los ya observados; pero si  $k$  es grande, esto limitaría la dinámica futura del modelo al no permitir un nuevo *modo* no antes observado.

En general, cuando se desea hacer predicciones se debe considerar que conforme  $k$  sea mayor habrá más incertidumbre y el pronóstico será menos preciso. En la práctica, cuando los datos se observan en un periodo corto de tiempo, por ejemplo el tipo de cambio que tiene periodicidad diaria, resulta conveniente calcular el pronóstico un punto adelante en el tiempo, y actualizarlo secuencialmente conforme se dispone de una nueva observación.

# Referencias

- C. Alexander (2001). *Market Models: A Guide to Financial Data Analysis*. John Wiley and Sons.
- L. Angeles, et al. (1982). *La devaluación de 1982*. Editorial Terra Nova.
- C. E. Antoniak (1974). ‘Mixtures of Dirichlet Processes with applications to Bayesian Nonparametric Problems’. *The Annals of Statistics* **2**(6):1152–1174.
- N. S. Balke & T. B. Fomby (1997). ‘Threshold cointegration’. *International Economic Review* **38**(3):627–645.
- D. Barber (2012). *Bayesian reasoning and machine learning*. Cambridge University Press.
- C. F. Baum, et al. (2004). ‘Nonlinear effects of exchange rate volatility on the volume of bilateral exports’. *Journal of Applied Econometrics* **19**(1):1–23.
- S. Bazdresch & A. M. Werner (2002). ‘El comportamiento del tipo de cambio en México y el régimen de libre flotación: 1996-2001’. Tech. Rep. 2002-09, Banco de México.
- M. J. Beal, et al. (2002). ‘The infinite hidden Markov model’. In T. G. Dietterich, S. Becker, & Z. Ghahramani (eds.), *Neural Information Processing Systems 14*, vol. 14, pp. 577–584. MIT Press.
- M. A. Berger (1993). *An introduction to probability and stochastic processes*. Springer-Verlag.
- R. Bhar & S. Hamori (2004). *Hidden Markov models: applications to financial economics*. Kluwer Academic Publishers.
- C. Bishop (2006). *Pattern recognition and machine learning*. Springer.
- D. Blackwell & J. B. MacQueen (1973). ‘Ferguson Distributions via Polya Urn Schemes’. *The Annals of Statistics* **1**(2):353–355.

## REFERENCIAS

---

- D. M. Blei & M. I. Jordan (2006). ‘Variational Inference for Dirichlet Process Mixture’. *Bayesian Analysis* **1**(1):121–144.
- T. Bollerslev (1986). ‘Generalized autoregressive conditional heteroscedasticity’. *Journal of Econometrics* **31**:307–327.
- G. E. P. Box, et al. (2015). *Time Series Analysis: Forecasting and Control*. Wiley, 5 edn.
- C. Bregler (1997). ‘Learning and Recognizing Human Dynamics in Video Sequences’. In *IEEE Computer Society Conference. Computer Vision and Pattern Recognition*.
- R. J. Brenner & K. F. Kroner (1995). ‘Arbitrage, Cointegration, and Testing the Unbiasedness Hypothesis in Financial Markets’. *Journal of Financial and Quantitative Analysis* **30**(1):23–39.
- E. Cárdenas (1996). *La política económica en México, 1950-1994*. Fondo de Cultura Económica.
- F. Caron, et al. (2008). ‘Bayesian Inference for Linear Dynamic Models with Dirichlet Process Mixtures’. *IEEE Transactions on signal processing* **56**(1):71–84.
- C. Carter & R. Kohn (1994). ‘On Gibbs sampling for state space models’. *Biometrika* **81**(3):541–553.
- C. Carter & R. Kohn (1996). ‘Markov chain Monte Carlo in conditionally Gaussian state space models’. *Biometrika* **83**(3):589–601.
- C. M. Carvalho & H. F. Lopes (2007). ‘Simulation based sequential analysis of Markov switching stochastic volatility models’. *Computational Statistics and Data Analysis* **51**:4526–4542.
- W. H. Chan (2003). ‘A correlated bivariate Poisson jump model for foreign exchange’. *Empirical Economics* **28**(4):669–685.
- C. W. S. Chen, et al. (2011). ‘A comparison of estimators for regression models with change points’. *Statistics and Computing* **21**(3):395–414.
- P. K. Clark (1982). ‘Inflation and the productivity decline’. *American Economic Review* **72**(2):149–154.
- E. Conrad Lamon III, et al. (1998). ‘Forecasting PCB concentrations in lake Michigan Salmonids: a dynamic linear model approach’. *Ecological applications* **8**(3):659–668.

## REFERENCIAS

---

- J. Córdoba & G. Ortíz (1979). ‘Aspectos deflacionarios de la devaluación del peso mexicano de 1976’. Tech. Rep. 9, Banco de México.
- R. D. Core-Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- A. Crimmins, et al. (2016). *The Impacts of Climate Change on Human Health in the United States: A Scientific Assessment*. U.S. Global Change Research Program.
- L. Ding & M. Vo (2012). ‘Exchange rates and oil prices: A multivariate stochastic volatility analysis’. *The Quarterly Review of Economics and Finance* **52**(1):15–37.
- M. Diniz, et al. (2012). ‘Cointegration: Bayesian Significance Test’. *Communications in Statistics: Theory and Methods* **41**(19):3562–3574.
- B. L. E. & T. Petrie (1966). ‘Statistical Inference for Probabilistic Functions of Finite State Markov Chains’. *The Annals of Mathematical Statistics* **37**(6):1554–1563.
- R. G. Edmonds Jr. & J. Y. So (2004). ‘Is exchange rate volatility excessive? An ARCH and AR approach’. *The Quarterly of Economics and Finance* **44**(1):122–154.
- R. F. Engle (1982). ‘Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation’. *Econometrica* **50**(4):987–1007.
- R. F. Engle & C. W. J. Granger (1987). ‘Co-Integration and Error Correction: Representation, Estimation, and Testing’. *Econometrica* **55**(5):251–276.
- M. D. Escobar (1994). ‘Estimating Normal Means with a Dirichlet Process prior’. *Journal of the American Statistical Association* **89**(425):268–277.
- M. D. Escobar & M. West (1995). ‘Bayesian Density Estimation and Inference Using Mixtures’. *Journal of the American Statistical Association* **90**(430):577–588.
- B. S. Everitt & T. Hothorn (2010). *A Handbook of Statistical Analyses Using R*. Chaptan and Hall/CRC, 2nd edn.
- J. R. Faria & F. G. Carneiro (2001). ‘Does high inflation affect growth in the long and short run?’. *Applied Economics* **4**(1):89–105.
- P. Fearnhead (2005). ‘Exact bayesian curve fitting and signal segmentation’. *IEEE Transactions on signal processing* **53**(6):2160–2166.

## REFERENCIAS

---

- P. Fearnhead (2006). ‘Exact and efficient bayesian inference for multiple change-point problems’. *Statistics and computing* **16**(2):203–213.
- R. M. Feldman & C. Valdez-Flores (2010). *Applied probability and stochastic processes*. Springer, 2nd edn.
- T. S. Ferguson (1973). ‘A Bayesian Analysis of Some Nonparametric Problems’. *The Annals of Statistics* **1**(2):209–230.
- T. S. Ferguson (1983). ‘Bayesian Density Estimation by Mixtures of Normal Distributions’. *Recent Advances in Statistics* pp. 287–302.
- N. Fiess & R. Shankar (2009). ‘Determinants of exchange rate regime switching’. *Journal of International Money and Finance* **28**(1):68–98.
- S. Fisher & F. Modigliani (1978). ‘Towards and understanding of the real effects and costs of inflation’. *Review of World Economics* **114**(4):810–833.
- E. Fox, et al. (2011a). ‘Bayesian Nonparametric Inference of Switching Dynamic Linear Models’. *IEEE Transactions on signal processing* **59**(4):1569–1585.
- E. Fox, et al. (2011b). ‘A sticky HDP-HMM with application to speaker diarization’. *The Annals of Applied Statistics* **5**(2A):1020–1056.
- B. Friedland (1986). *Control system design. An introduction to state-space methods*. McGraw-Hill.
- B. A. Frigiyik, et al. (2010). ‘Introduction to the Dirichlet Distribution and Related Processes’. Tech. Rep. UWEETR-2010-0006, University of Washington.
- S. Fruhwirth-Schnatter (1994). ‘Data augmentation and dynamic linear models’. *Journal of Time Series Analysis* **15**(2):183–202.
- D. Galar Pascual (2015). *Artificial intelligence tools. Decision support systems in condition monitoring and diagnosis*. CRC Press.
- S. J. Gershman & D. M. Blei (2012). ‘A tutorial on Bayesian Nonparametric Models’. *Journal of Mathematical Psychology* **56**:1–12.
- Z. Ghahramani & G. E. Hinton (1998). ‘Variational learning for switching state-space models’. *Neural Computation* **12**(4):963–996.
- Z. Ghahramani & G. E. Hinton (2000). ‘Variational learning for switching state-space models’. *Neural Computation* **12**(4):831–864.
- A. Ghosh & S. Phillips (1998). ‘Inflation, disinflation and growth’. Tech. Rep. WP/98/68, International Monetary Fund.

## REFERENCIAS

---

- J. P. Góngora Pérez (2012). ‘La formación bruta de capital fijo en México’. *Comercio Exterior* **62**(6).
- D. Görür & C. E. Rasmussen (2010). ‘Dirichlet Process Gaussian Mixture Models: Choise of the Base Distribution’. *Journal of Computer Science and Technology* **25**(4):615–626.
- C. W. J. Granger (1981). ‘Some properties of time series data and their use in econometric model specification’. *Journal of Econometrics* **16**(1):121–130.
- J. D. Hamilton (1989). ‘A new approach to the economic analysis of nonstationary time series and the business cycle’. *Econometrica* **57**(2):357–384.
- J. D. Hamilton (1990). ‘Analysis of time series subject to changes in regime’. *Journal of Econometrics* **45**(1-2):39–70.
- K. M. Hangos, et al. (2001). *Intelligent Control Systems. An introduction with examples*. Kluwer academic publishers.
- F. Hayashi (2000). *Econometrics*. Princeton University Press.
- H. V. Henderson & S. R. Searle (1981). ‘On deriving the inverse of a sum of matrices’. *SIAM Review* **23**(1):53–60.
- J. L. Hernández Mota (2011). ‘Política macroeconómica y crecimiento económico: la experiencia mexicana’. *Economía Informa* (371):24–42.
- F. L. Herrera, et al. (2011). ‘Volatilidad estocástica del tipo de cambio peso-dólar: el régimen flotante en México’. *Investigación Económica* **70**(276):19–50.
- M. D. Hoffman, et al. (2008). ‘Data-driven recomposition using the hierarchical Dirichlet process hidden Markov model’. In *Proc. International Computer Music Conference*.
- Z. Huang (2007). ‘The central bank and speculators in the foreign exchange market under asymmetric information: A strategic approach and evidence’. *Journal of Economics and Business* **59**(1):28–50.
- G. Huerta, et al. (2004). ‘A spatiotemporal model for Mexico city ozone levels’. *Journal of the Royal Statistical Society* **53**(2):231–248.
- J. H. Hung (1997). ‘Intervention strategies and exchange rate volatility: a noise trading perspective’. *Journal of International Money and Finance* **16**(5):779–793.
- H. Ishwaran & L. F. James (2001). ‘Gibbs Sampling Methods for Stick-Breaking Priors’. *Journal of the American Statistical Association* **96**(453):161–173.



## REFERENCIAS

---

- H. Ishwaran & L. F. James (2002). ‘Approximate Dirichlet Process Computing in Finite Normal Mixtures: Smoothing and Prior Information’. *Journal of Computational and Graphical Statistics* **11**(3):1–26.
- H. Ishwaran & M. Zarepour (2002). ‘Exact and approximate sum representations for the Dirichlet process’. *The Canadian Journal of Statistics* **30**(2):269–283.
- S. Jain & R. M. Neal (2004). ‘A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model’. *Journal of Computational and Graphical Statistics* **13**(1):158–182.
- A. Jasra, et al. (2005). ‘Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling’. *Statistical Science* **20**(1):50–67.
- S. Johansen (1991). ‘Estimation and hypothesis testing of cointegrating vectors in Gaussian vector autoregressive models’. *Econometrica* **59**(6):1551–1580.
- N. Jouchi (2013). ‘Stochastic volatility model with regime-switching skewness in heavy-tailed errors for exchange rate returns’. *Studies in Nonlinear Dynamics and Econometrics* **17**(5):499–520.
- B. H. Juang & L. R. Rabiner (1991). ‘Hidden Markov models for speech recognition’. *Technometrics* **33**(3):251–272.
- M. Kalli, et al. (2011). ‘Slice sampling mixture models’. *Statistics and Computing* **21**(1):93–105.
- R. E. Kalman (1960). ‘A new approach to linear filtering and prediction problems’. *Journal of Basic Engineering* **82**:35–45.
- R. E. Kalman (1963). ‘Mathematical description of linear dynamical systems’. *Journal of the Society for Industrial and Applied Mathematics* **1**(2):152–192.
- C. Kilic & F. Arica (2014). ‘Economic freedom, inflation rate and their impact o economic growth: a panel data analysis’. *Romanian Journal of Economic Forecasting* **17**(1):160–176.
- C. J. Kim (1994). ‘Dynamic linear models with Markov switching’. *Journal of Econometrics* **60**(1-2):1–22.
- F. Kleibergen & R. Paap (2002). ‘Priors, posteriors and bayes factors for a Bayesian analysis of cointegration’. *Journal of Econometrics* **111**(2):223–249.
- G. Kotsalis, et al. (2006). ‘Model reduction of discrete-time Markov jump linear systems’. In *Proc. American Control Conference*.

## REFERENCIAS

---

- A. Krogh, et al. (1994). ‘Hidden Markov Models in computational biology: Applications to protein modelling’. *Journal of Molecular Biology* **235**(4):1501–1531.
- A. Krogh, et al. (2001). ‘Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes’. *Journal of Molecular Biology* **305**(3):567–580.
- L. Kuo & B. Mallick (1998). ‘Variable selection for regression models’. *The Indian Journal of Statistics. Special Issue on Bayesian Analysis* **60**(1):65–81.
- J. Li, et al. (2010). ‘Bounded influence estimator for GARCH models: evidence from foreign exchange rates’. *Applied Economics* **42**(11):1437–1445.
- G. Mallik & A. Chowdhury (2001). ‘Inflation and economic growth: evidence from South Asian countries’. *Asian Pacific Development* **8**:123–135.
- R. S. Mamon & R. J. Elliott (eds.) (2014). *Hidden Markov Models in Finance Further Developments and Applications*, vol. II. Springer.
- C. Marsilli (2014). ‘Variable selection in predictive MIDAS models’. Document de travail, Banque de France.
- A. D. Martin, et al. (2017). *Markov Chain Monte Carlo (MCMC) Package*. R package version 1.4-0.
- K. McAlinn & M. West (2016). ‘Dynamic Bayesian Predictive Synthesis in Time Series Forecasting’. Tech. rep., Duke University.
- M. McKenzie & H. Mitchell (2002). ‘Generalized asymmetric power ARCH modelling of exchange rate volatility’. *Applied Financial Economics* **12**(8):555–564.
- L. Mondal (2013). ‘Volatility spillover between the RBI’s intervention and exchange rate’. *International Economic and Economic Policy* **11**(4):549–560.
- M. Moryson (1998). *Testing for random walk coefficients in regression and state space models*. Springer.
- V. M. R. Mugeo (2008). ‘Segmented: An R package to fit regression models with broken-line relationships’. *R News* **8**(1):20–25.
- P. Müller, et al. (2004). ‘A method for combining inference across related nonparametric bayesian models’. *Journal of the Royal Statistical Society* **66**(3):735–749.
- A. Musacchio (2012). ‘Mexico’s financial crisis of 1994-1995’. Tech. Rep. 12-101, Harvard Business School.
- NASA (2016a). ‘Global Climate Change. Vital Signs of the Planet. A blanket around the Earth’.

## REFERENCIAS

---

- NASA (2016b). ‘Global Climate Change. Vital Signs of the Planet. Global Temperature’.
- R. M. Neal (2000). ‘Markov Chain Sampling Methods for Dirichlet Process Mixture Models.’. *Journal of Computational and Applied Mathematics* **9**(2):249–265.
- A. Nikolsko Rzhevskyy & R. Prodan (2012). ‘Markov switching and exchange rate predictability’. *International Journal of Forecasting* **28**(2):353–365.
- S. M. Oh, et al. (2008). ‘Learning and inferring motion patterns using parametric segmental switching linear dynamic systems’. *International Journal of Computer Vision* **77**(1):103–124.
- R. Paap & H. K. van Dijk (2003). ‘Bayes estimates of Markov trends in possibly cointegrated series: an application to U.S. consumption and income’. *Journal of Business and Economic Statistics* **21**(4):547–563.
- S. Paul, et al. (1997). ‘Inflation and economic growth: a multi-country empirical analysis’. *Applied Economics* **29**(10):387–401.
- V. Pavlović, et al. (1999). ‘A dynamic Bayesian network approach to figure tracking using learned dynamic models’. In *International Conf. on Computer Vision*, vol. 99, pp. 94–101.
- V. Pavlović, et al. (2001). ‘Learning switching linear models of human motion.’. In *Advances in Neural Information Processing Systems*, vol. 13. Neural Information Processing Systems (NIPS) 2000.
- M. H. Pesaran, et al. (2001). ‘Bounds testing approaches to the analysis of level relationships’. *Journal of Applied Econometrics* **16**(3):289–326.
- G. Petris, et al. (2009). *Dynamic Linear Models with R*. Springer-Verlag.
- P. C. B. Phillips (1986). ‘Understanding Spurious Regression in Econometrics’. *Journal of Econometrics* **33**:311–340.
- B. Puza & S. Roberts (2013). ‘A Bayesian approach to modeling the interaction between air pollution and temperature’. *Annals of Epidemiology* **23**(4):198–203.
- L. R. Rabiner (1989). ‘A tutorial on hidden Markov models and selected applications in speech recognition’. In *Proceedings of the IEEE*, vol. 77.
- W. A. Risso & E. J. S. Carrera (2009). ‘Inflation and Mexican economic growth: long-run relation and threshold effects’. *Journal of Financial Economic Policy* **1**(3):246–263.

## REFERENCIAS

---

- S. Roberts (2004). ‘Interactions between particulate air pollution and temperature in air pollution mortality time series studies’. *Environmental Research* **96**(3):328–337.
- A. Rodríguez (2007). *Some Advances in Bayesian Nonparametric Modeling*. Ph.D. thesis, Duke University.
- P. Romero Lankao, et al. (2013). ‘Exploration of health risks related to air pollution and temperature in three Latin American cities’. *Social Science & Medicine* **83**:110–118.
- I. Ruiz & S. Pozo (2008). ‘Exchange rates and US direct investment into Latin America’. *Journal of International Trade and Economic Development* **17**(3):411–438.
- R. B. Schinazi (2014). *Classical and spatial stochastic processes*. Springer, 2nd edn.
- J. Sethuraman (1994). ‘A constructive definition of Dirichlet priors’. *Statistica Sinica* **4**:639–650.
- A. A. Shah, et al. (2012). ‘Empirical Analysis of Long and Short Run Relationship among Macroeconomic Variables and Karachi Stock Market: An Auto Regressive Distributive Lag (ARDL) Approach’. *Pakistan Journal of Social Sciences* **32**(2):323–338.
- P. L. Siklos & C. W. Granger (1996). *Temporary cointegration with an application to interest rate parity*. Discussion paper. University of California, San Diego, Department of Economic.
- B. W. Silverman (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, New York.
- D. J. Smyth (1995). ‘Inflation and total factor productivity in Germany’. *Weltwirtschaftliches Archiv* **131**(2):403–405.
- K.-A. Sohn & E. P. Xing (2007). ‘Hidden Markov Dirichlet process: Modeling genetic inference in open ancestral space’. *Bayesian Analysis* **2**(3):501–528.
- M. Stephens (2000). ‘Dealing with label switching in mixture models’. *Journal of the Royal Statistical Society* **62**(4):795–809.
- K. Sugita (2008). ‘Bayesian analysis of a Markov switching temporal cointegration model’. *Japan and the World Economy* **20**(2):257–274.
- S. J. Taylor (1986). *Modelling financial time series*. John Wiley.

## REFERENCIAS

---

- Y. W. Teh, et al. (2005). ‘Sharing clusters among related groups: Hierarchical Dirichlet processes’. In L. K. Saul, Y. Weiss, & L. Bottou (eds.), *Advances in Neural Information Processing Systems 17*, vol. 17, pp. 1385–1392. MIT Press.
- Y. W. Teh, et al. (2006). ‘Hierarchical Dirichlet processes’. *Journal of the American Statistical Association* **101**:1566–1581.
- B. Tims & R. Mahieu (2006). ‘A range-based multivariate stochastic volatility model for exchange rates’. *Econometric Reviews* **25**(2-3):409–424.
- M. A. Tinoco Zermeño, et al. (2014). ‘Growth, bank credit, and inflation in Mexico: evidence from an ARDL-bounds testing approach’. *Latin American Economic Review* **23**(8):1–22.
- C. Velasco Cruz, et al. (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.
- A. Viterbi (1967). ‘Error bounds for convolutional codes and an asymptotically optimum decoding algorithm’. *IEEE Transactions on Information Theory* **13**(2):260–269.
- S. G. Walker (2007). ‘Sampling the Dirichlet Mixture Model with Slice’. *Communications in Statistics-Simulation and Computation* **36**(1):45–54.
- G. R. Warnes, et al. (2015). *Various R Programming Tools*. R package version 3.5.0.
- M. West (1992). ‘Hyperparameter estimation in Dirichlet process mixture models’. Discussion Paper 92-A03, Duke University.
- M. West (2013). *Bayesian Dynamic Modelling*, chap. 8. Oxford University Press.
- M. West & J. Harrison (1997). *Bayesian Forecasting and Dynamic Models*. Springer, 2nd edn.
- WHO (2016). ‘Global Urban Ambient Air Pollution Database’.
- X. Xuan & K. P. Murphy (2007). ‘Modeling changing dependency structure in multivariate time series’. In *Proc. International Conference on Machine Learning*, pp. 1055–1062.
- G. U. Yule (1926). ‘Why do we sometimes get nonsense correlations between time series? A study in sampling and the nature of time series’. *Journal of the Royal Statistical Society* **89**(1):1–63.

## REFERENCIAS

---

- P. Zarchan & H. Musoff (2009). *Fundamentals of Kalman Filtering: A Practical Approach*, vol. 190. American Institute of Aeronautics and Astronautics, 3rd edn.
- Y. Zeng & S. Wu (eds.) (2013). *State-space models. Applications in Economics and Finance*. Springer.
- S. Zhang & J. Buongiorno (2010). ‘Effects of exchange rate volatility on export volume and prices of forest products’. *Canadian Journal of Forest Research* **40**(11):2069–2081.
- E. Zivot & J. Wang (2006). *Modeling Financial Time Series with S-Plus*. Springer, 2nd edn.

# Anexos

# Anexos A

## Procesos Dirichlet para modelos de mezclas

### A.1. Ejemplo 1

Considere el modelo (2.12). Sea  $y_i|\theta \sim N(\mu, V)$ , donde  $\theta = (\mu, V)$ . Considere  $\tau = 1/V$ ,  $\mu|\tau \sim N(m_0, [\tau\rho]^{-1})$  y  $\tau \sim \text{Ga}(a/2, 2/b)$ , donde  $a/2$  y  $2/b$  son parámetros de forma y escala, respectivamente, y  $\rho > 0$ . Entonces, la proporción  $q_0$  de la distribución condicional para  $\theta_i$ , dados los datos y  $\theta_{-i}$ , es proporcional a  $\alpha$  veces  $\int F(y_i, \theta) dG_0(\theta)$ , donde



## A.1. Ejemplo 1

---

$$\begin{aligned}
& \int F(y_i, \theta) dG_0(\theta) \\
&= \int \int \frac{\tau^{1/2}}{\sqrt{2\pi}} \exp \left\{ -\frac{\tau}{2} (y_i - \mu)^2 \right\} \frac{(\tau\rho)^{1/2}}{\sqrt{2\pi}} \exp \left\{ -\frac{\tau\rho}{2} (\mu - m_0)^2 \right\} \\
&\quad \frac{(b/2)^{a/2}}{\Gamma(a/2)} \tau^{a/2-1} \exp \left\{ -\frac{b}{2} \tau \right\} d\mu d\tau \\
&= \int \frac{\tau^{a/2-1}}{\sqrt{2\pi}} (\tau\rho)^{1/2} \frac{(b/2)^{a/2}}{\Gamma(a/2)} \exp \left\{ -\frac{\tau\rho}{2(1+\rho)} \left( m_0(m_0 - 2y_i) + y_i^2 + \frac{b(1+\rho)}{\rho} \right) \right\} \\
&\quad \frac{1}{\sqrt{(1+\rho)}} \left\{ \int \sqrt{\frac{\tau(1+\rho)}{2\pi}} \exp \left\{ -\frac{\tau(1+\rho)}{2} \left( \mu - \frac{y_i + \rho m_0}{1+\rho} \right)^2 \right\} d\mu \right\} d\tau \\
&= \int \frac{\tau^{a/2-1/2}}{\sqrt{2\pi}} (\rho)^{1/2} \frac{(b/2)^{a/2}}{\Gamma(a/2)} \frac{1}{\sqrt{(1+\rho)}} \\
&\quad \exp \left\{ -\frac{\tau\rho}{2(1+\rho)} \left( m_0(m_0 - 2y_i) + y_i^2 + \frac{b(1+\rho)}{\rho} \right) \right\} d\tau \\
&= \int \sqrt{\frac{\rho}{2\pi(1+\rho)}} \frac{(b/2)^{a/2}}{\Gamma(a/2)} \left\{ \frac{\rho(y_i - m_0)^2}{2(1+\rho)} + \frac{b}{2} \right\}^{-\frac{a+1}{2}} \tau^{\frac{a+1}{2}-1} \left\{ \frac{\rho(y_i - m_0)^2}{2(1+\rho)} + \frac{b}{2} \right\}^{\frac{a+1}{2}} \\
&\quad \exp \left\{ -\tau \left( \frac{\rho(y_i - m_0)^2}{2(1+\rho)} + \frac{b}{2} \right) \right\} d\tau \\
&= \sqrt{\frac{\rho}{2\pi(1+\rho)}} \frac{\Gamma(\frac{a+1}{2})(b/2)^{a/2}}{\Gamma(a/2)} \left\{ \frac{\rho(y_i - m_0)^2}{2(1+\rho)} + \frac{b}{2} \right\}^{-\frac{a+1}{2}} \\
&= \sqrt{\frac{\rho}{\pi(1+\rho)b}} \frac{\Gamma(\frac{a+1}{2})}{\Gamma(a/2)} \left[ 1 + \frac{\rho(y_i - m_0)^2}{(1+\rho)b} \right]^{-\frac{a+1}{2}}
\end{aligned}$$

## A.1. Ejemplo 1

La distribución a posteriori conjunta  $G(\theta|\mathbf{y})$  es Normal-Gamma:

$$\begin{aligned}
& G(\mu, \tau|\mathbf{y}) \\
&= \frac{\tau^{n/2}}{\sqrt{2\pi}} \exp \left\{ \frac{-\tau}{2} \sum (y_i - \mu)^2 \right\} \frac{(\tau\rho)^{1/2}}{\sqrt{2\pi}} \exp \left\{ -\frac{\tau\rho}{2} (\mu - m_0)^2 \right\} \\
&\quad \frac{(b/2)^{a/2}}{\Gamma(a/2)} \tau^{a/2-1} \exp \left\{ -\frac{b}{2}\tau \right\} \\
&\propto \frac{\tau^{\frac{n+a}{2}-1}}{\sqrt{2\pi}} (\tau\rho)^{1/2} \exp \left\{ -\frac{\tau}{2}(n+\rho) \left[ \mu^2 - 2\mu \frac{(\rho m_0 + n\bar{y})}{(n+\rho)} + \frac{(\rho m_0 + n\bar{y})^2}{(n+\rho)^2} \right] \right\} \\
&\quad \exp \left\{ -\frac{\tau}{2} \left[ b + \left( \sum y_i^2 + \rho m_0^2 \right) - \frac{(\rho m_0 + n\bar{y})^2}{(n+\rho)} \right] \right\} \\
&\propto \frac{\tau^{1/2}(n+\rho)^{1/2}}{\sqrt{2\pi}} \exp \left\{ -\frac{\tau}{2}(n+\rho) \left[ \mu - \frac{(\rho m_0 + n\bar{y})}{(n+\rho)} \right]^2 \right\} (\tau)^{\frac{n+a}{2}-1} \\
&\quad \frac{\rho^{1/2}}{(n+\rho)^{1/2}} \exp \left\{ -\frac{\tau}{2} \left[ b + \sum y_i^2 + \rho m_0^2 - \frac{(\rho m_0)^2 + 2\rho m_0 n\bar{y} + (n\bar{y})^2}{(n+\rho)} \right] \right\} \\
&\propto N\left(\mu^*, [\tau(n+\rho)]^{-1}\right) (\tau)^{\frac{n+a}{2}-1} \frac{\rho^{1/2}}{(n+\rho)^{1/2}} \\
&\quad \exp \left\{ -\frac{\tau}{2} \left[ b + \sum y_i^2 + \frac{\rho n}{(n+\rho)} \left( m_0^2 - 2m_0\bar{y} - \frac{(\sum y_i)^2}{\rho n} \right) \right] \right\} \\
&\propto N\left(\mu^*, [\tau(n+\rho)]^{-1}\right) (\tau)^{\frac{n+a}{2}-1} \frac{\rho^{1/2}}{(n+\rho)^{1/2}} \\
&\quad \exp \left\{ -\frac{\tau}{2} \left[ b + \sum (y_i - \bar{y})^2 + \frac{\rho n}{(n+\rho)} (\bar{y} - m_0)^2 \right] \right\} \\
&\propto N\left(\mu^*, [\tau(n+\rho)]^{-1}\right) \text{Ga}\left(a^*/2, b^*/2\right)
\end{aligned}$$

donde:

$$\begin{aligned}
\mu^* &= \frac{\rho m_0 + n\bar{y}}{(n+\rho)} \\
a^* &= a + n \\
b^* &= b + \sum (y_i - \bar{y})^2 + \frac{\rho n}{(n+\rho)} (\bar{y} - m_0)^2
\end{aligned}$$

Note que esta distribución, evaluada en  $y_i$  con  $n = 1$ , es la misma que  $G_i(\theta_i)$  de la Ec. (2.14), y  $G_i(\phi)$  de la Ec. (2.19). Mientras que si se evalúa en las observaciones asociadas con cada clase  $c$ , y haciendo  $n$  igual al número de componentes  $c_i$  que son iguales a  $c$ , corresponde a la Ec. (2.20).

## Anexos B

# Distribuciones a priori y posteriori de los parámetros dinámicos de un SLDS

Sea  $\mathbf{B}^{(k)}$  una matriz con  $n_k$  columnas, donde cada columna consiste de los vectores de estados  $\beta_t$  tales que  $z_t = k$ , y

$$\mathbf{B}^{(k)} = \mathbf{A}^{(k)}\mathbf{B}_{-1}^{(k)} + \mathbf{E}^{(k)} \quad k = 1, \dots, K, \quad (\text{B.1})$$

donde:

- $\mathbf{B}^{(k)}$ : una matriz de orden  $(p \times n_k)$  con columnas  $\beta_t$  tales que  $z_t = k$ .
- $\mathbf{B}_{-1}^{(k)}$ : una matriz de orden  $(p \times n_k)$  con columnas  $\beta_{t-1}$  tales que  $z_t = k$ .
- $\mathbf{A}^{(k)}$ : una matriz de orden  $(p \times p)$  correspondiente al  $k$ -ésimo modo.
- $\mathbf{E}^{(k)}$ : una matriz de orden  $(p \times n_k)$  de los vectores de ruido asociados a los  $\beta_t$ .
- $p$  es el tamaño del vector  $\beta_t \forall t$ .

## B.1. A priori conjugada de $\{\mathbf{A}^{(k)}, \Sigma^{(k)}\}$

Considere disponible una muestra de la secuencia de estados  $\boldsymbol{\beta}_{1:T}$ . La MNIW es una a priori conjugada a la verosimilitud definida en (B.1) para el conjunto de parámetros  $\{\mathbf{A}^{(k)}, \Sigma^{(k)}\}$ , tal que:

$$\begin{aligned} p(\mathbf{A}^{(k)} | \Sigma^{(k)}) &\sim \text{MN}(M, \Sigma^{(k)}, \mathbf{K}) \\ p(\Sigma^{(k)}) &\sim \text{IW}(n_0, S_0) \end{aligned}$$

Sea  $\mathbf{D}^{(k)} = \{\mathbf{B}^{(k)}, \mathbf{B}_{-1}^{(k)}\}$ . La verosimilitud del modelo (B.1) está dada por:

$$\begin{aligned} p(\mathbf{D}^{(k)} | \mathbf{A}^{(k)}, \Sigma^{(k)}) &\propto \frac{1}{|\Sigma^{(k)}|^{\frac{n_k}{2}}} \exp \left\{ -\frac{1}{2} \sum_{t:z_t=k} (\boldsymbol{\beta}_t - \mathbf{A}^{(k)} \boldsymbol{\beta}_{t-1})' (\Sigma^{(k)})^{-1} (\boldsymbol{\beta}_t - \mathbf{A}^{(k)} \boldsymbol{\beta}_{t-1}) \right\} \\ &= \frac{1}{|\Sigma^{(k)}|^{\frac{n_k}{2}}} \exp \left\{ -\frac{1}{2} \text{tr} \left[ (\Sigma^{(k)})^{-1} (\mathbf{B}^{(k)} - \mathbf{A}^{(k)} \mathbf{B}_{-1}^{(k)}) (\mathbf{B}^{(k)} - \mathbf{A}^{(k)} \mathbf{B}_{-1}^{(k)})' \right] \right\} \\ &= \frac{1}{|\Sigma^{(k)}|^{\frac{n_k}{2}}} \exp \left\{ -\frac{1}{2} \text{tr} \left[ (\mathbf{B}^{(k)} - \mathbf{A}^{(k)} \mathbf{B}_{-1}^{(k)})' (\Sigma^{(k)})^{-1} (\mathbf{B}^{(k)} - \mathbf{A}^{(k)} \mathbf{B}_{-1}^{(k)}) \right] \right\} \end{aligned}$$

que es proporcional a una distribución  $\text{MN}(\mathbf{A}^{(k)} \mathbf{B}_{-1}^{(k)}, \Sigma^{(k)}, \mathbf{I})$ . Entonces, la distribución condicional completa  $p(\mathbf{A}^{(k)} | \Sigma^{(k)}, \mathbf{D}^{(k)})$  es matriz normal, y  $p(\Sigma^{(k)} | \mathbf{D}^{(k)})$  es Wishart Inversa con los siguientes parámetros:

(1)

$$p(\mathbf{A}^{(k)} | \Sigma^{(k)}, \mathbf{D}^{(k)}) \propto p(\mathbf{A}^{(k)}, \mathbf{D}^{(k)} | \Sigma^{(k)}) \propto \text{MN}(S_{BB_{-1}}^{(k)} (S_{B_{-1}B_{-1}}^{(k)})^{-1}, \Sigma^{(k)}, (S_{B_{-1}B_{-1}}^{(k)})^{-1})$$

## B.1. A priori conjugada de $\{\mathbf{A}^{(k)}, \Sigma^{(k)}\}$

---

donde:

$$\begin{aligned}
p(\mathbf{A}^{(k)}, \mathbf{D}^{(k)} | \Sigma^{(k)}) &\propto \exp \left\{ -\frac{1}{2} \text{tr} \left[ \mathbf{K}^{-1} (\mathbf{A}^{(k)} - \mathbf{M})' (\Sigma^{(k)})^{-1} (\mathbf{A}^{(k)} - \mathbf{M}) \right. \right. \\
&\quad \left. \left. + (\mathbf{B}^{(k)} - \mathbf{A}^{(k)} \mathbf{B}_{-1}^{(k)})^T (\Sigma^{(k)})^{-1} (\mathbf{B}^{(k)} - \mathbf{A}^{(k)} \mathbf{B}_{-1}^{(k)}) \right] \right\} \\
&= \exp \left\{ -\frac{1}{2} \text{tr} \left[ (\Sigma^{(k)})^{-1} [(\mathbf{A}^{(k)} - \mathbf{M}) \mathbf{K}^{-1} (\mathbf{A}^{(k)} - \mathbf{M})' \right. \right. \\
&\quad \left. \left. + (\mathbf{B}^{(k)} - \mathbf{A}^{(k)} \mathbf{B}_{-1}^{(k)}) (\mathbf{B}^{(k)} - \mathbf{A}^{(k)} \mathbf{B}_{-1}^{(k)})' \right] \right\} \\
&= \exp \left\{ -\frac{1}{2} \text{tr} \left[ (\Sigma^{(k)})^{-1} [\mathbf{B}^{(k)} \mathbf{B}^{(k)'} - \mathbf{B}^{(k)} \mathbf{B}_{-1}^{(k)'} \mathbf{A}^{(k)'} - \mathbf{A}^{(k)} \mathbf{B}_{-1}^{(k)} \mathbf{B}^{(k)'} \right. \right. \\
&\quad \left. \left. + \mathbf{A}^{(k)} \mathbf{B}_{-1}^{(k)} \mathbf{B}_{-1}^{(k)'} \mathbf{A}^{(k)'} + \mathbf{A}^{(k)} \mathbf{K}^{-1} \mathbf{A}^{(k)'} - \mathbf{A}^{(k)} \mathbf{K}^{-1} \mathbf{M}' \right. \right. \\
&\quad \left. \left. - \mathbf{M} \mathbf{K}^{-1} \mathbf{A}^{(k)'} + \mathbf{M} \mathbf{K}^{-1} \mathbf{M}' \right] \right\} \\
&= \exp \left\{ -\frac{1}{2} \text{tr} \left[ (\Sigma^{(k)})^{-1} [\mathbf{A}^{(k)} S_{B_{-1}B_{-1}}^{(k)} \mathbf{A}^{(k)'} - 2S_{BB_{-1}}^{(k)} \mathbf{A}^{(k)'} + S_{BB}^{(k)}] \right] \right\} \\
&= \exp \left\{ -\frac{1}{2} \text{tr} \left[ (\Sigma^{(k)})^{-1} [(\mathbf{A}^{(k)} - S_{BB_{-1}}^{(k)} (S_{B_{-1}B_{-1}}^{(k)})^{-1}) S_{B_{-1}B_{-1}}^{(k)} \right. \right. \\
&\quad \left. \left. (\mathbf{A}^{(k)} - S_{BB_{-1}}^{(k)} (S_{B_{-1}B_{-1}}^{(k)})^{-1})' + S_{B|B_{-1}}^{(k)}] \right] \right\} \\
&\propto \exp \left\{ -\frac{1}{2} \text{tr} \left[ S_{B_{-1}B_{-1}}^{(k)} (\mathbf{A}^{(k)} - S_{BB_{-1}}^{(k)} (S_{B_{-1}B_{-1}}^{(k)})^{-1}) (\Sigma^{(k)})^{-1} \right. \right. \\
&\quad \left. \left. (\mathbf{A}^{(k)} - S_{BB_{-1}}^{(k)} (S_{B_{-1}B_{-1}}^{(k)})^{-1})' \right] \right\} \\
&\propto \text{MN}(S_{BB_{-1}}^{(k)} (S_{B_{-1}B_{-1}}^{(k)})^{-1}, \Sigma^{(k)}, (S_{B_{-1}B_{-1}}^{(k)})^{-1})
\end{aligned}$$

$$\begin{aligned}
S_{B_{-1}B_{-1}}^{(k)} &= \mathbf{B}_{-1}^{(k)} \mathbf{B}_{-1}^{(k)T} + \mathbf{K}^{-1} \\
S_{BB_{-1}}^{(k)} &= \mathbf{B}^{(k)} \mathbf{B}_{-1}^{(k)T} + \mathbf{M} \mathbf{K}^{-1} \\
S_{BB}^{(k)} &= \mathbf{B}^{(k)} \mathbf{B}^{(k)T} + \mathbf{M} \mathbf{K}^{-1} \mathbf{M}^T \\
S_{B|B_{-1}}^{(k)} &= S_{BB} - S_{BB_{-1}} (S_{B_{-1}B_{-1}})^{-1} S_{BB_{-1}}^T
\end{aligned}$$

(2)

$$p(\Sigma^{(k)} | \mathbf{D}^{(k)}) \propto p(\mathbf{D}^{(k)} | \Sigma^{(k)}) p(\Sigma^{(k)}) \propto \text{IW}(n_k + n_0, S_{B|B_{-1}}^{(k)} + S_0)$$

## B.1. A priori conjugada de $\{\mathbf{A}^{(k)}, \Sigma^{(k)}\}$

donde:

$$\begin{aligned}
 p(\mathbf{D}^{(k)}|\Sigma^{(k)}) &= \int_{\mathbf{A}^{(k)}} p(\mathbf{A}^{(k)}, \mathbf{D}^{(k)}|\Sigma^{(k)})d\mathbf{A}^{(k)} \\
 &\propto |\Sigma|^{-n_k/2} \exp \left\{ -\frac{1}{2}\text{tr}[(\Sigma^{(k)})^{-1}S_{B|B-1}^{(k)}] \right\} \\
 p(\mathbf{D}^{(k)}|\Sigma^{(k)})p(\Sigma^{(k)}) &\propto |\Sigma|^{-n_k/2} \exp \left\{ -\frac{1}{2}\text{tr}[(\Sigma^{(k)})^{-1}S_{B|B-1}^{(k)}] \right\} \\
 &\quad |\Sigma|^{-(n_0+p+1)/2} \exp \left\{ -\frac{1}{2}\text{tr}[S_0(\Sigma^{(k)})^{-1}] \right\} \\
 &= |\Sigma|^{-(n_0+n_k+p+1)/2} \exp \left\{ -\frac{1}{2}\text{tr}[(S_0 + S_{B|B-1}^{(k)})(\Sigma^{(k)})^{-1}] \right\} \\
 &\propto IW(n_k + n_0, S_{B|B-1}^{(k)} + S_0)
 \end{aligned}$$

Entonces, dada la secuencia de estados  $\beta_{1:T}$ , de modos  $z_{1:T}$ , y un conjunto de valores iniciales  $\{\mathbf{A}^{(k)}, \Sigma^{(k)}\}$ , la actualización de los parámetros dinámicos  $\{\mathbf{A}^{(k)}, \Sigma^{(k)}\}$  usando la distribución a priori conjugada MNIW se resume en el siguiente Algoritmo.

### Actualización de parámetros dinámicos con a priori MNIW

Para cada  $k \in \{1, \dots, K\}$ :

1. Construir  $\mathbf{B}^{(k)}$  y  $\mathbf{B}_{-1}^{(k)}$  como en (B.1).
2. Calcular:

$$\begin{aligned}
 S_{B-1B-1}^{(k)} &= \mathbf{B}_{-1}^{(k)}\mathbf{B}_{-1}^{(k)T} + K^{-1} \\
 S_{BB-1}^{(k)} &= \mathbf{B}^{(k)}\mathbf{B}_{-1}^{(k)T} + MK^{-1} \\
 S_{BB}^{(k)} &= \mathbf{B}^{(k)}\mathbf{B}^{(k)T} + MK^{-1}M^T \\
 S_{B|B-1}^{(k)} &= S_{BB} - S_{BB-1}(S_{B-1B-1})^{-1}S_{BB-1}^T
 \end{aligned}$$

3. Muestrear los parámetros dinámicos de las siguientes distribuciones:

$$\begin{aligned}
 \Sigma^{(k)} &\sim IW(n_k + n_0, S_{B|B-1}^{(k)} + S_0) \\
 \mathbf{A}^{(k)}|\Sigma^{(k)} &\sim MN(S_{BB-1}^{(k)}(S_{B-1B-1}^{(k)})^{-1}, \Sigma^{(k)}, (S_{B-1B-1}^{(k)})^{-1})
 \end{aligned}$$

## B.2. A priori *Automatic Relevance Determination* (ARD)

La a priori conjugada MNIW no permite identificar componentes del vector de estados irrelevantes para el modelo. La a priori ARD reduce el número de componentes haciendo cero aquellos cuya presencia no está soportada por los datos. A priori, se asumen distribuciones Gaussianas independientes sobre las columnas de la matriz  $\mathbf{A}^{(k)}$ , tal que (Fox et al., 2011a):

$$p(\mathbf{A}^{(k)}|\boldsymbol{\alpha}^{(k)}) = \prod_{j=1}^p N(\mathbf{a}_j^{(k)}; 0, 1/\alpha_j^{(k)} I_p) \quad (\text{B.2})$$

donde  $\alpha_j^{(k)} \sim \text{Gam}(a, b)$ . La expresión (B.2) penaliza las columnas de la matriz dinámica por una cantidad proporcional a  $\alpha_j^{(k)}$ . Cuando la evidencia en los datos sea insuficiente para soportar la presencia del  $j$ -ésimo componente del vector de estados,  $\alpha_j^{(k)}$  será grande y  $\mathbf{a}_j^{(k)} \rightarrow 0$ , implicando que ese componente no contribuye al sistema en el  $k$ -ésimo modo.

Para encontrar la distribución condicional completa de  $\mathbf{A}^{(k)}$  considere la siguiente equivalencia de la a priori:

$$p(\text{vec}(\mathbf{A}^{(k)})|\boldsymbol{\alpha}^{(k)}) = N(0, \Sigma_0^{(k)}) \quad (\text{B.3})$$

donde:  $\text{vec}(\mathbf{A}^{(k)})$  denota la vectorización de la matriz  $\mathbf{A}^{(k)}$ ,  $\boldsymbol{\alpha}^{(k)} = \{\alpha_1^{(k)}, \dots, \alpha_p^{(k)}\}$  y  $\Sigma_0^{(k)} = \text{diag}(\alpha_1^{(k)}, \dots, \alpha_1^{(k)}, \dots, \alpha_p^{(k)}, \dots, \alpha_p^{(k)})^{-1}$ , con  $p$  réplicas de cada  $\alpha_i^{(k)}$ , y  $p$  la dimensión del vector de estados. Entonces, la ecuación de estados puede reescribirse como:

$$\begin{aligned} \boldsymbol{\beta}_t &= [\beta_{t-1,1} I_p \quad \beta_{t-1,2} I_p \quad \cdots \quad \beta_{t-1,p} I_p] \text{vec}(\mathbf{A}^{(k)}) + \mathbf{e}_t^{(k)} \quad \forall t | z_t = k \\ &= \tilde{\mathbf{B}}_{t-1} \text{vec}(\mathbf{A}^{(k)}) + \mathbf{e}_t^{(k)} \end{aligned} \quad (\text{B.4})$$

## B.2. A priori Automatic Relevance Determination (ARD)

---

Note que:

$$\begin{aligned}
& \log\{p(\mathbf{D}^{(k)}, \mathbf{A}^{(k)} | \Sigma^{(k)}, \boldsymbol{\alpha}^{(k)})\} \\
& \propto -\frac{1}{2} \sum_{t|z_t=k} (\boldsymbol{\beta}_t - \tilde{\mathbf{B}}_{t-1} \text{vec}(\mathbf{A}^{(k)}))' (\Sigma^{(k)})^{-1} (\boldsymbol{\beta}_t - \tilde{\mathbf{B}}_{t-1} \text{vec}(\mathbf{A}^{(k)})) \\
& \quad - \frac{1}{2} (\text{vec}(\mathbf{A}^{(k)}))' (\Sigma_0^{(k)})^{-1} (\text{vec}(\mathbf{A}^{(k)})) \\
& = -\frac{1}{2} \sum_{t|z_t=k} \left[ (\boldsymbol{\beta}'_t (\Sigma^{(k)})^{-1} \boldsymbol{\beta}_t) - \text{vec}(\mathbf{A}^{(k)})' \tilde{\mathbf{B}}'_{t-1} (\Sigma^{(k)})^{-1} \boldsymbol{\beta}_t \right. \\
& \quad \left. - \boldsymbol{\beta}'_t (\Sigma^{(k)})^{-1} \tilde{\mathbf{B}}_{t-1} \text{vec}(\mathbf{A}^{(k)}) + \text{vec}(\mathbf{A}^{(k)})' \tilde{\mathbf{B}}'_{t-1} (\Sigma^{(k)})^{-1} \tilde{\mathbf{B}}_{t-1} \text{vec}(\mathbf{A}^{(k)}) \right] \\
& \quad - \frac{1}{2} \text{vec}(\mathbf{A}^{(k)})' (\Sigma_0^{(k)})^{-1} \text{vec}(\mathbf{A}^{(k)}) \\
& = -\frac{1}{2} \text{vec}(\mathbf{A}^{(k)})' \left[ (\Sigma_0^{(k)})^{-1} + \sum_{t|z_t=k} \tilde{\mathbf{B}}'_{t-1} (\Sigma^{(k)})^{-1} \tilde{\mathbf{B}}_{t-1} \right] \text{vec}(\mathbf{A}^{(k)}) \\
& \quad + \text{vec}(\mathbf{A}^{(k)})' \left[ \sum_{t|z_t=k} \tilde{\mathbf{B}}'_{t-1} (\Sigma^{(k)})^{-1} \boldsymbol{\beta}_t \right] - \frac{1}{2} \sum_{t|z_t=k} \boldsymbol{\beta}'_t (\Sigma^{(k)})^{-1} \boldsymbol{\beta}_t
\end{aligned}$$

por lo que la distribución deseada está dada por:

$$\begin{aligned}
p(\text{vec}(\mathbf{A}^{(k)}) | \mathbf{D}^{(k)}, \Sigma^{(k)}, \boldsymbol{\alpha}^{(k)}) & \propto p(\mathbf{D}^{(k)}, \mathbf{A}^{(k)} | \Sigma^{(k)}, \boldsymbol{\alpha}^{(k)}) \\
& \propto \exp \left\{ -\frac{1}{2} \left[ \text{vec}(\mathbf{A}^{(k)})' \left[ (\Sigma_0^{(k)})^{-1} + \sum_{t|z_t=k} \tilde{\mathbf{B}}'_{t-1} (\Sigma^{(k)})^{-1} \tilde{\mathbf{B}}_{t-1} \right] \right. \right. \\
& \quad \left. \left. \text{vec}(\mathbf{A}^{(k)}) - 2 \text{vec}(\mathbf{A}^{(k)})' \left[ \sum_{t|z_t=k} \tilde{\mathbf{B}}'_{t-1} (\Sigma^{(k)})^{-1} \boldsymbol{\beta}_t \right] \right] \right\} \\
& \propto N(\boldsymbol{\mu}, \Lambda),
\end{aligned}$$

donde:

$$\begin{aligned}
\boldsymbol{\mu} & = \Lambda \sum_{t|z_t=k} \tilde{\mathbf{B}}'_{t-1} (\Sigma^{(k)})^{-1} \boldsymbol{\beta}_t \\
\Lambda & = \left[ (\Sigma_0^{(k)})^{-1} + \sum_{t|z_t=k} \tilde{\mathbf{B}}'_{t-1} (\Sigma^{(k)})^{-1} \tilde{\mathbf{B}}_{t-1} \right]^{-1}.
\end{aligned}$$

Cada componente del vector de parámetros de precisión  $\boldsymbol{\alpha}^{(k)}$  tiene a priori una distribución gamma independiente,  $\text{Gam}(a, b)$ ; la condicional para el proceso de



### B.3. Distribución a priori y posteriori del ruido de medición

---

actualización es entonces:

$$\begin{aligned}
 p(\alpha_l^{(k)} | \mathbf{A}^{(k)}) &\propto p(\mathbf{A}^{(k)} | \boldsymbol{\alpha}^{(k)}) p(\alpha_l^{(k)}) \\
 &\propto |\Sigma_0^{(k)}|^{-1/2} \exp \left\{ -1/2 \text{vec}(\mathbf{A}^{(k)})' (\Sigma_0^{(k)})^{-1} \text{vec}(\mathbf{A}^{(k)}) \right\} \exp \left\{ -b \alpha_l^{(k)} \right\} (\alpha_l^{(k)})^{a-1} \\
 &\propto (\alpha_l^{(k)})^{(a+p/2-1)} \exp \left\{ -\alpha_l^{(k)} \left[ b + \sum_i (a_{il}^{(k)})^2 / 2 \right] \right\} \\
 &\propto \text{Gam} \left( a + p/2, b + \sum_i (a_{il}^{(k)})^2 / 2 \right).
 \end{aligned}$$

Para  $\Sigma^{(k)}$  se define a priori una distribución IW( $n_0, S_0$ ); la condicional dados  $\mathbf{A}^{(k)}$  y  $\mathbf{D}^{(k)}$  es también IW:

$$\begin{aligned}
 p(\Sigma^{(k)} | \mathbf{D}^{(k)}, \mathbf{A}^{(k)}) &\propto |\Sigma^{(k)}|^{-n_k/2} \exp \left\{ -\frac{1}{2} \sum_{t|z_t=k} (\boldsymbol{\beta}_t - \mathbf{A}^{(k)} \boldsymbol{\beta}_{t-1})' (\Sigma^{(k)})^{-1} (\boldsymbol{\beta}_t - \mathbf{A}^{(k)} \boldsymbol{\beta}_{t-1}) \right\} \\
 &\quad |\Sigma^{(k)}|^{-\frac{n_0+p+1}{2}} \exp \left\{ -\frac{1}{2} \text{tr}(S_0 (\Sigma^{(k)})^{-1}) \right\} \\
 &\propto |\Sigma^{(k)}|^{-\frac{n_0+n_k+p+1}{2}} \exp \left\{ -\frac{1}{2} \text{tr} \left[ (S_0 + S_{\beta|\beta-1}) (\Sigma^{(k)})^{-1} \right] \right\} \\
 &\propto \text{IW} \left( n_0 + n_k, S_0 + S_{\beta|\beta-1} \right)
 \end{aligned}$$

donde:

$$\begin{aligned}
 n_k &= |\{t | z_t = k, t = 1, \dots, T\}| \\
 S_{\beta|\beta-1} &= \sum_{t|z_t=k} (\boldsymbol{\beta}_t - \mathbf{A}^{(k)} \boldsymbol{\beta}_{t-1}) (\boldsymbol{\beta}_t - \mathbf{A}^{(k)} \boldsymbol{\beta}_{t-1})^T
 \end{aligned}$$

### B.3. Distribución a priori y posteriori del ruido de medición

La distribución Inversa-Gamma (IG) se usa frecuentemente en estadística Bayesiana como distribución a priori conjugada para parámetros de escala, tal como la varianza de una distribución normal. Sin embargo, en muchas aplicaciones es más conveniente trabajar con el inverso del parámetro de escala, por ejemplo, la precisión en una distribución normal o la *tasa* (*rate*) en una distribución Gamma

### B.3. Distribución a priori y posteriori del ruido de medición

---

(Gam). En estos casos, la distribución Gam se usa como una distribución a priori conjugada. Estas dos distribuciones, IG y Gam, están relacionadas de la siguiente manera: si  $X \sim \text{Gam}(\alpha, \beta)$ , donde  $\alpha$  es un parámetro de forma y  $\beta$  un parámetro de escala, tal que  $E(X) = \alpha\beta$ , entonces  $1/X \sim \text{IG}(\alpha, \beta^{-1})$ .

Dado el modelo (4.6), hacer inferencia sobre el parámetro de varianza del ruido de medición  $R^{(k)}$  implica calcular la distribución condicional

$$p(R^{(k)} | y_{1:T}, \boldsymbol{\beta}_{1:T}, z_{1:T}).$$

Asumiendo como distribución a priori  $\text{IG}(a_r, b_r)$ , es decir,  $1/R^{(k)} \sim \text{Gam}(a_r, 1/b_r)$ , es fácil verificar que, para toda  $k$ :

$$1/R^{(k)} \sim \text{Gam}\left(\frac{n_k}{2} + a_r, b_r + \frac{1}{2} \sum_{t:z_t=k} (y_t - X'_t \boldsymbol{\beta}_t)^2\right),$$

donde:  $n_k = |\{t : z_t = k, t = 1 \dots, T\}|$ .

# Glosario

**DP** Acrónimo de *Dirichlet process*. ix, xiv, 2, 14, 15, 16, 17, 19, 21, 22, 24, 25, 37, 38, 42, 44, 45, 46, 47, 84, 85, 86, 87, 88, 90, 96, 166

**DPMM** Acrónimo de *Dirichlet process mixture model*, traducido como proceso Dirichlet para modelos de mezclas. 5, 9, 22, 24, 25, 26, 36, 44, 166

**HDP** Acrónimo de *hierarchical Dirichlet process*, traducido como proceso Dirichlet jerárquico. x, xi, xv, 5, 6, 44, 45, 46, 47, 82, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 96, 97, 99, 106, 107, 109, 111, 115, 121, 122, 127, 136, 139, 142, 146, 149, 152, 153

**HMM** Acrónimo de *hidden Markov Models*, traducido como modelo de Markov oculto. x, xv, 1, 2, 5, 6, 50, 53, 54, 55, 56, 57, 58, 62, 64, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 93, 95, 96, 100, 107, 119, 153

**LDS** Acrónimo de *linear dynamical system*, traducido como sistema dinámico lineal. 1, 2, 5, 6, 50, 53, 55, 62, 63, 64, 69, 71, 76, 83, 84, 89, 100, 107, 136, 147, 148, 149, 166, 167, 168, 169

**SLDS** Acrónimo de *switching linear dynamical systems*, traducido como sistema dinámico lineal de cambio de régimen. ix, x, xi, xv, 1, 2, 3, 6, 7, 80, 81, 82, 83, 84, 89, 90, 91, 92, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 99, 101, 102, 103, 104, 105, 106, 107, 106, 107, 108, 109, 110, 109, 111, 112, 111, 113, 114, 115, 116, 115, 117, 118, 119, 120, 119, 121, 122, 123, 122, 124, 125, 126, 127, 128, 127, 136, 139, 142, 146, 147, 148, 149, 164, 165, 167, 168, 169, 170