00365

UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

POSGRADO EN CIENCIAS MATEMATICAS
FACULTAD DE CIENCIAS

UN CRITERIO PREDICTIVO DE SELECCION DE MODELOS PARA SERIES DE TIEMPO

T E S I S

QUE PARA OBTENER EL GRADO ACADEMICO DE

MAESTRO EN CIENCIAS MATEMATICAS

P R E S E N T A

JUAN CARLOS MARTINEZ OVANDO

DIRECTOR DE TESIS: DR. EDUARDO A. GUTIERREZ PEÑA

MEXICO, D.F. JUNIO 2004

Resumen

Un Criterio Predictivo de Selección de Modelos para Series de Tiempo

Juan Carlos Martínez Ovando

Departamento de Probabilidad y Estadística

IIMAS - UNAM

Junio, 2004

Supervisor: Dr. Eduardo A. Gutiérrez Peña

Actualmente existe un fructífero debate en torno al problema de selección Bayesiana de modelos, particularmente de modelos paramétricos. En este trabajo revisamos algunas de las soluciones que surgen de plantear este problema como un problema Bayesiano de decisión en un ambiente de incertidumbre, concentrándonos en el caso en que el objetivo final del análisis es de predicción. La solución del problema depende, entre otras cosas, de la perspectiva que se asuma respecto a la clase de modelos contendientes $\mathcal{M}$. Nosotros consideramos la perspectiva $\mathcal{M}$-abierta como la alternativa más honesta, en ella se considera que ninguno de los modelos dentro de la clase $\mathcal{M}$ puede ser visto como como el modelo generador del proceso en estudio. Gutiérrez-Peña y Walker (2001) propusieron una solución para este problema en el caso de variables aleatorias independientes (o intercambiables). En este trabajo nosotros aplicamos este criterio para analizar series de tiempo, con la variante particular de que usamos un modelo semiparamétrico para aproximar la "verdadera" distribución predictiva de la serie. Asumiendo esta perspectiva y utilizando la función de utilidad logarítmica se tiene que la solución óptima del problema consiste en elegir el modelo paramétrico que tenga la distribución predictiva más semejante a la del modelo semiparamétrico, en el sentido de Kullback-Leibler. Terminamos este trabajo presentando una aplicación práctica y una discusión.

Temas: Selección de modelos, series de tiempo, teoría de decisión, inferencia Bayesiana paramétrica y semiparamétrica.

Dedico este trabajo a mis amados padres

Gustavo y Lupita,

y a mis queridas hermanas

Angela y Mariana,

con mucho cariño.

Agradecimientos

La realización de este trabajo se vió influenciada por la presencia y participación directa o indirecta de diferentes personas. A todas ellas de antemano les expreso mi más cordial agradecimiento.

Agradezco a mi supervisor el Dr. Eduardo Gutiérrez Peña por el interés, el tiempo y la dedicación que me brindó durante la elaboración de esta tesis. Trabajar bajo su supervisión ha sido un privilegio y una constante fuente de aprendizaje y estímulo de superación. Su trato personal y comprensión fueron fundamentales para llevar a término este trabajo. Muchas gracias Eduardo.

También agradezco a mis sinodales, los Doctores Manuel Mendoza, Raúl Rueda, Alberto Contreras y Luis Enrique Nieto Barajas. Sus comentarios juiciosos y sus correcciones contribuyeron de manera sustantiva para este trabajo tuviera una mejor presentación, y despertaron en mí nuevos temas de reflexión. También agradezco a mis profesores, quienes a través de su dedicación han estimulado en mi el deseo de aprender y aplicar los conceptos de estadística.

En el periodo de estancia en la maestría tuve la oportunidad de convivir con espléndidas personas como compañero de clase y amigo, a todos ellos les estoy agradecido, especialmente a Eunice Campirán, Jessica Hernández, Hugo Villaseñor y Karim Anaya. Después de este periodo también he tenido la fortuna de laborar con personas muy especiales, a mis amigos y compañeros del Programa Oportunidades, en especial a Nora Jaimes y Laura Dávila, y recientemente del Banco de México, en especial a Graciela Ruiz y al Dr. Gabriel Vera, les estoy profundamente agradecido.

Agradezco el apoyo financiero que me proporcionó la DGEP al inicio de mis estudios de maestría. Quiero expresar mi agradecimiento al CONACyT por el apoyo financiero que me brindó para la realización de esta tesis a través del Proyecto de Investigación #32256-E.

Este trabajo no hubiese sido concebido sin el apoyo y estímulo incondicional que mis papás y mis hermanas me brindan constantemente, a ellos les estaré eternamente agradecido.

JUAN CARLOS MARTÍNEZ OVANDO

Inst. Inv. Mat. Aplic. y Sist. - UNAM

13 de Junio de 2004.

Contenido

Lista de figuras

Lista de cuadros

Capítulo 1

Introducción

Uno de los principales problemas en el análisis estadístico de datos reside en nuestra incertidumbre respecto al modelo que planeamos utilizar para reportar nuestros resultados. Generalmente éste es elegido dentro de una clase de modelos paramétricos. Bajo ciertas circunstancias es necesario elegir un modelo específico y conducir nuestro estudio y posibles decisiones futuras con base en él. La elección de un sólo modelo no es una tarea simple, pues en este proceso intervienen diferentes factores que deben ser considerados, tales como restricciones en los recursos para la obtención de los datos, la naturaleza del problema, su marco teórico, etc. Lo cierto es que por más exhaustivo que sea el proceso de selección subyace el problema de que en general ningún modelo planteado representa la "verdadera" naturaleza del fenómeno de interés.

Históricamente se han planteado diferentes soluciones o criterios para seleccionar un sólo modelo. Algunos de estos procedimientos se restringen a la comparación y discriminación de modelos pertenecientes a la misma familia de modelos paramétricos. En este trabajo nuestro interés se concentra en la selección de modelos estadísticos usando el enfoque Bayesiano de inferencia. En particular, el criterio que es de nuestro interés consiste en la selección de modelos considerando el problema de predicción como objetivo final del análisis, pensado que en la mayoría de los casos el objetivo final del análisis estadístico es el de predecir valores futuros de una variable aleatoria, i.e. encontrar un

modelo que reproduzca la naturaleza de los datos observados. Inclusive, de no ser éste el caso, es deseable que el modelo seleccionado tenga una buena capacidad predictiva, pues representaría de mejor forma la naturaleza de los datos observados y consecuentemente nuestro análisis tendría una mejor sustentabilidad. Bajo la perspectiva Bayesiana, las predicciones futuras de una variable aleatoria se obtienen por medio de una medida de probabilidad definida sobre el espacio de los posibles valores que puede tomar la variable de interés. La comparación y selección de modelos se realizará a través de la comparación de sus correspondientes distribuciones predictivas. Esta comparación tiene una justificación adicional, pues esencialmente éstas características de los modelos son comparables entre sí.

La selección del mejor modelo se realizará por medio de un procedimiento de toma de decisión bajo un ambiente de incertidumbre respecto al "verdadero" modelo del fenómeno de interés. Recientemente Gutiérrez-Peña y Walker (2001) propusieron un criterio para la selección de modelos considerando que ninguno de los modelos postulados es considerado como el modelo verdadero. Este criterio consiste en elegir el modelo que mejor se aproxime a un modelo no paramétrico o semiparamétrico, que en esencia resulta más flexible que los modelos paramétricos tradicionales y donde los supuestos estructurales del modelo son más relajados que en los modelos paramétricos. La comparación de los modelos paramétricos postulados con el modelo semiparamétrico se realiza mediante una función de utilidad adecuada, y la solución consiste en elegir el modelo que maximice la correspondiente utilidad esperada. En términos generales resulta en elegir el modelo paramétrico cuya densidad predictiva se aproxime mejor a la densidad predictiva del modelo no paramétrico flexible. Posteriormente Walker *et al.* (2001) extendieron este criterio para seleccionar no sólo un modelo paramétrico, sino una mezcla de éstos, bajo un criterio semejante al original.

En este trabajo aplicaremos el criterio predictivo de selección y mezcla de modelos propuesto por Gutiérrez-Peña y Walker (2001) para el análisis de series de tiempo. La sustentabilidad práctica de este criterio se basa en una adecuada especificación de un modelo no paramétrico o semiparamétrico flexible. En el análisis de series de tiempo ésta no es

una tarea simple, pues existen diferentes factores y características en los procesos para los cuales no resulta posible especificar una clase de modelos estrictamente no paramétrica. Una aproximación consiste en la determinación de modelos semiparamétricos que flexibilicen ciertos elementos del proceso de interés. La alternativa que proponemos consiste en la mezcla de dos modelos semiparamétricos, cada uno enfocado en modelar diferentes características del proceso de manera semiparamétrica. La mezcla se realiza a través del enfoque Bayesiano tradicional, de manera que los pesos en la mezcla estarán determinados por la naturaleza de los datos observados.

1.1 Estructura de la Tesis

Este trabajo está compuesto por seis capítulos y cuatro apéndices. El primer capítulo se introduce al problema de selección de modelos. En el Capítulo 2 daremos una breve introducción al enfoque Bayesiano de inferencia y predicción, haciendo énfasis en la clase de modelos paramétricos. Daremos también una introducción a algunos elementos operativos para su implementación, así como a algunas técnicas numéricas de aproximación que resultan de utilidad en diversos problemas prácticos. Al final del capítulo realizaremos una breve introducción al análisis de series de tiempo, así como a algunos métodos de inferencia Bayesiana sobre ciertos modelos representativos.

En el Capítulo 3 realizamos una breve revisión histórica de algunos criterios Bayesianos de selección y combinación de modelos. Revisaremos las propuestas de selección de modelos que han tenido una mayor aceptación bajo el enfoque Bayesiano, revisaremos las diferentes posturas que un analista puede asumir respecto a la clase de modelos postulados, brindaremos una breve descripción de los elementos que integran el problema de selección de modelos en un ambiente de incertidumbre, y argumentaremos que el criterio predictivo es una alternativa honesta, aunque computacionalmente demandante, para seleccionar modelos. Al final presentaremos dos ejemplos del método de selección y su comparación con otros criterios, dentro de la clase de los modelos autorregresivos lineales.

En el Capítulo 4 realizaremos una breve descripción del análisis semiparamétrico de series de tiempo con el objetivo de predecir valores futuros de la serie. Consideramos

fundamentalmente descomponer un modelo de series de tiempo en dos componentes, uno caracterizando la parte sistemática temporal del proceso y otro caracterizando la parte aleatoria relacionada con el proceso de manera semiparamétrica. Describiremos dos clases de modelos que modelan semiparamétricamente uno de estos dos componente de ésta descomposición. El primero modela la parte sistemática temporal de la serie mediante la adición de bases de onduletas radiales (o bases radiales en general), y el segundo modela semiparamétricamente el componente aleatorio a través de un proceso de árbol de Pólya. Al final del capítulo describimos la forma de mezclar ambos modelos con el propósito de crear un modelo predictivo más flexible, que sirva como modelo juez para la comparación de los modelos paramétricos postulados.

Una aplicación del criterio de selección se realiza en el Capítulo 5, donde se estudian y comparan diferentes modelos para la serie del Índice Metropolitano de Calidad del Aire (IMECA) de la Ciudad de México correspondiente al monitor de la zona centro. Sugerimos modelar esta serie mediante cinco modelos completamente paramétricos, cuyos elementos en cada caso tienen interpretaciones distintas. Consideramos modelos dinámicos con niveles sujetos a cambios aleatorios y por otro lado consideramos un modelo más flexible que incorpora un proceso de ruido aleatorio estable en lugar del tradicional ruido aleatorio Gaussiano, definiendo así un modelo más robusto en los errores. La comparación de éstos modelos se realiza siguiendo el procedimiento predictivo descrito al final del capitulo 3 y considerando en este caso a la mezcla de densidades predictivas *flexibles* como la densidad juez del proceso de comparación.

En el Capítulo 6 culminaremos el trabajo con una breve discusión sobre las ventajas y limitaciones que tiene el criterio predictivo de selección que hemos instrumentado, así como una revisión sobre los resultados empíricos obtenidos.

Al final de la tesis se incluyen algunos apéndices con información de apoyo para diferentes partes de este trabajo. En el Apéndice A se presentan las propiedades más importantes de la medida de divergencia de Kullback-Leibler de funciones de densidad. En el Apéndice B resumimos algunos resultados básicos de inferencia Bayesiana para el modelo de regresión lineal de rango completo, además de enunciar la definición y algunos

resultados generales relacionados con los modelos utilizados en el capítulo 5 con relación a algunas propiedades de las distribuciones α-estables y aproximaciones no paramétricas de densidades por medio de *kernels*. En el Apéndice C presentamos algunas medidas de diagnóstico predictivo de modelos, que empleamos en las aplicaciones. Finalmente, en el Apéndice D realizamos una breve descripción general del software desarrollado para las aplicaciones que realizamos durante este trabajo. Todos los códigos fueron desarrollados en MATLAB (MathWorks, 2000), y en algunos casos se emplearon funciones auxiliares que son compartidas gratuitamente a través su repositorio en la red[1]. Estos códigos se encuentra disponibles libremente para los lectores interesados y son distribuidas en un disco adicional a este documento.

[1] www.mathtools.net/MATLAB/Add-on_functions/

Capítulo 2

Preliminares

2.1 Paradigma Bayesiano

La estadística es el estudio de fenómenos bajo un estado de conocimiento o información incompleto. Los fundamentos teóricos de lo que en la actualidad se conoce como estadística Bayesiana tienen su origen con la publicación de un artículo del Reverendo Thomas Bayes en 1773, dos años después de su muerte. En ese trabajo, Thomas Bayes resolvió un problema de información inversa planteado por Bernoulli, que consiste en obtener información sobre réplicas independientes de variables aleatorias Bernoulli. Una década después, Laplace retomó las ideas de Bayes y desarrolló con mayor claridad lo que en la actualidad se conoce como el paradigma Bayesiano de inferencia.

Denotemos por $\mathcal{H}_1, \mathcal{H}_2, ..., \mathcal{H}_p$ a una colección de proposiciones o hipótesis excluyentes y exhaustivas, y supongamos que deseamos realizar inferencias sobre éstas con base en un nivel de información denotado por $\mathcal{I}$, el cual resume nuestra percepción e información inicial respecto a estas hipótesis. Nuestro estado de información lo expresamos a través de una medida de probabilidad definida sobre el espacio de las hipótesis o proposiciones en cuestión, condicional en nuestro estado de información, que denotamos por $P(\mathcal{H}_i|\mathcal{I})$, y es tal que $P(\mathcal{H}_1|\mathcal{I})+\cdots+P(\mathcal{H}_p|\mathcal{I}) = 1$. Nuestro aprendizaje respecto a las hipótesis consiste en la actualización del conocimiento mediante la incorporación de nueva información

relevante, que denotamos por $\mathcal{D}$. Por simetría tenemos la siguiente relación

$$P(\mathcal{D}|\mathcal{I})\,P(\mathcal{H}_i|\mathcal{D},\mathcal{I}) = P(\mathcal{H}_i|\mathcal{I})\,P(\mathcal{D}|\mathcal{H}_i,\mathcal{I}), \tag{2.1}$$

para $i = 1,...,p$. Si $P(\mathcal{D}|\mathcal{I}) > 0$, i.e. la información relevante proporcionada por el entorno real es plausible, entonces nuestro estado de información actualizado es de la forma

$$P(\mathcal{H}_i|\mathcal{D},\mathcal{I}) = P(\mathcal{H}_i|\mathcal{I})\,\frac{P(\mathcal{D}|\mathcal{H}_i,\mathcal{I})}{P(\mathcal{D}|\mathcal{I})}. \tag{2.2}$$

La relación (2.2) es la representación matemática del proceso de aprendizaje y es conocida como el Teorema de Bayes, aún cuando Bayes no haya sido quien lo enunció formalmente. Esta relación muestra cómo la probabilidad inicial o *a priori* respecto a las hipótesis, $P(\mathcal{H}_i|\mathcal{I})$, es actualizada a la probabilidad final o *a posteriori*, $P(\mathcal{H}_i|\mathcal{D},\mathcal{I})$, como resultado de la incorporación de nueva información $\mathcal{D}$. El Teorema de Bayes puede ser aplicado repetidamente conforme nueva información $\mathcal{D}_1, \mathcal{D}_2, ...$ es obtenida, en cuyo caso la distribución final se convierte en la nueva información inicial para el caso siguiente, de forma que en cualquier instante la *plausibilidad* de la hipótesis $\mathcal{H}_i$ dependerá de la evidencia total disponible. De esta forma, captura la naturaleza secuencial del proceso de aprendizaje general que usualmente efectuamos en nuestra vida cotidiana.

Durante los años subsecuentes del siglo XVIII y del siguiente, el paradigma Bayesiano se encontró en un estado inerte debido a que esta teoría carecía de sustentabilidad teórica respecto al enfoque frecuentista de inferencia. No fue sino hasta el segundo tercio del siglo pasado en que Harold Jeffreys y Bruno de Finetti desarrollaron y formalizaron la teoría que actualmente se encuentra vigente. Ambos fueron defensores del paradigma Bayesiano, aunque tenían visiones distintas respecto a la conceptualización e interpretación de la probabilidad. Por un lado, Jeffreys defendió una postura *objetiva* sobre el tema, y por otro lado de Finetti propuso y formalizó una visión enteramente *subjetiva* donde se entiende que la probabilidad mide el grado de creencia respecto al fenómeno de interés del individuo quien la expresa. En este sentido, la información inicial en la relación del (2.2) es necesariamente subjetiva. La interpretación objetiva e impositiva de Jeffreys, desarrollada de manera más flexible por Richard Cox, se basa en un principio de consistencia,

el cual enuncia que dos individuos con el mismo nivel de información deban necesariamente reportar la misma apreciación inicial respecto a su incertidumbre de manera que las conclusiones que éstos generan necesariamente deben de ser completamente compatibles. Con el enfoque de de Finetti esta regla no necesariamente debe de cumplirse.

El enfoque Bayesiano ha evolucionado de manera sorprendente durante los años subsecuentes. Su uso nos provee de una herramienta útil de inferencia y sobre todo de predicción, que en general puede considerarse como el problema central del análisis estadístico. Una revisión detallada respecto a la evolución del paradigma Bayesiano, y en general del proceso de inferencia estadística, la podemos encontrar en Hald (1998). En las siguientes subsecciones describiremos los principios fundamentales de inferencia estadística Bayesiana y predicción.

2.1.1 Inferencia Estadística y el Proceso de Aprendizaje

Supongamos que una variable aleatoria de interés, denotada por Y, tiene una distribución de probabilidad en la familia $\mathcal{P} = \{p(y|\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, donde $\boldsymbol{\theta}$ es un parámetro que indiza la función de probabilidad de la variable aleatoria Y, y $\boldsymbol{\Theta}$ es un espacio parametral. Desde el enfoque Bayesiano el desconocimiento sobre el valor del parámetro de interés $\boldsymbol{\theta}$ es manifestado mediante la asignación de una medida de probabilidad, digamos $\pi(\boldsymbol{\theta})$, que representa nuestro nivel de información sobre el verdadero valor de éste. Denotemos por $\boldsymbol{y}$ a un conjunto de realizaciones observables de la variable Y, y denotemos a la distribución de probabilidad conjunta de $\boldsymbol{y}$ y $\boldsymbol{\theta}$ por $p(\boldsymbol{y}, \boldsymbol{\theta})$. Entonces, por las leyes básicas de probabilidad, se cumplen las siguientes relaciones

$$p(\boldsymbol{y}, \boldsymbol{\theta}) = p(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|\boldsymbol{y})p(\boldsymbol{y}) \tag{2.3}$$

donde $p(\boldsymbol{y}|\boldsymbol{\theta})$ es la función de probabilidad de la v.a. $\boldsymbol{y}$ condicional en $\boldsymbol{\theta}$; y $p(\boldsymbol{y})$ y $\pi(\boldsymbol{\theta})$ son las funciones de densidad marginales de $\boldsymbol{y}$ y $\boldsymbol{\theta}$ respectivamente.

De las ecuaciones anteriores es posible deducir que

$$\begin{aligned} \pi(\boldsymbol{\theta}|\boldsymbol{y}) &= \frac{p(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{p(\boldsymbol{y})} \\ &\propto p(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \\ &\propto \textit{verosimilitud} \times \textit{inicial} \end{aligned} \tag{2.4}$$

donde $\propto$ denota proporcionalidad en términos de $\boldsymbol{\theta}$; $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ es conocida como la distribución *final* (o *a posteriori*) del parámetro $\boldsymbol{\theta}$, condicional a la información muestral $\boldsymbol{y}$; $\pi(\boldsymbol{\theta})$ es la distribución *inicial* (o *a priori*) asignada al parámetro $\boldsymbol{\theta}$; y $p(\boldsymbol{y}|\boldsymbol{\theta})$ es la función de verosimilitud, vista como función de $\boldsymbol{\theta}$.

La distribución inicial del parámetro $\pi(\boldsymbol{\theta})$ cuantifica nuestro estado de información respecto al valor desconocido del parámetro $\boldsymbol{\theta}$. Este conocimiento lo actualizamos, mediante la aplicación del Teorema de Bayes, con la incorporación de información adicional relevante, por ejemplo una muestra aleatoria observada $\boldsymbol{y}$ de la variable de interés, que proporcione evidencia sobre el verdadero parámetro $\boldsymbol{\theta}$. De esta forma, nuestro conocimiento actualizado sobre el parámetro $\boldsymbol{\theta}$ es resumido en la distribución final, o *a posteriori*, $\pi(\boldsymbol{\theta}|\boldsymbol{y})$.

Consideremos ahora que el parámetro está particionado como $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, y que sólo un subconjunto de éste, digamos $\boldsymbol{\theta}_1$, es de interés inferencial; en este caso $\boldsymbol{\theta}_2$ es conocido como parámetro de ruido. Dada una muestra $\boldsymbol{y}$, es de interés encontrar la distribución marginal final de $\boldsymbol{\theta}_1$, sin prestar atención al valor de $\boldsymbol{\theta}_2$. De esta forma la inferencia que se realice sobre $\boldsymbol{\theta}_1$ deberá basarse en la distribución final de $\boldsymbol{\theta}_1$ condicional en $\boldsymbol{y}$, la cual obtenemos con el siguiente proceso de marginalización

$$\begin{aligned} \pi\left(\boldsymbol{\theta}_1|\boldsymbol{y}\right) &= \int \pi\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\boldsymbol{y}\right) d\boldsymbol{\theta}_2 \\ &= \int \pi\left(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \boldsymbol{y}\right) \pi\left(\boldsymbol{\theta}_2|\boldsymbol{y}\right) d\boldsymbol{\theta}_2 \\ &= \mathbb{E}_{\Theta_2|\boldsymbol{y}}\left\{\pi\left(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \boldsymbol{y}\right)\right\} \end{aligned}$$

donde Θ_2 es el espacio parametral de $\boldsymbol{\theta}_2$; $\pi\left(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \boldsymbol{y}\right)$ es la distribución final condicional de $\boldsymbol{\theta}_1$ dado $\boldsymbol{\theta}_2$ y $\boldsymbol{y}$; y $\pi\left(\boldsymbol{\theta}_1|\boldsymbol{y}\right)$ es la distribución final marginal de $\boldsymbol{\theta}_1$ dado $\boldsymbol{y}$.

La aplicación del paradigma Bayesiano nos permite establecer un procedimiento secuencial de actualización de la información sobre el parámetro de interés $\boldsymbol{\theta}$. Supongamos

que $\boldsymbol{y}_1$ es una realización de la v.a. Y. Aplicando el Teorema de Bayes (2.4) la distribución final de $\boldsymbol{\theta}$ dado $\boldsymbol{y}_1$ es $\pi(\boldsymbol{\theta}|\boldsymbol{y}_1)$. Si posteriormente se tiene acceso a otra realización de Y, denotada por $\boldsymbol{y}_2$, entonces la distribución final de $\boldsymbol{\theta}$ dado $\boldsymbol{y}_1$ y $\boldsymbol{y}_2$, que resume nuestro conocimiento sobre $\boldsymbol{\theta}$ actualizado por $\boldsymbol{y}_1$ y $\boldsymbol{y}_2$, puede expresarse como

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}_1, \boldsymbol{y}_2) \propto p(\boldsymbol{y}_2|\boldsymbol{\theta}, \boldsymbol{y}_1)\pi(\boldsymbol{\theta}|\boldsymbol{y}_1). \tag{2.5}$$

De la ecuación (2.5) se puede establecer un procedimiento de aprendizaje secuencial, si se considera a $\pi(\boldsymbol{\theta}|\boldsymbol{y}_1)$ como la nueva distribución inicial para $\boldsymbol{\theta}$, antes de observar $\boldsymbol{y}_2$.

2.1.2 Predicción

Uno de los objetivos centrales del análisis estadístico es el de predecir valores futuros de una variable aleatoria de interés Y condicional en la información histórica observada de la misma variable, $\boldsymbol{y}$ (Box, 1980), y posiblemente bajo las consideraciones de algunos otros elementos o factores adicionales relevantes. Usando el enfoque Bayesiano, los resultados siguiendo este objetivo, se resumen a través de una distribución de probabilidad definida sobre el espacio de las variables futuras, i.e. toda la información relevante sobre la variable futura, denotada por Y_f, estará resumida en $p(y_f|\boldsymbol{y})$, cuyo cálculo se obtiene de manera directa usando en Teorema de Bayes y un proceso simple de marginalización. De la ecuación (2.3) tenemos que

$$p(y_f, \boldsymbol{\theta}|\boldsymbol{y}) = p(y_f|\boldsymbol{\theta}, \boldsymbol{y})\pi(\boldsymbol{\theta}|\boldsymbol{y}),$$

donde $p(y_f|\boldsymbol{\theta}, \boldsymbol{y})$ es la densidad de la variable Y_f condicional en $\boldsymbol{\theta}$ y $\boldsymbol{y}$; y $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ es la distribución final de $\boldsymbol{\theta}$ dado $\boldsymbol{y}$. De esta forma, la distribución predictiva final la podemos calcular como

$$\begin{aligned} p(y_f|\boldsymbol{y}) &= \int p(y_f, \boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta} \\ &= \int p(y_f|\boldsymbol{\theta}, \boldsymbol{y})\pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta} \\ &= \mathbb{E}_{\Theta|\boldsymbol{y}}\left\{p(y_f|\boldsymbol{\theta}, \boldsymbol{y})\right\}. \end{aligned}$$

En el caso de variables aleatorias intercambiables, i.e. cuando $Y_1, ..., Y_n$ son condicionalmente independientes dado el parámetro $\boldsymbol{\theta}$, el cálculo de la densidad final de y_f se obtiene

a través de

$$p(y_f|\boldsymbol{y}) = \int p(y_f|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta},$$

en vista de la independencia condicional de y_f y $\boldsymbol{y}$ dado $\boldsymbol{\theta}$. El problema de inferencia y predicción es en esencia un problema de decisión estadística bajo un ambiente de incertidumbre. En la siguiente sección describiremos brevemente los elementos que conforman un problema de decisión y la solución Bayesiana óptima en el caso de inferencia o predicción puntual e inferencia y predicción general.

2.2 Elementos de la Teoría de Decisión

El problema estadístico de inferencia y predicción es básicamente un problema de decisión en un ambiente de incertidumbre: Un problema de decisión general está compuesto por un espacio de *estados de la naturaleza*, que denotaremos por $\boldsymbol{\Omega}$. En este espacio está definido el elemento sobre el cual reside nuestra incertidumbre y sobre el cual no tenemos ningún control. El espacio donde tenemos un control directo define nuestras diferentes alternativas o cursos de acción respecto al fenómeno o variable que nos interesa, y básicamente representa nuestras opciones disponibles en la búsqueda de un objetivo. Este espacio lo denotamos por $\mathcal{A}$. Cada trayectoria de decisión está compuesta por la pareja (a, ω) donde $a \in \mathcal{A}$ es la acción o postura que hemos asumido respecto a la cantidad que nos interesa $\boldsymbol{\omega} \in \boldsymbol{\Omega}$, sobre la cual, como ya mencionamos, carecemos de control. Desde luego, las acciones tomadas nos conducirán a obtener diferentes resultados, que son desconocidos, y debemos de definir una escala de preferencias de manera que nuestras acciones sean consistentes y coherentes. Esta escala de preferencias sobre todas las posibles trayectorias de decisión la podemos definir a través una función de utilidad (o pérdida según sea el caso), inducida por nuestra relación de preferencia particular, y denotada por $u : \mathcal{A} \times \boldsymbol{\Omega} \to \Re_+$, (o $l = -u$ en el caso de una función de pérdida).

Con el enfoque Bayesiano toda la información sobre el estado de la naturaleza, $\boldsymbol{\omega}$, está resumida en una medida de probabilidad $p(\cdot)$ condicional en toda la información relevante disponible al momento de la toma de decisiones. La solución Bayesiana óptima consiste en

elegir la acción $a^* \in \mathcal{A}$ que maximice (minimice) la utilidad (pérdida) esperada (Bernardo y Smith, 1994, Capítulo 2).

En las siguientes subsecciones describiremos brevemente cómo se pueden obtener soluciones Bayesianas óptimas al problema de inferencia y predicción usando esta herramienta de toma de decisiones.

2.2.1 Estimación y Predicción Puntual

En algunas circunstancias, cuando la distribución de una v.a. Y está caracterizada por un valor parametral $\boldsymbol{\theta}$ desconocido, es de interés encontrar un valor específico de $\boldsymbol{\theta}$, digamos $\boldsymbol{\theta}^*$, que describa convenientemente la distribución de probabilidad de la v.a. $\boldsymbol{Y}$.

Claramente este es un problema de toma de decisiones en un ambiente de incertidumbre, donde el espacio de *estados de la naturaleza* y el espacio de acciones coinciden con el espacio parametral $\boldsymbol{\Theta}$. En este caso asignamos una medida de penalización a la acción de elegir un valor específico $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ respecto al verdadero valor de $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ (Bernardo y Smith, 1994). Por su naturaleza, esta función es conocida como *función de pérdida* y es denotada por $l(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$, que al ser función de $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ es una variable aleatoria.

Como ya mencionamos, la estrategia Bayesiana óptima consiste en elegir el valor $\boldsymbol{\theta}^*$ que minimice la función de pérdida esperada respecto a la distribución final de $\boldsymbol{\theta}$ dados los datos $\boldsymbol{y}$, i.e. elegiremos $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ tal que

$$\boldsymbol{\theta}^* = \arg\min \left\{ \mathbb{E}_{\boldsymbol{\Theta}|\boldsymbol{y}} \left[l(\boldsymbol{\theta}, \boldsymbol{\theta}^*) | \boldsymbol{y} \right] \right\}.$$

Usando, por ejemplo, la función de pérdida cuadrática, se tiene que el estimador puntual Bayesiano $\boldsymbol{\theta}^*$ es la media de la distribución final de $\boldsymbol{\theta}$, i.e. $\boldsymbol{\theta}^* = \mathbb{E}(\boldsymbol{\theta}|\boldsymbol{y})$. Otros estimadores puntuales, como la mediana y la moda de la distribución final de $\boldsymbol{\theta}$, pueden obtenerse como una solución alternativa si se utilizan ciertas funciones de pérdida (Bernardo y Smith, 1994).

Si nuestro interés reside en pronosticar un valor de la v.a. $\boldsymbol{Y}$, con base en observaciones previas de la misma, entonces los estimadores de pronóstico Bayesiano los construiremos bajo el criterio anterior en términos de la distribución *predictiva final* de $\boldsymbol{Y}$.

2.2.2 Inferencia y Predicción General

Supongamos que el interés del análisis estadístico es el de inferir respecto a un *estado de la naturaleza*, denotado por $\omega \in \Omega$, que se rige de manera aleatoria y sobre el cual nuestra información es limitada e inclusive en algunos casos faltantes. Las decisiones en este caso consisten en proporcionar alguna aseveración estadística respecto al valor incierto de ω, que desde un enfoque Bayesiano es resumida en una medida de probabilidad. Desde luego estas aseveraciones estarán condicionadas en la información relevante disponible al momento de la toma de decisiones, la cual en este caso denotamos por D, y que en términos generales está constituida por un conjunto de datos observados relacionados con el problema. En este caso el espacio de acciones estará definido por $\mathcal{A} = \{p_i(\cdot|D) : i \in I\}$, donde $p_i(\cdot|D)$ es una medida de probabilidad definida en Ω, para $i \in I$ con I un conjunto índice. Así, el conjunto de todas las posibles trayectorias del problema de decisión estarán denotadas por el conjunto $\mathcal{C} = \{c_i : i \in I\}$, donde $c_i = \{p_i(\cdot|D), \omega\}$ para todo $\omega \in \Omega$. La especificación de un problema de decisión general requiere establecer una relación de preferencia que cuantifique la consecuencia de decidir por el modelo $p_i(\cdot|D)$ cuando el estado de la naturaleza es ω.

La relación de preferencias se define en términos de una función de puntaje (Bernardo y Smith, 1994, definición 3.15) $u : \mathcal{A} \times \Omega \to \Re$. Así, la solución Bayesiana óptima consiste en elegir la distribución (o densidad) de la clase $\mathcal{A}$ que maximice en I la utilidad esperada

$$\overline{u}(p_i(\cdot|D)) = \int u(p_i(\cdot|D), \omega) p(\omega|D) d\omega, \tag{2.6}$$

donde $p(\omega|D)$ es la "verdadera" densidad de ω condicional en los datos observados D.

Se dice que una función de puntaje es *propia* si la utilidad esperada máxima se obtiene cuando $\sup_{i \in I} \overline{u}(p_i(\cdot|D)) = \overline{u}(p(\cdot|D))$, i.e. cuando la opción óptima es la "verdadera" densidad (distribución) para ω, y la función de puntaje es *local* si $u(p_i(\cdot|D), \omega) = u(p_i(\omega|D))$ para todo $\omega \in \Omega$, i.e. si depende sólo del valor de densidad (distribución) evaluada en ω.

Bernardo (1979) demostró que si una función de puntaje es propia y local, entonces debe ser de la forma

$$u\left(p_i(\cdot|D), \omega\right) = A \log p_i(\omega|D) + B(\omega), \tag{2.7}$$

para todo $\omega \in \Omega$, con $A > 0$ una constante real y $B(\cdot)$ una función integrable respecto a $p(\cdot|D)$. La función (2.7) es conocida como *función de puntaje logarítmico.*

2.3 Integración de Monte Carlo

Como vimos en las secciones anteriores, resolver un problema estadístico con el enfoque Bayesiano consiste operativamente en resolver integrales. En la práctica, muchas de estas integrales pueden ser difíciles de trabajar analíticamente. A través de la historia se han propuesto diferentes métodos para resolver algunos problemas de integración con estas características, algunos de los cuales consisten en aproximaciones numéricas deterministas o analíticas a la integral de interés. Las aproximaciones analíticas se basan en la aproximación de Laplace y resultan particularmente útiles para el caso de modelos cuya distribución pertenece a la familia exponencial (e.g., Tierney y Kadane (1986)). En esta sección describiremos el método de Monte Carlo, que sirve para aproximar integrales complejas mediante técnicas de simulación estocástica. Para efectos prácticos supongamos que deseamos resolver una integral de la forma $\int g(\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$, donde $\boldsymbol{\theta} \in \Theta \subset \Re^p$ es una variable aleatoria, $\pi(\cdot)$ es la densidad de $\boldsymbol{\theta}$, y $g(\cdot)$ es una función real conocida e integrable respecto a π. En el enfoque Bayesiano $\pi(\boldsymbol{\theta})$ estará condicionada en la información relevante disponible al momento del análisis, denotada por D, que por simplicidad en la notación es omitida en el transcurso de esta sección. Los resultados de esta sección son aplicables en los casos en que $\boldsymbol{\theta}$ represente algunos parámetros asociados a un modelo, o cuando represente variables aleatorias observables.

El método de Monte Carlo se basa en el supuesto que seamos capaces de generar una muestra de tamaño N, $\{\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(N)}\}$, de la distribución $\pi(\boldsymbol{\theta})$[1]. Usando esta muestra podemos aproximar el valor de la integral de interés, la cual podemos interpretar como el valor esperado de g,

$$\mathbb{E}_\pi\left[g(\boldsymbol{\theta})\right] = \int g(\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}, \tag{2.8}$$

[1] Por simplicidad π denotará a la distribución de θ y a su densidad de manera indistinta.

mediante el promedio empírico

$$\widehat{\mathbb{E}}_{\pi}\left[g(\boldsymbol{\theta})\right] = \frac{1}{N}\sum_{i=1}^{N} g(\boldsymbol{\theta}^{(i)}). \tag{2.9}$$

El estimador (2.9), conocido como el estimador de Monte Carlo de (2.8), es un estimador insesgado y converge casi seguramente al valor de la integral de interés. Cuando la esperanza de $g^2(\cdot)$ es finita respecto a $\pi(\cdot)$, la convergencia de (2.9) puede medirse en términos de su varianza teórica

$$var\left[\widehat{\mathbb{E}}_{\pi}\left[g(\boldsymbol{\theta})\right]\right] = \frac{1}{N}\int \left[g(\boldsymbol{\theta}) - \mathbb{E}[g(\boldsymbol{\theta})]\right]^2 \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{2.10}$$

la cual puede estimarse usando la misma muestra mediante su contraparte muestral

$$\widehat{var}\left[\widehat{\mathbb{E}}_{\pi}\left[g(\boldsymbol{\theta})\right]\right] = \frac{1}{N^2}\sum_{i=1}^{N}\left[g(\boldsymbol{\theta}^{(i)}) - \widehat{\mathbb{E}}_{\pi}\left[g(\boldsymbol{\theta})\right]\right]^2. \tag{2.11}$$

En un contexto de inferencia Bayesiana generalmente conocemos la densidad $\pi(\cdot)$ salvo por una constante de normalización, que usualmente es difícil de calcular, y de hecho nos remonta al problema inicial de resolver una integral como (2.8) con $g(\cdot)$ igual a la función constante unitaria. En este caso es difícil generar datos de la distribución $\pi(\cdot)$ directamente, y por ende es difícil aplicar el método de Monte Carlo. A través de la historia se han propuesto diferentes alternativas para generar datos de densidades conocidas salvo por una constante de normalización. Algunos de éstos los describiremos brevemente en las siguientes subsecciones.

2.3.1 Muestreo por Importancia

El muestreo por importancia consiste en suponer que tenemos acceso a una densidad $p(\cdot)$ "semejante" a la densidad de interés $\pi(\cdot)$, conocida como *función de densidad de importancia*, de la cual es relativamente simple generar datos muestrales. La idea consiste en utilizar estos datos para aproximar integrales de la forma (2.8) usando el método de Monte Carlo.

La base central de este método consiste en suponer que el soporte de $p(\cdot)$ contiene al soporte de la densidad de interés $\pi(\cdot)$, en cuyo caso (2.8) puede ser re-expresada como

$$\int g(\boldsymbol{\theta})\frac{\pi(\boldsymbol{\theta})}{p(\boldsymbol{\theta})}p(\boldsymbol{\theta})d\boldsymbol{\theta} = \mathbb{E}_{p}\left[g(\boldsymbol{\theta})w(\boldsymbol{\theta})\right],$$

donde $w(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})/p(\boldsymbol{\theta})$. De esta forma, si podemos generar una muestra de la densidad $p(\cdot)$, de tamaño N, $\{\boldsymbol{\theta}^{(i)} : i = 1, ..., N\}$, entonces podemos aproximar (2.8) mediante

$$\widehat{\mathbb{E}}_{\pi}[g(\boldsymbol{\theta})] = \frac{1}{N}\sum_{i=1}^{N} g(\boldsymbol{\theta}^{(i)})w(\boldsymbol{\theta}^{(i)}). \tag{2.12}$$

Cuando $\pi(\cdot)$ es conocida salvo por una constante de normalización, la aproximación (2.12) no puede ser usada directamente, sin embargo podemos expresar (2.8) como el cociente de dos esperanzas

$$\mathbb{E}_{\pi}[g(\boldsymbol{\theta})] = \frac{\int g(\boldsymbol{\theta})\frac{\pi(\boldsymbol{\theta})}{p(\boldsymbol{\theta})}p(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int \frac{\pi(\boldsymbol{\theta})}{p(\boldsymbol{\theta})}p(\boldsymbol{\theta})d\boldsymbol{\theta}}, \tag{2.13}$$

en cuyo caso $\mathbb{E}_{\pi}[g(\boldsymbol{\theta})]$ puede aproximarse como el cociente de dos aproximaciones de Monte Carlo como

$$\widehat{\mathbb{E}}_{\pi}[g(\boldsymbol{\theta})] = \sum_{i=1}^{N} g(\boldsymbol{\theta}^{(i)})\widetilde{w}(\boldsymbol{\theta}^{(i)}),$$

donde $\widetilde{w}(\boldsymbol{\theta}^{(i)}) = w(\boldsymbol{\theta}^{(i)})/\sum_{j=1}^{M} w(\boldsymbol{\theta}^{(j)})$ son los pesos asociados a cada dato $\boldsymbol{\theta}^{(i)}$, con $w(\boldsymbol{\theta}^{(i)})$ definida como antes, para $i = 1, ..., M$. En este caso se hace evidente que no necesitamos la constante de normalización de la densidad de interés $\pi(\cdot)$.

La convergencia de (2.12) a (2.8) se garantiza si elegimos $p(\cdot)$ de manera que su soporte contenga al soporte de la densidad de interés $\pi(\cdot)$. Una consideración adicional para tener una convergencia más rápida es que el cociente $\pi(\boldsymbol{\theta})/p(\boldsymbol{\theta})$ esté acotado para todos los valores de $\boldsymbol{\theta}$, y que adicionalmente las colas de $p(\cdot)$ sean más pesadas respecto a las colas de $\pi(\cdot)$. Una descripción detallada de este método se encuentra en Ripley (1987). Geweke (1989) describe algunas condiciones adicionales.

2.3.2 Muestreo-Remuestreo por Importancia

Este método extiende de manera natural las aproximaciones del método de muestreo por importancia.

El algoritmo funciona en dos etapas. En la primera etapa suponemos que tenemos una muestra de tamaño N de una densidad de importancia $p(\cdot)$, al igual que en la subsección anterior, i.e. tenemos una muestra $\{\boldsymbol{\theta}^{(i)} : i = 1, ..., N\}$ de la densidad $p(\boldsymbol{\theta})$, donde

cada $\boldsymbol{\theta}^{(i)}$ tiene un peso asociado $\widetilde{w}(\boldsymbol{\theta}^{(i)})$, definido como en la subsección anterior. La segunda etapa del algoritmo consiste en generar N muestras con reemplazo de los valores $\{\boldsymbol{\theta}^{(i)} : i = 1, ..., N\}$ de acuerdo a sus correspondientes pesos $\{\widetilde{w}(\boldsymbol{\theta}^{(i)}) : i = 1, ..., N\}$. De esta forma podemos aproximar (2.8) por

$$\widehat{\mathbb{E}}\left[g(\boldsymbol{\theta})\right] = \sum_{j=1}^{N} g(\widetilde{\boldsymbol{\theta}}^{(j)})\widetilde{w}(\widetilde{\boldsymbol{\theta}}^{(j)}), \tag{2.14}$$

donde $\{\widetilde{\boldsymbol{\theta}}^{(j)} : j = 1, ..., N\}$ son una muestra de la variable discreta $\widetilde{\boldsymbol{\Theta}} = \{\boldsymbol{\theta}^{(i)} : i = 1, ..., N\}$, donde cada $\boldsymbol{\theta}^{(i)}$ tiene asociada una masa de probabilidad $\widetilde{w}(\boldsymbol{\theta}^{(i)})$.

Más aún, con este procedimiento podemos aproximar características de $\pi(\cdot)$ que no pueden ser expresadas en forma de esperanza, como cuantiles e intervalos de credibilidad, ya que la distribución que asigna una masa $\widetilde{w}(\boldsymbol{\theta}^{(i)})$ a $\boldsymbol{\theta}^{(i)}$ en $\widetilde{\boldsymbol{\Theta}}$ tiende en distribución a $\pi(\boldsymbol{\theta})$ cuando $N \to \infty$ (Smith y Gelfand, 1992).

Este procedimiento es flexible y rico, en el sentido que podemos obtener muestras aproximadas que nos permiten reconstruir a $\pi(\cdot)$, por ejemplo a través de histogramas o aproximaciones por *kernel* (vea el apéndice B.2). Además permite implementar el Teorema de Bayes de manera directa, donde $\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{y})$. Si podemos generar una muestra aleatoria $\{\boldsymbol{\theta}^{(i)} : i = 1, ..., N\}$ de $\pi(\boldsymbol{\theta})$, podemos actualizarla a través de la verosimilitud para obtener una muestra $\{\widetilde{\boldsymbol{\theta}}^{(j)}\}$ de tamaño N, que se distribuya aproximadamente como $\pi(\boldsymbol{\theta}|\boldsymbol{y})$, remuestreando de $\{\boldsymbol{\theta}^{(i)} : i = 1, ..., M\}$, donde cada $\boldsymbol{\theta}^{(i)}$ tiene asociado una masa de probabilidad definida como $\widetilde{w}(\boldsymbol{\theta}^{(j)}) = p(\boldsymbol{y}|\boldsymbol{\theta}^{(j)}) / \sum_{i=1}^{N} p(\boldsymbol{y}|\boldsymbol{\theta}^{(i)})$ para $j = 1, ..., N$.

2.3.3 Monte Carlo vía Cadenas de Markov

Otro método importante para generar muestras de una distribución de probabilidad $\pi(\cdot)$ de interés, es construyendo una cadena de Markov cuya distribución invariante sea nuestra distribución objetivo $\pi(\cdot)$.

Supongamos que podemos construir una cadena de Markov homogénea $(\boldsymbol{\theta}^{(n)})_{n\geq 1}$ en tiempo discreto, con un espacio de estados $\boldsymbol{\Theta} \subset \Re^p$. En un esquema general, esta cadena de Markov está determinada mediante una función $K : \boldsymbol{\Theta} \times \mathcal{B}(\boldsymbol{\Theta}) \to [0, 1]$ de transición

de estados, conocida como *kernel de transición*, donde $\mathcal{B}(\Theta)$ es el σ-álgebra de Borel inducido por Θ. En el caso que Θ sea continuo, el *kernel* de transición denota a la densidad condicional de transición, $K(\boldsymbol{\theta}, \boldsymbol{\theta}')$, tal que $P(\Theta \in A|\boldsymbol{\theta}) = \int_A K(\boldsymbol{\theta}, d\boldsymbol{\theta}')$. Cuando el espacio de estado Θ es discreto, el *kernel* de transición denota la probabilidad de transición $K(\boldsymbol{\theta}, \boldsymbol{\theta}') = P(\Theta^{(k+1)} = \boldsymbol{\theta}'|\Theta^{(k)} = \boldsymbol{\theta})$ para todo $\boldsymbol{\theta}$ y $\boldsymbol{\theta}' \in \Theta$ entre las iteraciones k y $k+1$.

La idea central del método de Monte Carlo vía Cadenas de Markov (MCCM) es que la cadena de transición definida por un *kernel* de transición $K(\cdot, \cdot)$ tenga a $\pi(\cdot)$, la distribución de interés, como distribución *invariante*[2]. Este enfoque de análisis de cadenas de Markov es inverso al enfoque tradicional, ya que debemos construir una cadena partiendo de la distribución invariante, en lugar de construir una cadena con un *kernel* arbitrario y verificar si cumple con las condiciones de estabilidad. Si somos capaces de definir un *kernel* de transición que satisfaga la condición de balance $K(\boldsymbol{\theta}, \boldsymbol{\theta}')\pi(\boldsymbol{\theta}) = K(\boldsymbol{\theta}', \boldsymbol{\theta})\pi(\boldsymbol{\theta}')$ para todo $\boldsymbol{\theta}$ y $\boldsymbol{\theta}' \in \Theta$, entonces tenemos que la cadena de Markov construida con este *kernel* tiene como densidad invariante a $\pi(\cdot)$ (Robert y Casella, 1999, Teorema 6.2.2). Si la cadena es *irreducible*[3] y *aperiódica*[4], entonces (Tierney, 1994; Robert y Casella, 1999)

- $\boldsymbol{\theta}^{(k)} \xrightarrow{d} \boldsymbol{\theta} \backsim \pi$, y
- $\frac{1}{N}\sum_{k=1}^{N} g(\theta^{(k)}) \to \int g(\theta)\pi(\theta)d\theta$, casi seguramente cuando $N \to \infty$.

En las siguientes subsecciones describiremos diferentes métodos para construir cadenas de Markov con estas características.

[2] π es la densidad invariante de la cadena de Markov definida por el *kernel* $K(\cdot, \cdot)$ si $\boldsymbol{\theta}^{(k)} \sim \pi$ implica que $\boldsymbol{\theta}^{(k+1)} \sim \pi$, i.e. $\lim_{k\to\infty} K^k(\boldsymbol{\theta}, A) = \pi(A)$, para todo $A \in \mathcal{B}(\Theta)$.

[3] Una cadena de Markov es *irreducible* (π-irreducible) si para todo $\boldsymbol{\theta} \in E \in \mathcal{B}(\Theta)$ tal que $\pi(E) > 0$ se tiene que para todo $A \in \mathcal{B}(\Theta)$ con $\pi(A) > 0$ existe algún entero n tal que $K^n(\boldsymbol{\theta}, A) > 0$, i.e. si existe la libertad de que la cadena se mueva sobre todo el espacio de estados.

[4] Una cadena de Markov es *aperiódica* si no existe una partición $\{E_0, ..., E_{d-1}\}$ del espacio de estados Θ tal que $K(\boldsymbol{\theta}, E_j) = 1$ para todo $\boldsymbol{\theta} \in E_{j-1}$, i.e. no existe una trayectoria determinista de visitas a subconjuntos de Θ.

2.3.4 Algoritmo de Metropolis-Hastings (M-H)

Para una cadena de Markov $(\boldsymbol{\theta}^{(k)})_{k\geq 1}$, elegimos una familia de densidades $q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ parametrizadas por $\boldsymbol{\theta}$, i.e. para un valor de $\boldsymbol{\theta}$ fijo $q(\boldsymbol{\theta}, \cdot)$ es una densidad con el mismo soporte que la densidad objetivo $\pi(\cdot)$. La elección de esta familia es arbitraria con el requisito que la cadena definida por la densidad de transición $P(\boldsymbol{\Theta}^{(k+1)} = \boldsymbol{\theta}' | \boldsymbol{\Theta}^{(k)} = \boldsymbol{\theta}) = q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ sea irreducible, y que satisfaga la condición de balance.

El algoritmo funciona de la siguiente manera. Dado un estado actual de la cadena, digamos $\boldsymbol{\Theta}^{(k)} = \boldsymbol{\theta}^{(k)}$, un valor $\boldsymbol{\theta}'$ es propuesto para el estado $\boldsymbol{\Theta}^{(k+1)}$ con base en la densidad de transición $q(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}')$, y es aceptado con una probabilidad

$$\alpha(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}') = \text{mín}\left(1, \frac{q(\boldsymbol{\theta}', \boldsymbol{\theta}^{(k)})\pi(\boldsymbol{\theta}')}{q(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}')\pi(\boldsymbol{\theta}^{(k)})}\right), \tag{2.15}$$

i.e. con probabilidad α el valor de la cadena en la iteración $k+1$ es $\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}'$, de lo contrario $\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)}$. Este esquema de muestreo define una cadena de Markov con un *kernel* de transición de la iteración k a $k+1$ dada por

$$K(\boldsymbol{\theta}^{(k)}, d\boldsymbol{\theta}') = q(\boldsymbol{\theta}^{(k)}, d\boldsymbol{\theta}')\alpha(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}') + \left(1 - \int \alpha(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}')q(\boldsymbol{\theta}^{(k)}, d\boldsymbol{\theta}')\right)\delta_{\boldsymbol{\theta}^{(k)}}(d\boldsymbol{\theta}').$$

Tierney (1994) demostró que una cadena de Markov construida de esta forma es reversible y aperiódica, con lo cual se tiene que $\pi(\cdot)$ es su correspondiente distribución estacionaria. Este algoritmo es particularmente útil en el contexto de inferencia Bayesiana, donde en algunas ocasiones $\pi(\boldsymbol{\theta})$ es conocida salvo su constante de normalización, pues la distribución de interés π sólo es usada a través del cociente $\pi(\boldsymbol{\theta}')/\pi(\boldsymbol{\theta}^{(k)})$ en (2.15).

Utilizando este esquema de muestreo es posible determinar diferentes algoritmos de actualización, dependiendo de la definición de la distribución q por utilizar. El algoritmo original considera un esquema de muestreo independiente, i.e. $q(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}') = q(\boldsymbol{\theta}')$, en cuyo caso el cociente en (2.15) se reduce al cociente $w(\boldsymbol{\theta}')/w(\boldsymbol{\theta}^{(k)})$, donde $w(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})/q(\boldsymbol{\theta})$ denota los pesos de importancia definidos previamente para aproximar integrales empleando como distribución de importancia a q. Otra alternativa consiste en definir q como una distribución simétrica, i.e. $q(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}') = q(\boldsymbol{\theta}', \boldsymbol{\theta}^{(k)})$, en cuyo caso el cociente en (2.15) se simplifica de la forma $\pi(\boldsymbol{\theta}')/\pi(\boldsymbol{\theta}^{(k)})$. Otra posibilidad consiste en definir $q(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}')$ a

través de la relación $\Theta = \boldsymbol{\theta}^{(k)} + \boldsymbol{Z}$, donde $\boldsymbol{Z}$ es una variable aleatoria con media cero y función de distribución r. En este caso q se define a través de una caminata aleatoria, de manera que el algoritmo se concentra en explorar vecindades contiguas al estado previo de la cadena en el espacio de estados de la cadena. Es deseable que la distribución r sea simétrica. Para el caso de espacios no acotados, la distribución Normal (multivariada) o t son dos alternativas útiles y simples.

2.3.5 Muestreador de Gibbs

En algunas ocasiones tenemos el interés de obtener una muestra de una distribución $\pi(\cdot)$ multivariada. De esta forma, para obtener una muestra de $\pi(\cdot)$ mediante MCCM es necesario construir una cadena de Markov con un espacio de estado multivariado. En este caso el muestreador de Gibbs resulta un método práctico para construir tales cadenas preservando las características antes mencionadas.

Para estos efectos supongamos que la distribución de interés π corresponde a una variable aleatoria p-dimensional Θ, y que por razones prácticas podemos descomponer este espacio en $q \leq p$ componentes, denotados por $\Theta_1, ..., \Theta_q$, algunos de éstos posiblemente multivariados, y denotemos por Θ_{-l} a los componentes de Θ menos el l-ésimo, para $l = 1, 2, ..., q$.

Dados los valores de la cadena en la iteración k, $\boldsymbol{\theta}^{(k)} = (\boldsymbol{\theta}_1^{(k)}, ..., \boldsymbol{\theta}_q^{(k)})$, producimos la transición al estado $\boldsymbol{\theta}^{(k+1)}$ mediante un muestreo sucesivo de las distribuciones condicionales completas mediante el siguiente esquema:

$$\begin{aligned} \boldsymbol{\theta}_1^{(k+1)} &\sim \pi(\boldsymbol{\theta}_1 | \boldsymbol{\theta}_2^{(k)}, ..., \boldsymbol{\theta}_q^{(k)}) \\ \boldsymbol{\theta}_2^{(k+1)} &\sim \pi(\boldsymbol{\theta}_2 | \boldsymbol{\theta}_1^{(k+1)}, \boldsymbol{\theta}_3^{(k)}, ..., \boldsymbol{\theta}_q^{(k)}) \\ &\vdots \\ \boldsymbol{\theta}_q^{(k+1)} &\sim \pi(\boldsymbol{\theta}_q | \boldsymbol{\theta}_1^{(k+1)}, ..., \boldsymbol{\theta}_{q-1}^{(k+1)}). \end{aligned} \tag{2.16}$$

La estructura de actualización de los componentes dentro del algoritmo (2.16) puede definirse de manera aleatoria o determinista, considerando que cada componente es actualizado en cada ciclo al menos una vez. De esta forma la transición del estado $\boldsymbol{\theta}^{(k)}$ al

estado $\boldsymbol{\theta}^{(k+1)}$ está determinada por:

$$K(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k+1)}) = \prod_{l=1}^{q} \pi(\boldsymbol{\theta}_l^{(k+1)} | \boldsymbol{\theta}_j^{(k+1)}, j < l, \boldsymbol{\theta}_l^{(k)}, j > l).$$

En este algoritmo debe de considerarse la estructura de dependencia de los componentes individuales, y es recomendable agrupar en un bloque a aquellos componentes escalares que estén altamente correlacionados, para evitar que la cadena retarde su entrada al periodo de estabilidad. El muestreador de Gibbs puede ser visto como un caso particular del algoritmo de Metropolis-Hastings. En este caso el proceso de actualización se realiza en cada uno de los q componentes de $\boldsymbol{\Theta}$ de la manera antes mencionada, entonces el valor propuesto de la cadena para el siguiente estado en cada l-ésimo componente es actualizado, con probabilidad 1, de la densidad $\pi(\boldsymbol{\theta}_l^{(k+1)} | \boldsymbol{\theta}_j^{(k+1)}, j < l, \boldsymbol{\theta}_l^{(k)}, j > l)$.

Para implementar el muestreador de Gibbs de manera directa es necesario conocer de manera cerrada cada uno de las distribuciones condicionales completas y tener la capacidad de muestrear datos de ellas también directamente. En algunas ocasiones no es posible obtener de manera cerrada algunas de las distribuciones condicionales completas, que se conocen salvo su constante de normalización. En este caso, es posible diseñar el muestreador de Gibbs incorporando en la etapa de muestreo de la condicional no normalizada la generación de muestras de una distribución instrumental e incorporando ésta al proceso de muestreo mediante un paso adicional de importancia para la aceptación de esta muestra. Este algoritmo se conoce como muestreador de Gibbs por Importancia (Müller, 1993). Supongamos, sin pérdida de generalidad, que no es posible obtener la distribución condicional completa del j-ésimo componente en (2.16), y que podemos definir una distribución instrumental $q(\boldsymbol{\theta}_j | \boldsymbol{\theta}_{-j})$ completamente determinada que aproxima a $\pi(\boldsymbol{\theta}_j | \boldsymbol{\theta}_{-j})$. Esta distribución se deriva de la distribución instrumental $q(\boldsymbol{\theta})$ que aproxima a la distribución final completa $\pi(\boldsymbol{\theta})$. Así, en la j-ésima etapa de muestreo correspondiente de (2.16), donde el estado parcialmente actualizado de la cadena es $\boldsymbol{\theta}_{-j} = (\boldsymbol{\theta}_1^{(k+1)}, ..., \boldsymbol{\theta}_{j-1}^{(k+1)}, \boldsymbol{\theta}_{j+1}^{(k)}, ..., \boldsymbol{\theta}_q^{(k)})'$, para la $k+1$-ésima iteración de la cadena, definimos el cociente de importancia $w(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})/q(\boldsymbol{\theta})$ y aceptamos la muestra $\boldsymbol{\theta}_j'$, generada por $q(\boldsymbol{\theta}_j | \boldsymbol{\theta}_{-j})$, con una probabilidad $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}') = \text{mín}\{1, w(\boldsymbol{\theta}')/w(\boldsymbol{\theta})\}$, donde $\boldsymbol{\theta}' = (\boldsymbol{\theta}_1^{k+1}, ..., \boldsymbol{\theta}_{j-1}^{(k+1)}, \boldsymbol{\theta}_j', \boldsymbol{\theta}_{j+1}^{(k)}, ..., \boldsymbol{\theta}_q^{(k)})'$ y $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^{k+1}, ..., \boldsymbol{\theta}_{j-1}^{(k+1)}, \boldsymbol{\theta}_j^{(k)}, \boldsymbol{\theta}_{j+1}^{(k)}, ..., \boldsymbol{\theta}_q^{(k)})'$, en otro

caso $\boldsymbol{\theta}_j^{(k+1)} = \boldsymbol{\theta}_j^{(k)}$. Los pesos $w(\boldsymbol{\theta})$ en la probabilidad de aceptación son los mismos que empleamos para aproximar integrales mediante el muestreo por importancia, previamente discutido, así que las consideraciones presentadas para la distribución instrumental q tienen el mismo significado en este esquema. El componente aleatorio de aceptación garantiza que la cadena tenga a π como distribución invariante.

Los esquemas de muestreo que describimos previamente representan dos alternativas flexibles para implementar procedimientos Bayesianos de inferencia. Para problemas donde no puedan ser empleados de manera directa, se pueden definir diferentes combinaciones de éstos o de algunas generalizaciones, en diferentes etapas y bajo ciertas restricciones. Esta combinaciones dan origen a lo que se conoce como métodos híbrido de MCCM (Tierney, 1994).

Por otro lado, éstos esquemas de muestreo están diseñados para generar muestras o una cadena de Markov, de una distribución π definida sobre un espacio de dimensión fija. En la siguiente sección describimos un método de muestreo diseñado medidas de probabilidad π definidas sobre un espacio de dimensiones cambiantes, en cuyo caso se generan muestras de medidas de probabilidad degeneradas en subespacios del espacio general de interés.

2.3.6 MCCM con Salto Reversible

En algunos problemas como mezcla o selección Bayesiana de modelos (vea el capítulo 3), la distribución de interés π está definida en un espacio parametral *general*, denotado por $\boldsymbol{\Theta}$, formado por la unión de los diferentes subespacios parametrales asociados a cada modelo considerado en la mezcla o selección. A su vez los modelos son indexados por un conjunto índice I (finito o numerable). De esta forma el espacio parametral está representado de la forma

$$\boldsymbol{\Theta} = \bigcup_{m \in I} \{m\} \times \boldsymbol{\Theta}_m, \tag{2.17}$$

donde $\boldsymbol{\Theta}_m \subset \Re^{n(m)}$ es el espacio parametral asociado el modelo indexado por m, y $n(m) \geq 1$ es su correspondiente dimensión. En este caso, es natural pensar en una

forma jerarquizada de la distribución de interés de la forma

$$\pi(m, \boldsymbol{\theta}_m|\boldsymbol{y}) = p(m|\boldsymbol{y})\pi(\boldsymbol{\theta}_m|m, \boldsymbol{y}), \tag{2.18}$$

donde $p(m|\boldsymbol{y})$ es la probabilidad final del modelo $m \in I$; $\boldsymbol{\theta}_m \in \boldsymbol{\Theta}_m$ es el parámetro asociado al modelo m; y $\pi(\boldsymbol{\theta}_m|m, \boldsymbol{y})$ es la densidad final de $\boldsymbol{\theta}_m$ dado el modelo m y la muestra $\boldsymbol{y}$.

Para implementar el muestreo por MCCM sobre $\boldsymbol{\Theta}$, se necesita definir una estrategia de saltos entre los diferentes subespacios de manera que (2.18) sea la distribución invariante de la cadena $\{(m^{(k)}, \boldsymbol{\theta}^{(k)}_{m^{(k)}})\}_{k\geq 1}$.

Green (1995) propuso una metodología para muestrear sobre distribuciones de la forma (2.18), que básicamente es una extensión del algoritmo de M-H, en el sentido que ambos construyen cadenas de Markov reversibles con distribución invariante π. Sin embargo, en este algoritmo la distribución propuesta para la evolución de la cadena en el espacio de estados de los parámetros es una distribución degenerada en uno de los subespacios del espacio general (2.17). La idea de este algoritmo consiste en generar un nuevo valor de la cadena, $(m', \boldsymbol{\theta}'_{m'})$, condicional en el estado de la cadena en la k-ésima iteración, $(m^{(k)}, \boldsymbol{\theta}^{(k)}_{m^{(k)}})$. La construcción del nuevo valor propuesto en la cadena preserva la estructura jerárquica de (2.18). En primer lugar, se genera un nuevo valor índice del modelo, m' mediante la distribución de transición $J(m^{(k)}, m')$ entre los índices de los modelos.

La segunda etapa del muestreo, condicional en m', consiste en obtener una muestra $\boldsymbol{\theta}'_{m'}$ de $\pi(\boldsymbol{\theta}_{m'}|m', \boldsymbol{y})$. El problema en esta etapa consiste en definir de manera adecuada el salto entre los dos subespacios $\boldsymbol{\Theta}_{m^{(k)}}$ y $\boldsymbol{\Theta}_{m'}$ de manera que el nuevo estado de la cadena, $\boldsymbol{\theta}'_{m'}$, dependa del estado actual de la cadena, $\boldsymbol{\theta}^{(k)}_{m^{(k)}}$. La idea de Green consiste en definir un emparejamiento de las dimensiones de ambos espacios y suponer que existe una biyección entre ellos, que puede definirse a través de la incorporación de variables auxiliares de la forma $g_{m^{(k)},m'} : \boldsymbol{\Theta}_{m^{(k)}} \times \boldsymbol{U}_{m^{(k)},m'} \rightarrow \boldsymbol{\Theta}_{m'} \times \boldsymbol{U}_{m',m^{(k)}}$ que mapea entre los espacios emparejados expandidos, donde $\boldsymbol{U}_{m^{(k)},m'}$ y $\boldsymbol{U}_{m',m^{(k)}}$ denotan variables auxiliares definidas de manera que $\dim(\boldsymbol{\Theta}_{m^{(k)}} \times \boldsymbol{U}_{m^{(k)},m'}) = \dim(\boldsymbol{\Theta}_{m'} \times \boldsymbol{U}_{m',m^{(k)}})$.

Supongamos que el estado propuesto en la iteración $k+1$ es $(m', \boldsymbol{\theta}'_{m'})$, la probabilidad

de aceptación de estos valores como el nuevo estado de la cadena está dada por

$$\alpha_{m^{(k)},m'}(\boldsymbol{\theta}^{(k)}_{m^{(k)}}, \boldsymbol{\theta}'_{m'}) = \min\left(1, r_{m^{(k)},m'}(\boldsymbol{\theta}^{(k)}_{m^{(k)}}, \boldsymbol{\theta}'_{m'})\right), \tag{2.19}$$

donde

$$\begin{aligned} r_{m^{(k)},m'}(\boldsymbol{\theta}^{(k)}_{m^{(k)}}, \boldsymbol{\theta}'_{m'}) &= \frac{p(m')\pi(\boldsymbol{\theta}'_{m'}|m')J(m', m^{(k)})}{p(m^{(k)})\pi(\boldsymbol{\theta}^{(k)}_{m^{(k)}}|m^{(k)})J(m^{(k)}, m')} \\ &\times \frac{q_{m',m^{(k)}}(\boldsymbol{\theta}'_{m'}, \boldsymbol{u}_{m',m^{(k)}})}{q_{m^{(k)},m'}(\boldsymbol{\theta}^{(k)}_{m^{(k)}}, \boldsymbol{u}_{m^{(k)},m'})} \\ &\times \mathcal{J}_{g_{m^{(k)},m'}}, \end{aligned} \tag{2.20}$$

con

$$\mathcal{J}_{g_{m^{(k)},m'}} = \left|\det \frac{\partial g_{m^{(k)},m'}(\boldsymbol{\theta}_{m^{(k)}}, \boldsymbol{u}_{m^{(k)},m'})}{\partial \boldsymbol{\theta}_{m^{(k)}} \boldsymbol{u}_{m^{(k)},m'}}\right|, \tag{2.21}$$

así, con probabilidad $\alpha_{m^{(k)},m'}(\boldsymbol{\theta}^{(k)}_{m^{(k)}}, \boldsymbol{\theta}'_{m'})$ aceptamos los valores propuestos como el nuevo estado de la cadena en la iteración $k+1$.

Green (1995) y Waagepetersen y Sorensen (2001) brindan una descripción detallada de este algoritmo de muestreo. Este algoritmo esta diseñado para utilizar la información del estado actual en la propuesta de movimiento en cada iteración, aunque podemos utilizar ciertas modificaciones para simplificar la generación de estados propuestos. De acuerdo con Godsill (2001) es posible proponer nuevos posibles valores de espacio parametral de cada modelo de manera independiente, preservando que la distribución (2.18) sea la distribución invariante de la cadena. En este caso no es necesario realizar un emparejamiento de los espacios parametrales de cada modelo y de esta manera el Jacobiano (2.21) sería eliminado de la probabilidad de aceptación del movimiento. Esta es una alternativa simple, sin embargo la convergencia de la cadena podría ser más lenta, de acuerdo a la naturaleza y estructura de los modelos, como es apuntado por Godsill (2001).

En algunos problemas específicos las distribuciones finales de cada uno de los modelos, i.e. $\pi(\boldsymbol{\theta}_m|m, \boldsymbol{y})$ para cada m, son conocidas de manera cerrada. En este caso, si podemos generar muestras directamente de las distribuciones $\pi(\boldsymbol{\theta}_{m'}|m', \boldsymbol{y})$, eliminamos la necesidad de emparejar los espacios parametrales entre los modelos y simplificamos la expresión de la probabilidad de aceptación de movimiento propuesto, como

$$\alpha_{m^{(k)},m'}(\boldsymbol{\theta}^{(k)}_{m^{(k)}}, \boldsymbol{\theta}'_{m'}) = \left(1, \frac{p(m')p(\boldsymbol{y}|m')J(m', m^{(k)})}{p(m^{(k)})p(\boldsymbol{y}|m^{(k)})J(m^{(k)}, m')}\right), \tag{2.22}$$

donde $p(\boldsymbol{y}|m) = \int p(\boldsymbol{y}|\boldsymbol{\theta}_m, m)\, \pi(\boldsymbol{\theta}_m|m) d\boldsymbol{\theta}_m$ denota la *verosimilitud integrada* del modelo m.

Mediante el MCCMSR, podemos ahorrarnos el trabajo de muestrear paralelamente en cada subespacio, y obtener una muestra que estará concentrada en los espacios que tenga una mayor probabilidad final, sin la necesidad de evaluar explícitamente éstas.

Consideraciones Generales

Existen diferentes consideraciones que se deben tomar en cuenta al momento de implementar algún algoritmo de muestreo mediante MCCM. Supongamos que $\{\boldsymbol{\theta}^{(k)}\}_{k\geq 1}$ es una trayectoria de una cadena de Markov con distribución invariante π. La primera consideración sobre el uso de MCCM es determinar cuándo la cadena de la cual estamos simulando entra en su fase de equilibrio. Esta no es una tarea simple, y empíricamente es difícil asegurar este comportamiento, sin embargo existen diferentes método para su monitoreo, como por ejemplo, graficar la trayectoria o traza de la cadena y los promedios actualizados de cada uno de sus componentes. En la práctica es usual definir un periodo inicial de longitud considerable de manera que la cadena presumiblemente entre en su fase de equilibrio, para tratar de garantizar que la cadena no se afecte por el valor inicial de la cadena. Por otro lado, suponiendo que la cadena se encuentra dentro de su fase de equilibrio, es evidente que debido a la estructura de Markov los datos muestreados no son independientes. Para reducir este efecto podemos obtener submuestras espaciadas de la trayectoria de la cadena simulada, de manera que la autocorrelación entre los datos no sea significativa. La longitud del submuestreo es arbitraria y se determina a partir de un análisis exploratorio de la trayectoria de la cadena. Alternativamente, para garantizar una muestra independiente de π, es posible generar un gran número de cadenas de manera simultánea y conservar los valores observados de cada cadena después de un periodo o longitud adecuada, de manera que la cadena este en su fase de equilibrio. Esta alternativa es poco eficiente ya que implica un costo computacional demasiado elevado durante su implementación.

Para el caso de MCCM con salto reversible, la verificación de convergencia de la cadena

es aún más complicada. En principio, es posible monitorear la convergencia de la cadena para el movimiento entre modelos o en aquellas cantidades o parámetros comunes entre ellos. Pero dado que las visitas de los modelos son aleatorias, el número de muestras dentro de cada modelos es aleatorio y para algunos de los modelos las cantidades observadas no son suficientes para monitorear la convergencia de la cadena dentro de esos modelos en particular.

En la siguiente sección realizaremos una breve descripción acerca de las series de tiempo. Describiremos algunos de los modelos más usados para el análisis y predicción de series de tiempo, así como algunos métodos de inferencia y predicción Bayesiana.

2.4 Series de Tiempo

Una serie de tiempo es un caso particular de un *Proceso Estocástico* $\{Y(t) : t \in T\}$ donde T es un conjunto índice y Y la variable aleatoria asociada a los elementos de T. La asociación de las variables con el conjunto índice se determina de manera única, i.e. a cada elemento del conjunto índice le corresponde sólo una variable aleatoria. Cuando el conjunto índice T representa una escala de tiempo el proceso $\{Y(t) : t \in T\}$ es conocido como una *Serie de Tiempo.* Si el conjunto T es medido en una escala continua se dice que es una *Serie de Tiempo Continua*, cuando el conjunto T es un subconjunto de los números enteros se dice que es una *Serie de Tiempo Discreta*, y su notación es $\{Y_t : t \in T\}$.

Se dice que una serie de tiempo $\{Y_t\}$ es *estacionaria estricta* si para todo par de números $s > 0$ y $q > 0$ se cumple $(Y_{1,\ldots,}Y_s) \stackrel{d}{=} (Y_{1+q,\ldots,}Y_{s+q})$, i.e. la distribución de cualquier segmento de la serie es invariante en el tiempo. Esta característica es difícil de verificar empíricamente, es por eso que en la práctica el concepto de estacionariedad se relaja al ser sustituido por un concepto alternativo y flexible pero que representa la naturaleza de la definición y que es más sencillo de verificar. En este sentido decimos que una serie de tiempo $\{Y_t\}$ es *Estacionaria Débil* si $\mathbb{E}(Y_t)$ y $V(Y_t)$ existen y son independientes de $t \in T$, y $Cov\,(Y_t, Y_{t-k})$ es independiente de t para toda $k \in T$.

El ejemplo más simple de un proceso estacionario es el de *ruido blanco*, usado para modelar factores de ruido o alteraciones estocásticas no controlables. Este proceso se

define como una sucesión de variables aleatorias $\{\varepsilon_t\}$ independientes e idénticamente distribuidas (*iid*) con media cero y varianza constante en el tiempo. Cuando además se supone que las variables tienen una distribución Gaussiana o Normal, se dice que $\{\varepsilon_t\}$ es una sucesión de *ruido blanco Gaussiano*, generalmente representado como $\varepsilon_t \overset{iid}{\sim} N(\varepsilon_t|0,\sigma^2)$. A continuación describiremos algunos modelos de series de tiempo capaces de representar una gran variedad de procesos.

2.4.1 Modelos Autorregresivos Lineales

Los modelos autorregresivos han sido ampliamente usados para analizar series de tiempo. En estos modelos se supone que el nivel de la serie de interés depende de valores pasados de la misma serie en cada tiempo t. Esta es una forma simple e intuitiva de definir alguna dependencia en series de tiempo. La dependencia de la serie con ciertas realizaciones pasadas del mismo proceso puede definirse de diferentes formas, siendo la forma lineal la manera más simple y usual de definir dicha dependencia.

Sea $\{Z_t\}$ una serie de tiempo escalar y μ_t su nivel medio al tiempo t, posiblemente variable en el tiempo. Se dice que la serie de tiempo es autorregresiva lineal de orden $p \geq 0$, denotado por AR(p), si

$$Z_t = \mu_t + \sum_{i=1}^{p} \phi_i \left(Z_{t-i} - \mu_{t-i}\right) + \varepsilon_t \tag{2.23}$$

donde $\{\varepsilon_t\}$ es una sucesión de variables aleatorias no correlacionadas, con media cero y varianza σ^2, y $\boldsymbol{\phi} = (\phi_1, ..., \phi_p)'$ es el vector de parámetros de autorregresión, usualmente constantes. Se supone que la sucesión de variables aleatorias $\{\varepsilon_t\}$ son condicionalmente independientes de los valores pasados de la serie $\{Z_t\}$. En los modelos tradicionales se supone que la sucesión de perturbaciones aleatorias $\{\varepsilon_t\}$ sigue un proceso ruido blanco Gaussiano.

Un proceso autorregresivo lineal, definido como en (2.23), puede expresarse de forma simplificada a través del polinomio de retraso $\boldsymbol{\phi}(B) = 1 - \sum_{j=1}^{p} \phi_j B^j$, de la forma

$$\boldsymbol{\phi}(B) Y_t = \varepsilon_t, \tag{2.24}$$

donde $Y_t = Z_t - \mu_t$ para todo t, y B es el operador de retraso usual, definido como $B^n Y_t = Y_{t-n}$ para todo $n \geq 0$. El polinomio característico puede factorizarse como $\phi(B) = \prod_{i=1}^{p}(1 - \alpha_i B)$ con los coeficientes α_i reales o complejos. La condición para que un proceso autorregresivo $\{Y_t\}$ sea estacionario es que todas las raíces del polinomio $\phi(B)$, que son las soluciones a la ecuación $1 - \phi_1 B - \ldots - \phi_p B^p = 0$, sean de módulo mayor a uno, o equivalentemente que todos los coeficientes α_i sean menores o iguales a uno en módulo. De manera natural, un proceso autorregresivo lineal es *no estacionario* si el módulo de alguna de las raíces del polinomio característico $\phi(B)$ es menor o igual a uno, o equivalentemente que alguno de los coeficientes α_i sea mayor a uno en módulo.

Supongamos que $\{y_t\}_{t=1}^{T}$ es la realización de un proceso AR(p) entre los tiempos 1 y T. En este caso, si suponemos que los primeros p valores iniciales del proceso son conocidos, el modelo (2.24) puede expresarse de manera matricial como

$$\boldsymbol{y}_T = \boldsymbol{Y}_{T,p}\boldsymbol{\phi} + \boldsymbol{\varepsilon}_T, \tag{2.25}$$

donde $\boldsymbol{y}_T = (y_{p+1}, ..., y_T)'$, $\boldsymbol{\varepsilon}_T = (\varepsilon_{p+1}, ..., \varepsilon_T)'$ con distribución $N_{T-p}(\boldsymbol{\varepsilon}_T|\mathbf{0}, \sigma^2 \boldsymbol{I}_{T-p})$, donde $\boldsymbol{I}_n$ denota la matriz identidad de dimensión n, y

$$\boldsymbol{Y}_{T,p} = \begin{pmatrix} y_p & y_{p-1} & \cdots & y_1 \\ y_{p+1} & y_p & \cdots & y_2 \\ \vdots & \vdots & & \vdots \\ y_{T-1} & y_{T-2} & \cdots & y_{T-p} \end{pmatrix},$$

con $\boldsymbol{\phi}$ definido como antes. Las cantidades desconocidas del modelo, suponiendo que el orden del proceso p es fijo, son $(\boldsymbol{\phi}, \sigma^2)$. Las inferencias sobre el modelo (2.25) son en esencia similares al problema de inferencia en el modelo de regresión usual (vea el apéndice B.1).

Análisis Bayesiano

La especificación de la distribución inicial para $(\boldsymbol{\phi}, \sigma^2)$ se puede realizar con base en nuestra información inicial sobre el modelo. Debido a la naturaleza del proceso y a la forma distribucional del proceso de los errores, el análisis se simplifica si asignamos una

distribución inicial conjugada sobre $(\boldsymbol{\phi}, \sigma^2)$ que es de la forma Normal-Gamma Inversa (vea el apéndice B).

Alternativamente podemos asignar una distribución inicial de referencia $\pi(\boldsymbol{\phi}, \sigma^2) \propto 1/\sigma^2$. En este caso la distribución final de $(\boldsymbol{\phi}, \sigma^2)$ es

$$\pi(\boldsymbol{\phi}, \sigma^2|\boldsymbol{y}_T) = N_p(\boldsymbol{\phi}|\boldsymbol{f}, \sigma^2\boldsymbol{B}^{-1})GaI(\sigma^2|a/2, b/2), \tag{2.26}$$

donde $\boldsymbol{B} = (\boldsymbol{Y}'_{T,p}\boldsymbol{Y}_{T,p})$, $\boldsymbol{f} = \boldsymbol{B}^{-1}(\boldsymbol{Y}'_{Y,p}\boldsymbol{y}_T)$, $a = T - p$, y $b = (\boldsymbol{y}_T - \boldsymbol{Y}_{T,p}\boldsymbol{f})'(\boldsymbol{y}_T - \boldsymbol{Y}_{T,p}\boldsymbol{f})$.

En ambos casos la distribución predictiva un paso adelante, para Y_{T+1}, es conocida de manera cerrada. Particularmente en el caso del uso de la distribución inicial de referencia, ésta es

$$p(y_{T+1}|\boldsymbol{y}_T) = St(y_{T+1}|\boldsymbol{y}'_{T-p+1:T}\boldsymbol{f}, \frac{a}{b}(1 + \boldsymbol{y}'_{T-p+1,T}\boldsymbol{B}^{-1}\boldsymbol{y}_{T-p+1:T}), T - p), \tag{2.27}$$

donde $\boldsymbol{y}_{T-p+1:T} = (y_T, ..., y_{T-p+1})'$. Por otro lado la distribución predictiva conjunta para $s \geq 2$ pasos adelante de $\boldsymbol{y}_{T+1:T+s} = (y_{T+1}, ..., y_{T+s})'$ es difícil de encontrar analíticamente, sin embargo podemos generar muestras de ésta mediante un procedimiento secuencial de simulación. También podemos aproximar su densidad predictiva conjunta mediante el método de Monte Carlo, a través de la distribución predictiva Rao-Blackwellizada (Gelfand y Smith, 1990). El análisis descrito anteriormente no requiere de la imposición de estacionariedad en los componentes de autorregresión, dejando abierta la posibilidad de que los datos contribuyan sobre la información de estaciornariedad del proceso.

A continuación presentamos la clase de los modelos autorregresivos y de promedios móviles. Esta clase es particularmente útil, pues puede representar modelos autorregresivos de un orden grande mediante una representación parsimoniosa de orden menor.

2.4.2 Modelos ARMA

Los modelos autorregresivos de promedios móviles (ARMA) es una generalización de los modelos autorregresivos, que consiste en la combinación de esta clase de modelos con la clase de los modelos de promedio móviles (MA), cuya representación caracteriza a un proceso estocástico $\{Z_t\}$ mediante una suma finita ponderada de choques aleatorios

estrictamente independientes, i.e. $Z_t = \mu_t + \varepsilon_t - \psi_1\varepsilon_{t-1} - \ldots - \psi_q\varepsilon_{T-q}$, donde μ_t es el nivel del proceso al tiempo t, $\{\varepsilon_t\}$ es un proceso de ruido blanco, y $\boldsymbol{\psi} = (\psi_1, ..., \psi_q)'$ son los coeficientes de promedios móviles, denotado por MA(q). El proceso de promedios móviles para $\{Y_t\}$, definido para $Y_t = Z_t - \mu_t$, es por definición estacionario, pero existe una condición que relaciona este proceso con un proceso autorregresivo. Esta condición se conoce como la condición de invertibilidad, en cuyo caso un proceso MA(q) puede representarse como un proceso autorregresivo de orden infinito. Un proceso MA(q) puede representarse de manera alternativa como $Y_t = \boldsymbol{\psi}(B)\varepsilon_t$, donde $\boldsymbol{\psi}(B) = 1 - \psi_1 B, \ldots - \psi_q B^q$ denota el polinomio de retraso del proceso. Se dice que este proceso es invertible si las raíces del polinomio de retraso $\boldsymbol{\psi}(B)$, que son las soluciones a la ecuación $1 - \psi_1 B, \ldots - \psi_q B^q = 0$, sean de módulo mayor o igual a uno.

Un modelo ARMA(p,q) es una composición de un modelo AR(p) y un modelo MA(q), representado por

$$\boldsymbol{\phi}(B)Y_t = \boldsymbol{\psi}(B)\varepsilon_t, \tag{2.28}$$

donde $\boldsymbol{\phi}(B)$ y $\boldsymbol{\psi}(B)$ son los polinomios de retraso, de orden p y q respectivamente, definidos como antes, y $\{\varepsilon_t\}$ es un proceso de ruido blanco. Se dice que el proceso ARMA(p,q) es estacionario si el componente AR(p) satisface la condición de estacionariedad, en cuyo caso el proceso puede representarse mediante un proceso MA de orden infinito.

A continuación presentamos brevemente un esquema de inferencia y predicción Bayesiana de un modelo ARMA(p,q). Este procedimiento considera el supuesto de estacionariedad e invertibilidad del proceso. A diferencia de la inferencia del modelo AR(p) anterior, el análisis no puede realizarse de manera analítica cerrada, sin embargo con la parametrización utilizada es posible implementar el muestreador de Gibbs de manera simple.

Sea $\{y_t\}_{t=1}^T$ la trayectoria observada del proceso $\{Y_t\}$ entre los tiempos 1 y T. Supongamos que el proceso está asociado a un modelo ARMA(p,q). Consideremos que los primeros p valores del proceso son fijos y conocidos, y que los primeros q valores del proceso latente $\{\varepsilon_t\}$ son iguales a cero. Supongamos que $\{\varepsilon_t\}$ sigue un proceso de ruido

blanco Gaussiano, con una varianza σ^2 desconocida. Dado el orden del proceso, p y q, y los primeros p valores del proceso conocidos, las cantidades relevantes desconocidas de interés son $(\boldsymbol{\phi}, \boldsymbol{\psi}, \sigma^2)$. En este caso la verosimilitud del modelo puede expresarse como $l(\boldsymbol{\phi}, \boldsymbol{\psi}, \sigma^2|\boldsymbol{y}_T) \propto \sigma^{(T-p)} \exp\{-1/2\sigma^2 \sum_{t=p+1}^{T} \varepsilon_t\}$, con $\varepsilon_t = y_t - \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \psi_j \varepsilon_{t-j}$, para $t = p+1, \ldots, T$.

Alternativamente podemos representar el proceso (2.28) de manera matricial como

$$\boldsymbol{y}_T = \boldsymbol{Y}_{T,p}\boldsymbol{\phi} + \boldsymbol{\Psi}_{T,q}\boldsymbol{\varepsilon}_T, \tag{2.29}$$

con $\boldsymbol{y}_T$, $\boldsymbol{Y}_{T,p}$, $\boldsymbol{\phi}$ y $\boldsymbol{\varepsilon}_T$ definidos como antes, y

$$\boldsymbol{\Psi}_{T,q} = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ \psi_1 & 1 & 0 & 0 & \cdots & 0 & 0 \\ \psi_2 & \psi_1 & 1 & 0 & \cdots & 0 & 0 \\ \psi_3 & \psi_2 & \psi_1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \psi_1 & 1 \end{pmatrix}.$$

Bajo el supuesto que $\{\varepsilon_t\}$ sigue un proceso de ruido blanco Gaussiano con varianza σ^2, la función de verosimilitud para $(\boldsymbol{\phi}, \boldsymbol{\psi}, \sigma^2)$ es de la forma

$$l(\boldsymbol{\phi}, \boldsymbol{\psi}, \sigma^2|\boldsymbol{y}_T) \propto \sigma^{-(T-p)} |\boldsymbol{V}|^{1/2} \exp\left\{-1/2\sigma^2(\boldsymbol{y}_T - \boldsymbol{Y}_{T,p}\boldsymbol{\phi})'\boldsymbol{V}(\boldsymbol{y}_T - \boldsymbol{Y}_{T,p}\boldsymbol{\phi})\right\}, \tag{2.30}$$

donde $\boldsymbol{V} = (\boldsymbol{\Psi}_{T,q}\boldsymbol{\Psi}'_{T,q})^{-1}$. En este caso, los coeficientes de promedios móviles contribuyen a la verosimilitud a través de la matriz $\boldsymbol{V}$, induciendo así una reparametrización natural del modelo. La recuperación de los coeficientes a partir de la matriz $\boldsymbol{V}$ se obtiene mediante la descomposición de Cholesky de la matriz inversa de ésta última. Esta recuperación es necesaria para realizar predicciones sobre valores futuros del proceso.

Análisis Bayesiano

Supongamos que los parámetros relevantes del modelo, que en este caso bajo la reparametrización inducida por (2.30) son $(\boldsymbol{\phi}, \boldsymbol{V}, \sigma^2)$, tienen una distribución inicial no informativa $\pi(\boldsymbol{\phi}, \boldsymbol{V}, \sigma^2) \propto 1/\sigma$. Así, obtenemos que la distribución final para estos parámetros

es

$$\pi(\boldsymbol{\phi}, \boldsymbol{V}, \sigma^2|\boldsymbol{y}_T) \propto \sigma^{-(T-p+1)} |\boldsymbol{V}|^{1/2} \exp\left\{-1/2\sigma^2(\boldsymbol{\phi} - \boldsymbol{f})'\boldsymbol{B}(\boldsymbol{\phi} - \boldsymbol{f}) + \boldsymbol{s}\right\}, \qquad (2.31)$$

donde $\boldsymbol{B} = (\boldsymbol{Y}'_{T,p}\boldsymbol{V}\boldsymbol{Y}_{T,p})$, $\boldsymbol{f} = \boldsymbol{B}^{-1}(\boldsymbol{Y}_{T,p}\boldsymbol{V}\boldsymbol{y}_T)$, y $\boldsymbol{s} = (\boldsymbol{y}_T - \boldsymbol{Y}_{T,p}\boldsymbol{f})'\boldsymbol{V}(\boldsymbol{y}_T - \boldsymbol{Y}_{T,p}\boldsymbol{f})$. A pesar de que no es posible determinar la distribución final conjunta de los parámetros de manera cerrada, la representación (2.31) permite determinar de manera simple las distribuciones condicionales completas para los tres bloques de parámetros $(\boldsymbol{\phi}, \boldsymbol{V}, \sigma^2)$. Así la distribución condicional completa para el conjunto de coeficientes de autorregresión es $\pi(\boldsymbol{\phi}|\boldsymbol{V}, \sigma^2, \boldsymbol{y}_T) = N_p(\boldsymbol{\phi}|\boldsymbol{f}, \sigma^2\boldsymbol{B})$; la distribución final condicional completa para la matriz de covarianzas $\boldsymbol{V}$ es $\pi(\boldsymbol{V}|\boldsymbol{\phi}, \sigma^2, \boldsymbol{y}_T) = Wishart(\boldsymbol{V}|\boldsymbol{\Sigma}, \nu, q)$, con la matriz $\boldsymbol{\Sigma} = \sigma^2\left((\boldsymbol{y}_T - \boldsymbol{Y}_{T,p}\boldsymbol{\phi})(\boldsymbol{y}_T - \boldsymbol{Y}_{T,p}\boldsymbol{\phi})'\right)^{-1}$ y $\nu = q+2$; y la distribución final condicional completa de σ^2 es $\pi(\sigma^2|\boldsymbol{\phi}, \boldsymbol{V}, \boldsymbol{y}_T) = GaI(\sigma^2|a/2, b/2)$ con $a = T - p$ y $b = (\boldsymbol{y}_T - \boldsymbol{Y}_{T,p}\boldsymbol{\phi})'\boldsymbol{V}(\boldsymbol{y}_T - \boldsymbol{Y}_{T,p}\boldsymbol{\phi})$.

La predicción de valores futuros k pasos adelante del proceso se obtiene mediante un proceso de simulación secuencial de la distribución

$$p(\boldsymbol{y}_{T+1:T+k}|\boldsymbol{\phi}, \boldsymbol{\psi}, \sigma^2, \boldsymbol{y}_T) = p(y_{T+1}|\boldsymbol{\phi}, \boldsymbol{\psi}, \sigma^2, \boldsymbol{y}_T) \prod_{t=T+2}^{T+k} p(y_t|\boldsymbol{\phi}, \boldsymbol{\psi}, \sigma^2, \boldsymbol{y}_T, \boldsymbol{y}_{T+1:t-1}),$$

a partir de muestras de la distribución final de $(\boldsymbol{\phi}, \boldsymbol{V}, \sigma^2)$ obtenidas mediante el muestreador de Gibbs.

En la actualidad existe una gran variedad de modelos que básicamente son generalizaciones de los dos modelos básicos que hemos descrito, particularmente en el caso de los modelos AR. Durante el transcurso de este trabajo describiremos algunos de estos modelos junto con sus correspondientes procedimientos de actualización y predicción. El análisis que hemos considerado de estos dos modelos considera que el orden de los componentes AR y MA son fijos. Un problema importante del análisis de series de tiempo con estos modelos consiste en la determinación del orden de los mismos. La determinación del orden de los modelos es un caso particular del problema de selección de modelos. En el siguiente capítulo realizaremos una breve descripción de algunos métodos Bayesiano de selección de modelos, y utilizaremos estos modelos para ejemplificar el funcionamiento de algunos de ellos.

Capítulo 3

Selección de Modelos

3.1 Antecedentes

Al estudiar un problema estadístico, debemos reconocer nuestra ignorancia respecto al modelo que "verdaderamente" representa o genera al fenómeno en estudio, aún cuando se haya realizado un buen análisis exploratorio previo. Es por esto que una de las etapas fundamentales del análisis consiste en la selección y determinación de un modelo estadístico representativo del proceso. Tradicionalmente éste se obtiene de una clase de modelos paramétricos contendientes. El problema de cómo seleccionar un modelo ha estado presente a lo largo de la historia de la ciencia, particularmente en la rama de la estadística, ya que nuestro nivel de información respecto a casi cualquier fenómeno de la naturaleza es incompleto y por tanto las abstracciones que realicemos sobre éste, representadas en este caso a través de los modelos estadísticos, son imperfectas. Por otro lado tenemos el problema de determinación de un modelo estadístico, que consiste en la evaluación de la capacidad de ajuste del modelo o reproducción del fenómeno de interés con base en un conjunto de datos observados. En el apéndice C describimos algunas medidas para evaluar y determinar el comportamiento de un modelo estadístico. Algunas de esas medidas han sido empleadas para seleccionar un modelo, aunque su uso no siempre ha sido adecuado, pues en realidad no representan una solución de un problema de selección de

modelos visto como un problema de decisión. En este sentido, una de las soluciones más representativas que se han planteado consiste en maximizar la función de verosimilitud de los diferentes modelos contendientes, con la posibilidad de incorporar un elemento de penalización sobre la complejidad del mismo, entendida ésta como el número de parámetros involucrados en el modelo. Dos de los criterios más utilizados en este sentido son el criterio de información de Akaike (AIC) (Akaike, 1973) y el criterio Bayesiano de información (BIC) (Schwarz, 1978). Estos dos criterios fueron desarrollados para la comparación de modelos dentro de una misma clase paramétrica y de la misma forma estructural. La posibilidad de incorporar una medida de penalización para la complejidad de los modelos ha sido planteada en diferentes escenarios, sin embargo no es una tarea simple y hasta el momento no existe una regla bien establecida sobre este procedimiento. Bajo el enfoque Bayesiano existen diferentes propuestas que han gozado de cierta aceptación entre sus practicantes; entre éstos se encuentran los factores de Bayes, los cuales describiremos brevemente en las siguientes secciones.

La selección de un sólo modelo puede producir sesgos en el análisis. Una alternativa para eliminar este efecto consiste en realizar una combinación de diferentes modelos dentro de la clase de modelos contendientes o en una clase representativa menor, con el objetivo de flexibilizar los supuestos del análisis. Esta alternativa es particularmente atractiva para resolver un problema de predicción, donde se considera que cada modelo considera ciertos supuestos particulares. Con el enfoque Bayesiano de inferencia y predicción la combinación de modelos se obtiene en forma de mezclas de distribuciones finales o distribuciones predictivas, según sea el caso. Para el primer caso, debemos tomar en cuenta que la característica de interés necesariamente debe tener el mismo significado en cada modelo considerado en la mezcla. El problema de mezcla de modelos es conocido recientemente como promedio Bayesiano de modelos (PBM), y es considerada como una solución alternativa al problema de selección de un modelo en un ambiente de incertidumbre. El reciente auge de este método se debe en gran parte a los avances de métodos computacionales que hacen posible el cálculo de las mezclas, o de características de éstas que puedan ser expresadas en forma de una esperanza respecto a la mezcla final.

Nuestro interés central es la selección de modelos con énfasis en el análisis de series de tiempo, un problema que ha generado gran interés a lo largo de la historia. Como ya mencionamos, cada modelo propuesto está sustentado por un contexto o marco teórico que involucra una serie de supuestos particulares respecto al entorno de la serie de interés o respecto al proceso interno que genera a la misma, i.e. cada modelo representa un escenario distinto que el analista está dispuesto a considerar.

Por otro lado, la combinación de modelos se basó inicialmente en encontrar una forma de combinar pronósticos puntuales, generados por diferentes modelos, a través de combinaciones lineales convexas de dichos estimadores puntuales de predicción. Básicamente estos métodos consistieron en determinar formas óptimas para combinar la información proporcionada por diferentes "sistemas expertos". Una primera aproximación se debe a Bates y Granger (1969), quienes propusieron combinar los modelos mediante su promedio ponderado, sujeto a que los componentes de la combinación minimicen la varianza inducida por la transformación. Esta propuesta consiste en suponer que conocemos estimadores puntuales de predicción, preferentemente insesgados, y estimaciones relativas a su precisión o variabilidad para cada uno de los modelos de la combinación. Bates y Granger (1969) supusieron que las predicciones generadas por los modelos son independientes entre sí. A partir de esta propuesta surgieron diferentes modificaciones y generalizaciones para combinar pronósticos puntuales, tanto con el enfoque frecuentista como Bayesiano. Este último tuvo una gran aceptación debido a que permite incorporar elementos subjetivos dentro de la mezcla, lo que para la fecha resultaba atractivo para los practicantes de este método (e.g. Bunn (1975) y Bordley (1982)). Anandaligan y Chen (1989) propusieron reportar la mezcla de modelos a través de una pseudo densidad predictiva, donde la densidad predictiva final de la variable de interés es obtenida a través de condicionar en los valores predictivos puntuales obtenidos de los diferentes modelos, vistos como una muestra jerárquica del modelo final. Sin embargo esta propuesta no es estrictamente Bayesiana, por lo que recibió un gran número de críticas. En Clemen (1989) podemos encontrar una revisión histórica a la fecha de los diferentes métodos de combinación de pronósticos que brevemente hemos descrito, tanto desde un punto de vista frecuentista como Bayesiano.

A pesar de los intentos de generar un método estándar de combinación de modelos que sea eficiente bajo diferentes criterios de optimalidad, ninguna de las propuestas consideradas había resumido la información relevante del problema de predicción de la manera como Jeffreys (1961) lo conceptualizó por medio de las mezclas de las distribuciones predictivas finales de los modelos. La propuesta de Jeffreys (1961) eventualmente se convirtió en la base de lo que hoy conocemos como el promedio Bayesiano de modelos, donde la distribución predictiva final de la variable futura se calcula como una mezcla de densidades predictivas finales, donde los pesos de la mezcla están determinados por las probabilidades finales de cada modelo. Más adelante discutiremos que esta forma de mezclar modelos es óptima bajo ciertos supuestos específicos sobre la clase de modelos portulada y la definición del objetivo final del problema de selección. Este procedimiento de combinación de modelos ha sido empleado recientemente para mezclar modelos con la misma forma estructural (e.g. Madigan y Raftery (1994), Madigan y York (1995) y Clyde (1999) entre otros), aunque también existen propuestas para mezclar modelos con formas estructurales distintas (e.g. Min y Zellner (1993), Draper (1995) y Walker *et al.* (2001)).

La forma de plantear un problema de selección de modelos depende de la perspectiva que asumamos respecto a la clase de los modelos contendientes. En la siguiente sección presentamos una breve descripción de algunas de estas perspectivas.

3.2 Perspectivas

En esta sección describimos algunas perspectivas respecto a la clase de modelos contendientes. Denotemos por $\mathcal{M} = \{M_k : k \in K\}$ a la clase de modelos paramétricos por comparar, donde K es un conjunto índice (finito o numerable), de la que deseamos seleccionar sólo uno. Cada modelo M_k es sólo una representación probabilística de cantidades observables, denotadas por Y, y cantidades no observables (parámetros), denotadas por $\boldsymbol{\theta}_k \in \boldsymbol{\Theta}_k \subset \Re^{n(k)}$, donde $\boldsymbol{\Theta}_k$ es el espacio parametral asociado al modelo M_k de dimensión finita y $n(k) \geq 1$ es su dimensión correspondiente. Bajo el enfoque Bayesiano, cada

modelo M_k está definido como

$$M_k = \{p_k(y|\boldsymbol{\theta}_k), \pi_k(\boldsymbol{\theta}_k)\}, \tag{3.1}$$

donde $p_k(y|\boldsymbol{\theta}_k)$ denota la densidad (o distribución) condicional de la variable aleatoria Y dado el parámetro $\boldsymbol{\theta}_k$ y el modelo M_k, y $\pi_k(\boldsymbol{\theta}_k)$ es la densidad (o distribución) inicial de los parámetros $\boldsymbol{\theta}_k$ condicional en el modelo M_k, para cada $k \in K$.

Para que el análisis sea lo más completo posible, es recomendable que la clase $\mathcal{M}$ contenga una colección amplia de modelos, deseablemente con diferentes formas estructurales. La incorporación de la distribución inicial $\pi_k(\boldsymbol{\theta}_k)$ en la definición (3.1) es un factor importante de determinación del modelo, y en sí mismo caracteriza de manera independiente a M_k, pero su especificación no debe ser considerada como el objetivo final del análisis. La comparación de modelos estadísticos caracterizados por diferentes distribuciones iniciales y una misma verosimilitud corresponde a una área de estudio independiente, que estudia precisamente la robustez del modelo ante variaciones en las asignaciones de las distribuciones iniciales sobre los parámetros. Esta área se concentra fundamentalmente en estudiar la sensibilidad del modelo, y de las densidades finales de interés, ante alteraciones de la distribución inicial sobre los parámetros.

En general existen al menos tres diferentes perspectivas o enfoques respecto a la clase $\mathcal{M}$ que podemos asumir al momento de comparar diferentes modelos estadísticos (Bernardo y Smith, 1994, págs. 383-385). Una de éstas consiste es suponer que existe un modelo dentro de la clase $\mathcal{M}$ que es el modelo "verdadero", en el sentido que representa al proceso que genera los datos observados. En la discusión inicial de este capítulo describimos que nuestro desconocimiento sobre la naturaleza del fenómeno de estudio nos limita a tener que definir modelos aproximados a la realidad. Suponer que algún modelo es verdadero puede conducirnos a errores y sesgos importantes en el análisis. Esta perspectiva es conocida como $\mathcal{M}$-cerrada. En este caso la elección de los modelos se reduce a buscar y elegir dentro de la clase $\mathcal{M}$ al modelo "verdadero". Esta perspectiva es empleada implícitamente en algunos criterios de selección de modelos.

Una perspectiva más honesta para la mayoría de los problemas estadísticos prácticos consiste en suponer que ninguno de los modelos de la clase es el modelo verdadero, i.e.

cada modelo es solamente una aproximación e interpretación del proceso subyacente al fenómeno que nos interesa, sobre el cual desconocemos la forma estructural. En este caso, la búsqueda y selección de un modelo consiste en la comparación y selección de uno de éstos con base en algún criterio de optimalidad o discrepancia respecto al modelo "verdadero" dentro de la clase $\mathcal{M}$. Esta perspectiva es conocida como $\mathcal{M}$-abierta.

En algunos casos especiales, como en simulación y reconocimiento e identificación de señales, sabemos que existe un modelo verdadero y en ocasiones éste puede ser accesible para nosotros pero por diferentes circunstancias, como costo de resolución o la necesidad obtener inferencias en un lapso de tiempo corto, necesitamos aproximarlo mediante una representación parsimoniosa. En este caso la comparación de modelos consiste en elegir el modelo más simple posible y que sea una buena aproximación del modelo verdadero. Este enfoque es conocido como el enfoque $\mathcal{M}$-completo.

Nosotros consideramos que el problema de selección de modelos es esencialmente un problema de decisiones en un ambiente de incertidumbre respecto a diferentes características de interés. Diferentes criterios de selección surgen como resultado de definir diferentes espacios de decisión, y sobre todo de definir diferentes espacios de 'estados de la naturaleza', sobre el cual nuestro conocimiento es limitado o inexistente. La definición de estos espacios dependen fundamentalmente del objetivo final de nuestro análisis y, sobre todo, de la perspectiva que estemos dispuestos a asumir sobre la clase $\mathcal{M}$. En la siguiente sección describiremos algunas de las soluciones Bayesianas de este problema que han recibido una gran aceptación en la práctica .

3.3 Selección de Modelos como un Problema de Decisión

Denotemos por D_t a la información disponible al momento de seleccionar un modelo, donde t es el tiempo de la toma de decisión o selección. La información D_t incluye los datos observados del proceso de interés hasta el tiempo t y la información adicional relevante para el problema. Tradicionalmente en esquemas cerrados, i.e. aquellos donde la

inferencias y predicción depende en cada tiempo exclusivamente de los datos observados sin incorporar elementos adicionales de intervención, si Y es la variable aleatoria de interés y el tiempo de la toma de decisión es t, la información relevante estará compuesta por los datos observados hasta este tiempo, i.e. $D_t = \{y_1, ..., y_t\}$. En esta sección presentaremos una breve descripción de los elementos del problema de decisión en ciertos criterios de selección.

3.3.1 Criterio Bayesiano de Máxima Probabilidad

El problema más simple de selección de modelos consiste en elegir sólo un modelo de la clase $\mathcal{M}$ sin considerar ningún objetivo final específico mas que el de selección. En este caso los elementos del problema de decisión se definen considerando como el espacio 'estados de la naturaleza' a la clase $\mathcal{M}$, y el espacio de acciones como la misma clase $\mathcal{M}$. Nuestra incertidumbre inicial sobre el verdadero modelo de la clase es resumida mediante la asignación de una probabilidad $p(M_k) = p(k)$, que se interpreta como la probabilidad de que el modelo M_k sea el "verdadero" modelo. Considerando la información disponible al tiempo t, nuestro conocimiento acerca del verdadero modelo es actualizado mediante el Teorema de Bayes como

$$p(k|D_t) \propto p_k(\boldsymbol{y}_t)p(k), \tag{3.2}$$

donde $\boldsymbol{y}_t = (y_1, ..., y_t)'$ son los datos observados hasta el tiempo t, y $p_k(\boldsymbol{y}_t)$ es la *verosimilitud integrada* del modelo M_k, que a su vez es calculad como:

$$p_k(\boldsymbol{y}_t) = \int p_k(\boldsymbol{y}_t|\boldsymbol{\theta}_k)\pi_k(\boldsymbol{\theta}_k)d\boldsymbol{\theta}_k, \tag{3.3}$$

Bajo la perspectiva $\mathcal{M}$-cerrada, es natural definir una función de utilidad 0-1, i.e.

$$u(k, \omega) = \begin{cases} 1, & \text{si } \omega = k, \\ 0, & \text{si } \omega \neq k. \end{cases} \tag{3.4}$$

La solución Bayesiana óptima consiste en elegir el modelo M_k que maximice la función de utilidad esperada final, condicional en la información D_t, que es

$$\begin{aligned} \overline{u}(k) &= \sum_{j \in K} u(k, j)p(j|D_t) \\ &= p(k|D_t), \end{aligned} \tag{3.5}$$

donde $p(j|D_t)$ denota nuestro conocimiento actualizado respecto al modelo M_j. De esta forma el modelo óptimo es aquel que tenga la mayor probabilidad final. Este criterio se basa fundamentalmente en la consistencia del procedimiento Bayesiano y de la probabilidad final del supuesto modelo verdadero, denotado por M_T. Conforme nuestro estado de información se incremente la probabilidad $p(T|D_t)$ tenderá asintóticamente a uno, siempre que el modelo M_T se encuentre dentro de la clase de modelos contendientes.

3.3.2 Momios Finales y Factores de Bayes

En algunas ocasiones es deseable comparar dos modelos entre sí. Denotemos por M_k y M_h a los dos modelos contendientes de la clase $\mathcal{M}$. La cuantificación natural de la comparación entre estos modelos se calcula como el cociente de sus correspondientes probabilidades finales de la forma

$$\frac{p(k|D_t)}{p(h|D_t)} = \frac{p_k(\boldsymbol{y}_t)}{p_h(\boldsymbol{y}_t)} \times \frac{p(k)}{p(h)}. \tag{3.6}$$

Este cociente surge como la solución Bayesiana óptima del contraste de hipótesis entre los modelos M_k y M_h, usando la función de utilidad 0-1 (3.4). La relación (3.6) muestra que el cociente de probabilidades finales es resultado de la actualización del cociente de probabilidades iniciales a través del cociente de sus correspondientes verosimilitudes integradas. Este último cociente es conocido como el *Factor de Bayes* (FB) del modelo M_k respecto al modelo M_h, que se define como

$$FB_{kh} = \frac{p_k(\boldsymbol{y}_t)}{p_h(\boldsymbol{y}_t)}, \tag{3.7}$$

donde

$$p_k(\boldsymbol{y}_t) = \int p_k(\boldsymbol{y}_t|\boldsymbol{\theta}_k)\pi_k(\boldsymbol{\theta}_k)d\boldsymbol{\theta}_k. \tag{3.8}$$

Los usuarios de este método interpretan al factor de Bayes como una medida de soporte de los datos en favor del modelo M_k respecto al modelo M_h. Si consideramos una distribución inicial propia no informativa en la clase $\mathcal{M}$, que representa nuestro estado de ignorancia sobre el verdadero modelo, entonces el factor de Bayes (3.7) es la solución óptima del contraste Bayesiano de los modelos M_k y M_h, usando la función de utilidad 0-1.

Muchos problemas prácticos requieren de modelos paramétricos complicados para su modelación, en cuyo caso no se dispone de información inicial sobre los parámetros $\boldsymbol{\theta}_k$ asociados al modelo, de manera que no es fácil asignarles una distribución inicial que represente nuestro estado de información inicial. Una práctica común en este caso consiste en asignar distribuciones iniciales difusas o no informativas. Algunos de los métodos para asignar distribuciones iniciales no informativas pueden producir distribuciones impropias. En este caso la integral $\int p_k(\boldsymbol{y}_t|\boldsymbol{\theta}_k)\pi_k(\boldsymbol{\theta}_k)d\boldsymbol{\theta}_k$ puede no estar determinada, y por consiguiente el factor de Bayes (3.7) resultaría indeterminado, debido a las constantes multiplicativas arbitrarias que se pueden relacionar a la distribución inicial $\pi_k(\boldsymbol{\theta}_k)$.

Una solución para corregir parcialmente el problema de indeterminación consiste en obtener una submuestra de los datos observados con la que que se pueda tener una distribución final propia, que a su vez sería considerada como inicial al incorporar los datos restantes. Siguiendo la idea del aprendizaje secuencial del paradigma Bayesiano, ésta puede usarse como una nueva distribución inicial, y entonces podemos proceder a comparar los modelos con la muestra restante. Así, podemos calcular el factor de Bayes con la muestra restante. El factor de Bayes calculado con este procedimiento es conocido como *factor de Bayes parcial* (FBP), cuyo nombre se debe a que hace referencia sólo a un subconjunto de los datos observados, y se define como

$$FBP_{hk} = \frac{p_h(\boldsymbol{y}_{(-m)}|\boldsymbol{y}_{(m)})}{p_k(\boldsymbol{y}_{(-m)}|\boldsymbol{y}_{(m)})}, \tag{3.9}$$

donde m es el tamaño de la muestra de entrenamiento, $\boldsymbol{y}_{(m)} = (y_1, ..., y_m)$ es la muestra de entrenamiento y $\boldsymbol{y}_{(-m)} = (y_{m+1}, ..., y_n)$ es la muestra restante de tamaño $n - m$. Así la nueva verosimilitud condicional integrada se calcula como

$$p_k(\boldsymbol{y}_{(-m)}|\boldsymbol{y}_{(m)}) = \int p_k(\boldsymbol{y}_{(-m)}|\boldsymbol{\theta}_k, \boldsymbol{y}_{(m)})\pi_k(\boldsymbol{\theta}_k|\boldsymbol{y}_{(m)})d\boldsymbol{\theta}_k, \tag{3.10}$$

donde $\pi_k(\boldsymbol{\theta}_k|\boldsymbol{y}_{(m)})$ denota la densidad final de los parámetros, que es propia, condicional en la muestra de entrenamiento. Este procedimiento, además de hacer un doble uso de los datos, conlleva el problema de cómo determinar el tamaño y la muestra inicial de entrenamiento $\boldsymbol{y}_{(m)}$.

Siguiendo este procedimiento podemos definir diferentes tamaños para la muestra de entrenamiento. Berger y Pericchi (1996) propusieron determinar un tamaño de muestra mínimo que garantice que las distribuciones finales de todos los modelos contendientes sean propias y posteriormente proceder a calcular los factores de Bayes parciales obtenidos a través de todas las posibles muestra de entrenamiento de este tamaño mínimo. Berger y Pericchi propusieron comparar dos modelos a través del promedio, aritmético o geométrico, de estos factores de Bayes parciales. Los factores de Bayes producidos con este procedimiento son conocidos como factores de Bayes intrínsecos.

O'Hagan (1995, sección 2.3) propuso un procedimiento alternativo para resolver el problema de la indeterminación del factor de Bayes. Ésta consiste incorporar una densidad fraccionaria en el cálculo de la verosimilitud integrada del modelo M condicional en la muestra de entrenamiento, y así definir una verosimilitud integrada fraccionaria de la forma $p_k^{(\alpha)}(\boldsymbol{y}_{(-m)}|\boldsymbol{y}_{(m)}) = \int p_k^{\alpha}(\boldsymbol{y}_{(-m)}|\boldsymbol{\theta}_k, \boldsymbol{y}_{(m)})\pi(\boldsymbol{\theta}_k|\boldsymbol{y}_{(m)}, k)d\boldsymbol{\theta}_k$, donde $\alpha = \frac{m}{n}$. Así, la comparación entre dos modelos se efectúa mediante el cociente de sus correspondientes verosimilitudes integradas fraccionarias, y es conocido como factor de Bayes fraccionario.

En el contexto de series de tiempo se tiene una estructura de dependencia definida entre las variables aleatorias. Sabemos que, para un modelo M, podemos efectuar la siguiente descomposición de la densidad conjunta,

$$p(y_1, ..., y_T|M) = \prod_{t=1}^{T} p(y_t|\boldsymbol{y}_{t-1}, M), \tag{3.11}$$

donde $p(y_t|\boldsymbol{y}_{t-1}, M)$ denota la densidad predictiva del proceso un tiempo adelante en el tiempo $t-1$ y $\boldsymbol{y}_{t-1} = (y_1, ..., y_{t-1})'$ denota la trayectoria observada del proceso hasta el tiempo $t-1$. Geweke (1995) propuso sustituir la densidad anterior en la definición del factor de Bayes (3.7), de la que se produce una evidente descomposición en el producto de los factores de Bayes para cada tiempo t, conocido como *factor de Bayes predictivo* ($FBpred$), de la forma

$$FBpred_{hk} = \prod_{t=1}^{T} FBpred_{hk}(t), \tag{3.12}$$

donde

$$FBpred_{hk}(t) = \frac{p_h(y_t|\boldsymbol{y}_{t-1})}{p_k(y_t|\boldsymbol{y}_{t-1})}, \tag{3.13}$$

con

$$p_M(y_t|\boldsymbol{y}_{t-1}) = \int p_M(y_t|\boldsymbol{\theta}, \boldsymbol{y}_{t-1})\pi(\boldsymbol{\theta}|\boldsymbol{y}_{t-1})d\boldsymbol{\theta}, \tag{3.14}$$

para $t = 1, ..., T$ y cada $M \in \mathcal{M}$. Esta descomposición ofrece un panorama secuencial de actualización y comparación de modelos que empata con la naturaleza secuencial de la forma como se observan los datos. Si existe algún problema de indeterminación en los factores de Bayes predictivos al inicio de la serie, podemos definir una segmento inicial de la misma como una muestra de entrenamiento, y proceder a calcular los factores de Bayes predictivos parciales. De manera análoga se puede obtener la descomposición del cociente de probabilidades finales (3.6).

En nuestra opinión, los factores de Bayes permiten comparar modelos sin la necesidad de imponer una estructura de anidamiento entre ellos, entendiendo por anidamiento a aquellos modelos que tienen la misma forma estructural con la diferencia de incorporar una cantidad diferente de parámetros y donde algunos de éstos puede verse como un caso particular de uno más general de la misma clase, pero si se desea comparar más de dos modelos se debe realizar una comparación dos a dos de todos las posibles combinaciones de ellos. Se han desarrollado diferentes métodos numéricos para estimar factor de Bayes principalmente para modelos con la misma forma estructural, en este caso podemos utilizar estas aproximaciones para estimar las probabilidades finales de cada modelo en los casos en que éstas no puedan calcularse o estimarse directamente de manera eficiente, de acuerdo al resultado analítico propuesto por Berger y Pericchi (1996). Por otro lado, mediante el uso de los factores de Bayes podemos instrumentar un procedimiento *ad-hoc* de selección de modelos mediante la comparación dos a dos entre todos los modelos de la clase $\mathcal{M}$, y elegir el modelo que tenga la "mayor evidencia" empírica a su favor, aunque en sí mismo este procedimiento no sea un método general de selección. Por otro lado, en su uso se supone implícitamente la cuestionable perspectiva $\mathcal{M}$-cerrada, aunque algunos usuarios de este método no lo vean así, interpretándolos solamente como una medida de soporte de los datos en favor de un modelo respecto a otro. Solamente el factor de Bayes geométrico puede interpretarse como una "aproximación" a la solución del problema de comparación de modelos en la perspectiva $\mathcal{M}$- abierta (vea por ejemplo Key *et al.* (1999)

y la discusión de A. F. M. Smith en O'Hagan (1995)). Un problema adicional respecto al uso de los factores de Bayes es que generalmente la verosimilitud integrada (3.8) es difícil de calcular de manera cerrada, aunque puede ser aproximada por el método de Laplace en su forma exponencial (Gelfand y Dey, 1994) o mediante otros métodos numéricos.

Uno de los problemas centrales del análisis estadístico es el de predicción. Este debe considerarse de manera explicita en la selección de modelos. En la siguiente sección describiremos la solución Bayesiana de selección para el caso particular donde el objetivo central del análisis es de predicción.

3.4 Selección Predictiva de Modelos

En muchas ocasiones, el objetivo final del análisis estadístico es el de producir predicciones de valores futuros de una variable aleatoria de interés Y (observable), como en el análisis de series de tiempo. Con el enfoque Bayesiano, estas predicciones se pueden reportar de diferentes formas, por ejemplo en la forma de predicciones puntuales, cuantiles o regiones, o en la forma más general a través de la distribución predictiva final. De acuerdo con Box (1980), las únicas características comparables entre diferentes modelos estadísticos son las distribuciones predictivas finales generadas por cada modelo. De esta forma, el criterio de comparación y selección del mejor modelo, tanto en la perspectiva cerrada o abierta sobre $\mathcal{M}$, debe basarse en la capacidad de predicción de cada uno de éstos, y en la identificación del modelo que genere las mejores predicciones, de acuerdo a algún criterio de optimalidad. A continuación describimos algunos de los componentes del problema de selección cuando el objetivo final del análisis es el de predicción.

3.4.1 Espacio de 'Estados de la Naturaleza' y Espacio de Acciones

Denotemos por $Y_f \in \mathcal{Y}_f \subset \Re$ a la variable aleatoria futura sobre la que deseamos inferir. En el contexto de toma de decisiones $\mathcal{Y}_f$ es el espacio de 'estados de la naturaleza'. En este caso tenemos un espacio de acciones subsecuentes, i.e. en la primera etapa de las

acciones debemos elegir un modelo dentro de la clase $\mathcal{M}$, i.e. $\mathcal{A}_1 = \mathcal{M}$; y en la segunda etapa debemos inferir respecto a la variable aleatoria Y_f. El problema puede simplificarse si definimos como espacio de acciones a la clase

$$\mathcal{P} = \{p_k(\cdot|D_t) : k \in K\}, \tag{3.15}$$

formada por todas las funciones predictivas finales generadas por los modelos de la clase $\mathcal{M}$, las cuales están definidas sobre el espacio $\mathcal{Y}_f \subset \Re$. La solución Bayesiana óptima a este problema de decisión consiste en elegir el modelo de la clase $\mathcal{M}$ que maximice la utilidad esperada final $\overline{u}(p_k(\cdot|D_t))$, que se obtiene como

$$\overline{u}(p_k(\cdot|D_t)) = \int u(p_k(\cdot|D_t), \omega)p(\omega|D_t)d\omega, \tag{3.16}$$

donde $p(\omega|D_t)$ denota la "verdadera" densidad predictiva final de Y_f, que en la mayoría de los casos reales es desconocida para nosotros.

3.4.2 Perspectiva $\mathcal{M}$-cerrada

En la perspectiva Bayesiana $\mathcal{M}$-cerrada suponemos que uno de los modelos en $\mathcal{M}$ es el modelo "verdadero", y nuestro desconocimiento respecto a cuál de estos modelos es el verdadero se resume en las correspondientes probabilidades finales de cada modelo en $\mathcal{M}$. Así, toda la información disponible respecto a la densidad predictiva final "verdadera" $p(\omega|D_t)$ está resumida en la forma de una mezcla de densidades predictivas finales de los modelos en la clase $\mathcal{P}$ ponderadas por la correspondiente probabilidad final de cada modelo, de hecho esta mezcla es el estimador Bayesiano bajo esta perspectiva, i.e.

$$p(\omega|D_t) = \sum_{k \in K} p_k(\omega|D_t)p(k|D_t). \tag{3.17}$$

La forma explícita de (3.17) involucra la solución de una integral complicada

$$p(\omega|D_t) = \sum_{k \in K} p(k|D_t) \int p_k(\omega|D_t, \boldsymbol{\theta}_k)\pi(\boldsymbol{\theta}_k|k, D_t)d\boldsymbol{\theta}_k, \tag{3.18}$$

que generalmente es difícil o imposible de resolver analíticamente. En estos caso no se puede calcular de manera cerrada la densidad (3.17), y por consiguiente tampoco podemos

calcular la utilidad esperada final (3.16) de manera cerrada. Para calcular esta integral se recurre generalmente a diferentes métodos numéricos de integración. en la sección 2.3 describimos brevemente algunos de estos métodos.

3.4.3 Perspectiva $\mathcal{M}$-abierta

Una postura realista consiste en aceptar nuestro desconocimiento respecto a la función predictiva $p(\omega|D_t)$. Como Gutiérrez-Peña (1997) y Gutiérrez-Peña y Walker (2001) apuntaron, este es un elemento de incertidumbre importante en el análisis, y de hecho es el elemento de incertidumbre fundamental en el criterio predictivo de selección. El espacio de 'estados de la naturaleza' corresponde a todas las posibles distribuciones (o densidades) predictivas

$$\mathcal{P}' = \{p(\cdot|D_t) : p(\cdot|D_t) \text{ esté definida en el espacio } \mathcal{Y}_f\}. \tag{3.19}$$

La cuantificación de nuestra incertidumbre sobre los elementos del espacio $\mathcal{P}'$ es una medida de probabilidad definida sobre un espacio de funciones de densidad. Para poder calcular la utilidad esperada (3.16) necesitamos estimar la densidad (o distribución) $p(\omega|D_t)$. Bernardo y Smith (1994) estimaron la distribución verdadera mediante la función de distribución empírica. Para el caso más general de variables aleatorias intercambiables, (Gutiérrez-Peña y Walker, 2001) modelaron la incertidumbre sobre la distribución verdadera a través de un proceso Dirichlet, que define una medida de probabilidad sobre el conjunto de todas las distribuciones definidas sobre un espacio particular. La medida de probabilidad final sobre la verdadera distribución es también un proceso Dirichlet. Los procesos Dirichlet pueden definirse de manera que nos aproximemos tanto como queramos a la distribución empírica. Para una introducción a los procesos de Dirichlet vea Ferguson (1974) o Schervish (1995, sección 1.6).

La solución del problema de decisión bajo la perspectiva Bayesiana $\mathcal{M}$-abierta depende de la forma como modelemos nuestra incertidumbre sobre la distribución que denota la verdadera distribución del proceso. Generalmente esto se debe hacer usando modelos Bayesianos no paramétricos o semiparamétricos, entendidos como aquellos modelos más flexibles donde los supuestos estructurales del modelo son más relajados a los de su contra-

parte paramétrica, y donde el espacio parametral es de dimensión infinita o excesivamente grande. Pospondremos por el momento la discusión sobre estos modelos, para retomarlos en el siguiente capítulo en el contexto del análisis de series de tiempo.

3.4.4 Funciones de Utilidad Compatibles

En las subsecciones anteriores revisamos algunos criterios para seleccionar un modelo con base en un enfoque predictivo. El elemento faltante para definir el problema de decisión consiste en determinar una función de utilidad apropiada para el espacio de consecuencias, que refleje las preferencias de un modelo a otro. Como revisamos en la subsección 2.2.2, la función de utilidad en este caso debe ser una función de puntaje, de preferencia local y propia. Muchos autores proponen utilizar la *función de puntaje logarítmico* (vea Bernardo (1979) y Bernardo y Smith (1994)), en cuyo caso la utilidad esperada de cada modelo es de la forma

$$\overline{u}(p_k(\cdot|D_t)) = A\mathbb{E}_{Y_f|D_t}\left[\log p_k(y_f|D_t)\right] + \mathbb{E}_{Y_f|D_t}\left[B(y_f)\right], \tag{3.20}$$

donde $A > 0$ es un número real positivo, y $B(\cdot)$ es una función real definida sobre $\mathcal{Y}_f$ que es medible respecto a la distribución $p(\cdot|D_t)$.

El modelo óptimo es aquel que maximice (3.20) o equivalentemente, eliminando los términos comunes a todos los modelos, el modelo que maximice

$$\int \log\left\{p_k(y_f|D_t)\right\} p(y_f|D_t)dy_f, \tag{3.21}$$

donde $p(y_f|D_t)$ denota la "verdadera" densidad predictiva de la variable Y_f dada la información D_t, que es equivalente a minimizar la entropía de $p_k(y_f|D_t)$ respecto a $p(y_t|D_t)$. Las esperanzas en la relación (3.20) o (3.21) se calculan respecto al PBM bajo la perspectiva $\mathcal{M}$-cerrada, o mediante una aproximación no paramétrica o semiparamétrica bajo la perspectiva abierta. La función de puntaje logarítmico ha sido utilizada por diferentes autores (Bernardo, 1979; San Martini y Spezzaferri, 1984; Bernardo y Smith, 1994; Laud e Ibrahim, 1995; Gutiérrez-Peña, 1997; Key *et al.*, 1999; Gutiérrez-Peña y Walker, 2001) en el contexto de selección de modelos.

Otra función de utilidad apropiada puede ser la *función de puntaje cuadrática* (Bernardo y Smith, 1994, sección 3.4.2.), donde se tiene que el modelo óptimo, eliminando los términos irrelevantes (ver el apéndice A), es aquel que maximiza

$$-\int \{p(y_f|D_t) - p_k(y_f|D_t)\}^2 \, dy_f. \tag{3.22}$$

Este último criterio corresponde al caso de elegir el modelo que minimice la discrepancia cuadrática entre su correspondiente densidad predictiva final respecto a la densidad predictiva "verdadera".

Por otro lado, maximizar (3.21) es equivalente a minimizar la divergencia de Kullback-Leibler de la densidad $p_k(y_f|D_t)$ respecto a la verdadera distribución $p(y_f|D_t)$. Como Bernardo y Smith (1994, sección 6.1.5) notaron, para reportar inferencias a través de la densidad predictiva, podemos definir un criterio más general eligiendo el modelo que minimice una medida de discrepancia entre su correspondiente densidad predictiva final respecto a la mezcla PBM (3.17), en el caso $\mathcal{M}$-cerrada, o a una densidad predictiva semiparámetrica o paramétrica que estime a la verdadera distribución, en el caso $\mathcal{M}$-abierta. En el apéndice A damos una breve introducción a la medida de discrepancia de Kullback-Leibler, con algunas de sus características más importantes.

3.5 Selección de Modelos para Series de Tiempo

Uno de los principales objetivos al analizar series de tiempo es el de generar u obtener información sobre valores futuros de la serie. Para poder reportar resultados de manera sustentada es necesario que el modelo propuesto tenga una buena capacidad predictiva. Aún cuando el objetivo final del análisis no sea el de predicción, por ejemplo en el caso de suavizamiento o descomposición de series de tiempo, es deseable y necesario que el modelo propuesto posea también una capacidad predictiva alta pues da validez a su utilización, ya que presumiblemente captura algunos elementos importante del "verdadero" modelo que rige su evolución. Generalmente se proponen diferentes modelos para tales fines, algunos pertenecientes a la misma familia parametral y otros con características y formas estructurales distintas. En este caso es necesario definir un procedimiento de

comparación y selección de modelos con base en su capacidad predictiva. En esta sección extenderemos el criterio predictivo de selección de modelos descrito en la sección anterior, en las perspectivas $\mathcal{M}$-cerrada y $\mathcal{M}$-abierta, y discutiremos algunos detalles generales sobre su implementación.

Consideremos una serie de tiempo en tiempo discreto $\{Y_t : t = 1, 2, ...\}$ y sea $\{y_t\}_{t=1}^{T}$ la trayectoria observada de este proceso entre el tiempo 1 y el tiempo T. Supongamos que después de realizar un análisis de inspección previo somos capaces de definir una clase de modelos paramétricos contendientes, $\mathcal{M} = \{M_k : k \in K\}$. Supongamos que deseamos seleccionar sólo un modelo con base en su capacidad predictiva respecto a la trayectoria futura del proceso en un horizonte de longitud $s \geq 1$.

3.5.1 Criterio Predictivo $\mathcal{M}$-cerrada

El espacio de acciones es la clase $\mathcal{M}$ de modelos contendientes y el espacio de 'estados de la naturaleza' es el conjunto de las posibles trayectorias futuras del proceso en el espacio $\mathcal{Y}_{T+1:T+s} \subset \Re^s$. Trabajando con la función de puntaje logarítmico, el modelo óptimo es aquel que maximiza

$$\int \log \left\{ p_k(\boldsymbol{y}_{T+1:T+s}|\boldsymbol{y}_T) \right\} p(\boldsymbol{y}_{T+1:T+s}|\boldsymbol{y}_T) d\boldsymbol{y}_{T+1:T+s}, \tag{3.23}$$

donde $\boldsymbol{y}_T = (y_1, ..., y_T)'$, $\boldsymbol{y}_{T+1:T+s} = (y_{T+1}, ..., y_{T+s})'$, $p_k(\boldsymbol{y}_{T+1:T+s}|\boldsymbol{y}_T)$ es la densidad predictiva de la trayectoria $\boldsymbol{Y}_{T+1:T+s}$ condicional en la trayectoria observada $\boldsymbol{y}_T$ en el modelo M_k, y $p(\boldsymbol{y}_{T+1:T+s}|\boldsymbol{y}_T)$ es la "verdadera" densidad predictiva de esa trayectoria. Como ya mencionamos, podemos considerar al menos dos perspectivas distintas respecto a la clase $\mathcal{M}$.

Considerando la perspectiva $\mathcal{M}$-cerrada, la densidad predictiva $p(\boldsymbol{y}_{T+1:T+s}|\boldsymbol{y}_T)$ puede estimarse como la mezcla de las diferentes densidades predictivas de los modelos de la clase $\mathcal{M}$ ponderada por sus correspondientes probabilidades finales, i.e.

$$p(\boldsymbol{y}_{T+1:T+s}|\boldsymbol{y}_T) = \sum_{k \in K} p_k(\boldsymbol{y}_{T+1:T+s}|\boldsymbol{y}_T) p(k|\boldsymbol{y}_T), \tag{3.24}$$

donde $p(k|\boldsymbol{y}_T) \propto p_k(\boldsymbol{y}_T)p(k)$ es la probabilidad final del modelo M_k dada la trayectoria observada $\boldsymbol{y}_T$. Generalmente es difícil calcular $p_k(\boldsymbol{y}_T)$ de manera cerrada. En este caso

podemos reexpresar (3.24) en su forma extendida como

$$\begin{aligned} p(\boldsymbol{y}_{T+1:T+s}|\boldsymbol{y}_T) &= \sum_{k\in K} p(k|\boldsymbol{y}_T) \int p_k(\boldsymbol{y}_{T+1:T+s}|\boldsymbol{y}_T, \boldsymbol{\theta}_k)\pi_k(\boldsymbol{\theta}_k|\boldsymbol{y}_T)d\boldsymbol{\theta}_k \\ &= \sum_{k\in K} \int p_k(\boldsymbol{y}_{T+1:T+s}|\boldsymbol{y}_T, \boldsymbol{\theta}_k)p(\boldsymbol{\theta}_k, k|\boldsymbol{y}_T)d\boldsymbol{\theta}_k, \end{aligned} \tag{3.25}$$

donde $p(\boldsymbol{\theta}_k, k|\boldsymbol{y}_T) = \pi_k(\boldsymbol{\theta}_k|\boldsymbol{y}_T)p(k|\boldsymbol{y}_T)$ es una medida de probabilidad definida en el espacio de los modelos y sus correspondientes espacios parametrales, similar a (2.17). Si somos capaces de generar una muestra de tamaño N de $p(\boldsymbol{\theta}_k, k|\boldsymbol{y}_T)$, que denotamos por $\left\{(\boldsymbol{\theta}^{(n)}, k^{(n)}) : n = 1, .., N\right\}$, entonces podemos aproximar la mezcla (3.25) como

$$p(\boldsymbol{y}_{T+1:T+s}|\boldsymbol{y}_T) \approx \frac{1}{N}\sum_{n=1}^{N} p_{k^{(n)}}(\boldsymbol{y}_{T+1:T+s}|\boldsymbol{y}_T, \boldsymbol{\theta}^{(n)}), \tag{3.26}$$

y en cuyo caso la utilidad esperada (3.23) puede ser aproximada por

$$\frac{1}{N}\sum_{n=1}^{N} \int \log\left\{p_k(\boldsymbol{y}_{T+1:T+s}|\boldsymbol{y}_T)\right\} p_{k^{(n)}}(\boldsymbol{y}_{T+1:T+s}|\boldsymbol{y}_T, \boldsymbol{\theta}^{(n)})d\boldsymbol{y}_{T+1:T+s}, \tag{3.27}$$

si la densidad $p_k(\boldsymbol{y}_{T+1:T+s}|\boldsymbol{y}_T)$ puede expresarse de manera cerrada.

La implementación computacional con este enfoque es intensa, y requiere sobre todo de definir métodos de simulación eficientes. Los problemas de integración son casos particulares, como veremos en el capítulo 5.

Bajo la perspectiva $\mathcal{M}$-cerrada, la mezcla (3.25) resume toda la información relevante respecto a $Y_{T+1:T+s}$. En términos generales, esta alternativa permite incorporar los elementos de incertidumbre respecto al "verdadero" modelo que genera los datos. Este método ha sido considerado como una solución al problema de selección de modelos bajo un ambiente de incertidumbre, inclusive si consideramos en la clase $\mathcal{P}$ a todas las posibles mezclas de de las distribuciones predictivas de los modelos de la clase $\mathcal{M}$. En este caso la solución óptima Bayesiana resulta ser la mezcla Bayesiana $p(\boldsymbol{y}_{T+1:T+s}|\boldsymbol{y}_T)$ de todos los modelos contendientes en la clase, dada por (3.24).

La idea de combinar los modelos a través de (3.24) es simple, aunque en la práctica se presentan diferentes dificultades para su implementación en problemas prácticos no triviales, debido a dos factores fundamentalmente. Uno de ellos es que el número de

modelos en la mezcla puede ser demasiado grande, lo que imposibilita el cálculo del promedio predictivo de modelos (3.24), y para implementarlo se requiere del cálculo de la probabilidad final de cada modelo, que como vimos a través del Teorema de Bayes se obtiene como

$$p(k|\boldsymbol{y}_T) \propto p(k)p_k(\boldsymbol{y}_T), \tag{3.28}$$

donde $p(k)$ denota la probabilidad inicial del k-ésimo modelo y $p_k(\boldsymbol{y}_T)$ denota la verosimilitud integrada del k-ésimo modelo, que como vimos se calcula como

$$p_k(\boldsymbol{y}_T) = \int p_k(\boldsymbol{y}_T|\boldsymbol{\theta}_k)\pi_k(\boldsymbol{\theta}_k)d\boldsymbol{\theta}_k, \tag{3.29}$$

con $\boldsymbol{\theta}_k$ el conjunto de parámetros que indexan a la familia parametral a la que pertenece la distribución (densidad) $p_k(\cdot|\boldsymbol{\theta}_k)$ de la variable aleatoria o serie de tiempo de interés bajo el modelo k-ésimo, y $\pi_k(\boldsymbol{\theta}_k)$ es la densidad inicial sobre este conjunto de parámetros bajo el mismo modelo. El segundo problema se relaciona con el cálculo de las probabilidades finales de los modelos. Como mencionamos también, la solución de esta integral en sí misma no es una tarea sencilla, ya que incluso el estimador de Monte Carlo directo puede ser inestable. Desde luego, también es posible hacer uso de la teoría relacionada con los factores de Bayes, y calcular las probabilidades finales a partir de los estimadores numéricos sobre todos los factores de Bayes que se pueden definir en la clase de modelos contendientes. Una dificultad adicional se tiene cuando la densidad $_k(\boldsymbol{y}_T|\boldsymbol{\theta}_k)$ no es expresable de manera analítica. Estas dificultades complican el cálculo de las probabilidades finales de los modelos, y por ende el cálculo de la mezcla Bayesiana de modelos.

Para reducir el número de modelos contemplados en la mezcla, Madigan y Raftery (1994) propusieron un método que consiste en depurar el número de modelos de la clase $\mathcal{M}$ para ser mezclados, a través de un criterio heurístico de discriminación con base en la comparación de las distribuciones finales de cada modelo respecto al modelo con la probabilidad final máxima en una primera etapa, y en una segunda instancia se eliminan los modelos más complejos que tengan una probabilidad final similar a la de algún modelo más simple. Este método es conocido como la ventana de Ockham, cuyo nombre se atribuye al principio de parsimonia implementado en la segunda instancia del método invocando el principio de parsimonia de William de Ockham. Por otro lado, en el contexto

de regresión lineal y regresión lineal generalizada, Clyde (1999) propone algunos métodos de optimización en el cálculo de (3.24) con base en transformaciones ortogonales de la matriz de regresión, de manera que los métodos computacionales implementados presentan una mejora significativa en su ejecución. Estos dos procedimientos están diseñados para mezclar modelos dentro de la misma familia de modelos paramétricos.

Por otro lado, si no se desean calcular explícitamente las probabilidades finales de los modelos, es posible generar muestras de $p(k, \boldsymbol{\theta}_k|\boldsymbol{y}_T)$ mediante algoritmos sofisticados complicados, con la restricción adicional en algunos de éstos que las distribuciones iniciales sobre los parámetros asociados a cada modelo sea propia. Pro ejemplo, consideremos el algpritmo de Monte Carlo vía Cadenas de Markov con salto reversible (MCCMSR) propuesto por Green (1995) que brevemente describimos en la sección 2.3.6. Otra alternativa es propuesta por Madigan y York (1995), la que puede ser vista como un caso particular del algoritmo propuesto por Green (1995) para el caso en que los saltos se definan solamente entre las etiquetas de los modelos. La idea consiste en generar una cadena de Markov irreducible con el conjunto de índices de los modelos de la clase $\mathcal{M}$ como el espacio de estados de la cadena y distribución invariante $p(k|\boldsymbol{y}_T)$. Habiendo obtenido una trayectoria de la cadena $\{k^{(n)}\}$ de longitud n, podemos aproximar (3.24) por Monte Carlo como

$$\widehat{p}(y_{T+1:T+s}|\boldsymbol{y}_T) = \frac{1}{N}\sum_{n=1}^{N} p_{k^{(n)}}(y_{T+1:T+s}|\boldsymbol{y}_T), \tag{3.30}$$

cuando $p_k(y_T|\boldsymbol{y}_T)$ es conocida de manera cerrada para todos $k \in K$. Este último estimador converge con probabilidad uno a $p(y_{T+1:T+s}|\boldsymbol{y}_T)$ cuando el tamaño de la muestra $N \to \infty$. La forma como se construye la cadena considera una distribución de transición simétrica en el conjunto de índices que es definida solamente en una vecindad de un modelo en cada iteración, y considerando la distribución de transición propuesta para el movimiento entre los diferentes espacios parametrales a través de la distribución final de los parámetros de manera independiente, la cual en sus aplicaciones es conocida de manera cerrada. Así, la probabilidad de aceptación del movimiento en un cambio de dimensiones es

$$\alpha(k^{(m)}, k') = \text{mín}\left(1, \frac{p(k'|\boldsymbol{y}_T)J(k', k^{(m)})}{p(k^{(m)}|\boldsymbol{y}_T)J(k^{(m)}, k')}\right), \tag{3.31}$$

donde $J(k, k')$ denota la probabilidad de transición del modelo M_k al modelo $M_{k'}$.

Los métodos que hemos descrito brevemente han sido desarrollados y aplicados en la combinación de modelos anidados, aunque pueden extenderse a clase de modelos más ricas en cuanto a las formas estructurales consideradas. Además, en éstos se supone de manera implícita las perspectiva $\mathcal{M}$-cerrada, que como ya hemos mencionado en secciones anteriores, puede conducirnos a sesgos en el análisis.

Draper (1995) propuso abordar el problema del promedio Bayesiano de modelos con una perspectiva más general, suponiendo que las formas estructurales de los modelos sean diferentes entre sí de manera que cada modelo genera una distribución predictiva "distinta" a los demás, y que cada modelo contenga información estructral diferente y particular. Para resolver el problema de incertidumbre del modelo, Draper (1995) propuso extender un modelo del cual tengamos información inicial respecto a un comportamiento eficiente, y extender la incertidumbre de este modelo mediante la adición de un parámetros adicional en el análisis, de manera que el modelo inicial sea visto como un caso particular de la clase más general. Esta alternativa se conoce como *extensión del modelo.* La extensión como la propuesta por Draper surge es ampliamente usada en mucho contextos y surge de manera natural para algunos casos particulares, por ejemplo en el caso de modelos localización y escala, particularmente en el modelo Gaussiano, esta extensión es ampliamente aplicada de manera que la distribución predictiva pueda ser vista como una mezcla continua de escales de modelos Gaussianos. Generalmente esta mezcla está definida en términos del parámetro de escala; de esta manera la distribución predictiva resulta definida como una mezcla de escala de distribuciones Gaussianas. Este tipo de análisis puede extenderse a otras familias paramétricas, dando en el caso mencionado origen a una extensión continua del modelo.

3.5.2 Criterio Predictivo $\mathcal{M}$-semiabierta

Al abordar el problema de selección de modelos con la perspectiva $\mathcal{M}$-abierta, lo que se hace de manera implícita es asignar una probabilidad uno a la clase de modelos paramétricos $\mathcal{M}$, y una probabilidad igual a cero a los modelos paramétricos fuera de esa clase. Este puede no ser un problema serio, si se tiene información sustentable de que

los modelos incorporados en la clase son buenos modelos estadísticos para el fenómeno estudiado. Por otro lado se tiene que en realidad se está dejando fuera, en la mayoría de los casos, a modelos que posiblemente representen una mejor alternativa a los propuestos en la clase. En este sentido, el sentido de optimalidad del criterio es cuestionable, pues se están dejando fuera modelos que posiblemente sean mejores alternativas a las propuestas. Por ejemplo, en el caso de series de tiempo es común restringir la clase $\mathcal{M}$ a una subclase de modelos simétricos y unimodales, aunque en algunos casos se tiene un persistente efecto de sesgo manifestado por los datos, y en ese caso si sólo se consideraran los modelos simétricos estaríamos incurriendo en sesgos por los modelos, a pesar del procedimiento de selección. Para reducir este posible sesgo, se incorporan a la clase algunos modelos con distribuciones sesgadas en alguna o ambas direcciones. De manera semejante podemos pensar en el sesgo de los modelos cuando se eligen a los modelos con distribuciones unimodales, asignando una probabilidad cero a los modelos con distribuciones con dos o más modas. En este sentido, de nuevo la optimalidad del modelo seleccionado puede ser cuestionable, ya que se está comparando con una clase de modelos que posiblemente se amplia pero no lo suficiente.

Por otro lado en la perspectiva $\mathcal{M}$-abierta, la selección de modelos se realiza suponiendo que ninguno de los modelos dentro de la clase $\mathcal{M}$ es el modelo verdadero, y para cada modelo paramétrico se evalúa su capacidad predictiva respecto a una clase mucho más flexible que la de los modelos paramétricos dentro de la clase, que de alguna forma representa una mejor alternativa a los modelos paramétricos en términos de predicción. En la concepción original de Bernardo y Smith (1994) la clase flexible corresponde a todas las posibles distribuciones de probabilidad definidas sobre el espacio de la variable aleatoria futura. Para el caso de variables independientes, Bernardo y Smith (1994) aproximaron esta distribución con la distribución empírica. Posteriormente Gutiérrez-Peña y Walker (2001) emplearon los procesos Dirichlet para asignar una distribución inicial e inferir sobre la que sería la verdadera distribución del proceso, para el caso de variables aleatorias intercambiables. En el primer caso se asigna una probabilidad uno a una familia más flexible, y en el segundo caso a la clase de todas las posibles distribuciones discretas.

En estos dos casos, el sentido de optimalidad del criterio de selección tiene un mayor sustento, pues en realizada se está eligiendo un modelo respecto a una clase mucho más amplia de modelos. Más adelante discutiremos como podemos aplicar esta perspectiva en el contexto de series de tiempo.

Por otro lado, considerando la perspectiva $\mathcal{M}$-abierta al igual que en la perspectiva $\mathcal{M}$-cerrada, podemos extender la definición del espacio de acciones al incorporar todas las posibles mezclas de las densidades predictivas de los modelos en la clase $\mathcal{M}$. En la perspectiva $\mathcal{M}$-abierta la mezcla tiene más sentido que en la perspectiva cerrada, ya que idealmente la clase de modelos a la que se le asigna probabilidad uno es mucho más flexible y contiene formas estructurales más generales que en el caso cerrado, y la mezcla óptima puede incluso no coincidir con la mezcla Bayesiana de los modelos. La mezcla o combinación predictiva de modelos está definida como

$$p_\eta(y_{T+1}|\boldsymbol{y}_T) = \sum_{k\in K} p_k(y_{T+1}|\boldsymbol{y}_T)\eta(k), \tag{3.32}$$

donde $\eta(\cdot)$ es una distribución de probabilidad definida sobre el conjunto índice de los modelos de la clase $\mathcal{M}$, que define la distribución de la mezcla. Alternativamente podemos bajar el nivel del problema de selección para el caso de querer elegir sólo un modelos paramétrico, en cuyo caso solamente se consideran las distribuciones de la mezclas degeneradas en cada modelo $k \in K$, mientras que consideramos también a la mezcla de modelos paramétricos como una alternativa interesante para tratar la incertidumbre de un modelo manteniendo ciertas estructuras paramétricas específicas.

Es natural pensar que esta mezcla no necesariamente coincide con la mezcla definida a través de los pesos dados por las probabilidades finales de cada modelo. En este caso la mezcla puede ser definida en términos de un conjunto índice finito o numerable, como en el caso de (3.32) o de manera continua como propone Draper (1995), y revisado posteriormente por Walker *et al.* (2001), en cuyo caso la mezcla (3.32) estará definida como una integral definida a su vez sobre el espacio índice de los modelos.

En este caso, la función $\eta(\cdot)$ que define la mezcla es desconocida. Para realizar la mezcla es necesario definirla de manera óptima bajo un cierto criterio de optimalidad. En este caso podemos implementar el criterio propuesto por Walker *et al.* (2001), que es una

generalización del criterio de selección predictivo propuesto por Gutiérrez-Peña y Walker (2001), que consiste en ver el problema de selección de la distribución de la mezcla como un problema de decisión estadística con los siguientes elementos:

- **Espacio de decisiones**

El *espacio de decisiones*, $\mathcal{A}$, consiste en el conjunto de todas las posibles mezclas predictivas de la forma (3.32). Debemos notar que para cada modelo, la densidad predictiva $p_k(y_{T+1}|\boldsymbol{y}_T)$ es calculada de manera paralela, y al momento de la toma de decisiones son conocidas, ya sea de manera cerrada o mediante una aproximación, usualmente a través de las técnicas de Monte Carlo. Entonces todas las posibles mezclas estarán definidas por la distribución de mezcla $\eta(\cdot)$. De esta manera el espacio de acciones puede simplificarse definiendo así a $\mathcal{A}$ como el conjunto de todas las posibles funciones de probabilidad definidas sobre el espacio índice de los modelos, que denotaremos por

$$\mathcal{A} = \{\eta(\cdot)|\ \eta \text{ es una medida de probabilidad definida sobre } K\}. \tag{3.33}$$

- **Espacio de 'estados de la naturaleza'**

Si nuestro interés es el de obtener predicciones respecto a las variables $Y_{T+1:T+s}$, entonces el espacio de *estados de la naturaleza* es $\mathcal{Y}_{T+1:T+s} \subset \Re$, que es el espacio de los posibles valores que puede tomar la variable $Y_{T+1:T+s}$. Debido a que deseamos realizar predicciones de manera general, nuestro estado de incertidumbre relevante reside en la verdadera densidad (o distribución) predictiva de $Y_{T+1:T+s}$ condicional en $\boldsymbol{y}_T$. Podemos pensar que la serie al tiempo t está determinada por una relación de localización y escala de la forma

$$y_t = f(y_{t-1}, ..., y_{t-p}) + \tau\varepsilon_t, \tag{3.34}$$

donde $f : \Re^p \to \Re$ es una función suave desconocida que determina la forma estructural del nivel de la serie, τ es una parámetro de escala desconocido de los errores aleatorios, y $\{\varepsilon_t\}$ es una sucesión de ruido blanco con media cero y varianza uno. De manera general, podemos suponer que el parámetro de escala τ puede variar en el tiempo, como una función de los datos pasados. Asimismo, podemos incorporar variables exógenas en la

dependencia de la serie definida por (3.34). De esta forma, los nuevos elementos relevantes de incertidumbre en (3.34) están dados por $f(\cdot)$ y τ. Así, la función de densidad de la serie al tiempo $T+1$ estará determinada como

$$p(y_{T+1}|\boldsymbol{y}_T, f, \tau) = p(y_{T+1}|f(\boldsymbol{y}_T), \tau). \tag{3.35}$$

Podemos aproximar las relaciones definidas por (3.34) mediante procedimientos Bayesianos paramétricos más flexibles o semiparamétricos, de manera que seamos capaces de obtener una densidad predictiva final para $Y_{T+1:T+s}$ condicional en $\boldsymbol{y}_T$ mediante un proceso de marginalización de la aproximación a (3.35) respecto a todos los elementos irrelevantes de su aproximación. Este método se basa fundamentalmente en nuestra capacidad para poder definir o determinar un modelo semiparamétrico que sea más flexible a los modelos paramétricos considerados para la mezcla. Gutiérrez-Peña (1997) consideró modelar semiparamétricamente el componente f a través de procesos Gaussianos, considerando el componente distribucional de los errores completamente paramétrico a través de una distribución Gaussiana. De manera alternativa podemos elegir modelar semiparamétricamente el componente distribucional de los errores y modelar de manera completamente paramétrica el componente f. En nuestro conocimiento, a la fecha no existe un modelo Bayesiano que modele semiparamétricamente ambos componentes. En el capítulo 4 realizaremos una breve descripción acerca de algunos modelos semiparamétricos de inferencia en series de tiempo.

- **Función de utilidad**

Utilizaremos la función de utilidad de puntaje logarítmico, de la cual hemos descrito ya algunas propiedades relevantes (ver el apéndice A). De manera simplificada, eliminando todos los elementos irrelevante, ésta puede resumirse de la forma

$$u(p_\eta(\cdot|\boldsymbol{y}_T), y_{T+1}) = \log\left\{p_\eta(y_{T+1}|\boldsymbol{y}_T))\right\}. \tag{3.36}$$

Finalmente la solución óptima Bayesiana consiste en elegir la mezcla que maximice la utilidad esperada

$$\overline{u}(p_\eta(\cdot|\boldsymbol{y}_T)) = \int \log\left\{p_\eta(y_{T+1}|\boldsymbol{y}_T))\right\} p(y_{T+1}|\boldsymbol{y}_T) dy_{T+1}, \tag{3.37}$$

donde $p(y_{T+1}|\boldsymbol{y}_T)$ es la densidad predictiva de la serie al tiempo T, que como ya mencionamos surge mediante un proceso de marginalización de la aproximación a (3.35).

Como observaron Walker *et al.* (2001), si nuestro objetivo es estrictamente el de producir predicciones de una variable futura, entonces la mezcla de modelos con este criterio es necesariamente más eficiente que la selección predictiva de sólo un modelo. En este caso, la densidad predictiva de cada modelo particular k en K, puede ser vista como una mezcla (3.32) con una distribución de probabilidad degenerada en el modelo k-ésimo, i.e. $\eta(k) = 1$. Así, la utilidad obtenida a través de la selección predictiva será menor o igual a la obtenida mediante la mezcla óptima.

La implementación de este método de combinación depende fundamentalmente de los modelos que deseemos incorporar en la mezcla. Las características de éstos inducirán la forma mínima de la aproximación a la "verdadera" densidad predictiva que debemos utilizar para obtener (3.37).

3.6 Ejemplo: Selección del Orden en Modelos AR

Para ilustrar los criterios descritos anteriormente consideramos el problema simple, pero importante, de seleccionar el orden dentro de una clase de modelos autorregresivos lineales. Sea $\{Y_t : t = 1, 2, ...\}$ una serie de tiempo escalar en tiempo discreto y sea $\{y_t\}_{t=1}^{T}$ la trayectoria de la serie observada entre los tiempos 1 y T. Los modelos contendientes en este problema pertenecen a la clase $\mathcal{M} = \{M_k : k = p_{\text{mín}}, ..., p_{\text{máx}}\}$ donde cada modelo, M_k, está determinado por la ecuación

$$y_t = \sum_{j=1}^{k} \phi_{j,k} y_{t-j} + \varepsilon_{t,k}, \tag{3.38}$$

con $\boldsymbol{\phi}_k = (\phi_{1,k}, ..., \phi_{k,k})'$ un vector de coeficientes de autorregresión desconocido, y $\{\varepsilon_{t,k}\}$ es un proceso de ruido blanco Gaussiano con precisión τ_k desconocida.

Bajo el enfoque Bayesiano cada modelo estará completamente determinado con la incorporación de una distribución inicial sobre los elementos de incertidumbre del modelo, que en este caso son parámetros $(\boldsymbol{\phi}_k, \tau_k)$, para cada modelo M_k, y de esta forma el modelo

Bayesiano queda completamente determinado por

$$M_k = \left\{ p_k(y_t|\boldsymbol{y}_{t-1}, \boldsymbol{\phi}_k, \tau_k), \pi_k(\boldsymbol{\phi}_k, \tau_k) \right\}, \tag{3.39}$$

donde $p_k(y_t|\boldsymbol{y}_{t-1}, \boldsymbol{\phi}_k, \tau_k) = N(y_t|\boldsymbol{y}'_{t-k:t}\boldsymbol{\phi}_k, \tau_k)$, con $\boldsymbol{y}_{t-k:t} = (y_{t-1}, ..., y_{t-k})'$, y $\pi_k(\boldsymbol{\phi}_k, \tau_k)$ es la distribución inicial de $(\boldsymbol{\phi}_k, \tau_k)$ condicional en el k-ésimo modelo. En la notación de (3.39) el subíndice denota un condicionamiento implícito de las densidades respecto a la forma estructural del k-ésimo modelo. Para efectos prácticos, supongamos que los valores iniciales de la serie hasta el orden máximo de comparación, $\boldsymbol{y}_{p_{\text{máx}}} = (y_{p_{\text{máx}}}, ..., y_1)'$, son conocidos. En este caso la clase de modelos por comparar, $\mathcal{M}$, está formada por todos los modelos autorregresivos lineales de orden $p_{\text{mín}}$ a $p_{\text{máx}}$.

Una ventaja que tiene el empleo del criterio de selección $\mathcal{M}$-abierta, es que no es necesario que asignemos distribuciones iniciales propias en los parámetros, mientras que para el caso $\mathcal{M}$-cerrada, en caso de querer instrumentar el muestreador MCCMSR, es necesario que las distribuciones iniciales sean propias. En nuestro caso particular asignamos distribuciones iniciales no informativas sobre los parámetros de cada modelo (vea la sección 2.4.1), de manera que la distribución final de los parámetros para cada modelo M_k es

$$\pi_k(\boldsymbol{\phi}_k, \tau_k|\boldsymbol{y}_T) = N_k(\boldsymbol{\phi}_k|\boldsymbol{f}_k, \tau_k^{-1}\boldsymbol{B}_k^{-1})Ga(\tau_k|a_k/2, b_k/2), \tag{3.40}$$

con $\boldsymbol{f}_k = (\boldsymbol{Y}'_{T,k}\boldsymbol{Y}_{T,k})^{-1}(\boldsymbol{Y}'_{T,k}\boldsymbol{y}_T)$, $\boldsymbol{B}_k = (\boldsymbol{Y}'_{T,k}\boldsymbol{Y}_{T,k})$, $a_k = T - p_{\text{máx}} - k$, y $b_k = (\boldsymbol{y}_T - \boldsymbol{Y}_{T,k}\boldsymbol{f}_k)'(\boldsymbol{y}_T - \boldsymbol{Y}_{T,k}\boldsymbol{f}_k)$. Y la distribución predictiva un tiempo adelante al tiempo T para el mismo modelo k-ésimo es:

$$p_k(y_{T+1}|\boldsymbol{y}_T) = St(y_{T+1}|\boldsymbol{y}'_{T-k:T}\boldsymbol{f}_k, \frac{b_k}{a_k}(1 + \boldsymbol{y}'_{T-k:T}\boldsymbol{B}_k^{-1}\boldsymbol{y}_{T-k:T}), T - p_{\text{máx}} - k). \tag{3.41}$$

En la siguiente sección describimos dos criterios predictivos de determinación del orden de un modelo AR. Para el caso del criterio predictivo $\mathcal{M}$-cerrada consideramos la asignación de distribuciones iniciales propias pero difusas sobre los parámetros de cada modelo, donde estas pueden calcularse de manera directa mediante la representación matricial en forma de regresión de cada modelo (vea al apéndice B.1). Posteriormente ejemplificaremos estos criterios con un proceso AR Gaussiano simulado y una serie económica

real. Al final presentamos la comparación con otros criterios de determinación del orden de un modelo autorregresivo que son ampliamente utilizados, como el criterio AIC (Akaike, 1973) y el criterio BIC (Schwarz, 1978).

3.6.1 Selección Predictiva del Modelo

La determinación del orden de un modelo AR es en esencia un problema de selección de modelos, por ende estructuraremos este problema mediante el esquema de teoría de decisión que describimos en la sección 3.5. Si consideramos realizar predicciones un tiempo adelante de la serie tenemos que los elementos del problema de decisión son:

- **Espacio de Acciones**

El espacio de acciones está compuesto por todas las posibles distribuciones predictivas generadas por los modelos de la clase $\mathcal{M}$,

$$\mathcal{A} = \{p_k(\cdot|\boldsymbol{y}_T) : k = p_{\text{mín}}, ..., p_{\text{máx}}\}, \tag{3.42}$$

donde las densidades predictivas $p_k(\cdot|\boldsymbol{y}_T)$ están definidas sobre $\mathcal{Y}_{T+1}$, de acuerdo a (3.41). Para esta caso particular éstas son conocidas de manera analítica cerrada y no consideramos en esta clase las posibles mezclas de algunas de estas densidades como una alternativa dentro del proceso de decisión.

- **Espacio de 'estados de la naturaleza'**

Como ya mencionamos, el espacio de 'estados de la naturaleza' en ambos enfoques respecto a la clase $\mathcal{M}$ es el espacio de posibles valores futuros del proceso en el tiempo $T+1$, i.e. el espacio $\mathcal{Y}_{T+1} \subset \Re$.

- **Función de Utilidad**

Utilizamos la función de puntaje logarítmico como función de utilidad

$$u\left(p_k(\cdot|\boldsymbol{y}_T), y_{T+1}\right) = \log\left\{p_k(y_{T+1}|\boldsymbol{y}_T)\right\}. \tag{3.43}$$

De esta forma, el modelo óptimo es el que maximice en $\mathcal{A}$ la utilidad esperada

$$\overline{u}(p_k(\cdot|\boldsymbol{y}_T)) = \int \log\left(p_k(y_{T+1}|\boldsymbol{y}_T)\right) p(y_{T+1}|\boldsymbol{y}_T) dy_{T+1}, \tag{3.44}$$

donde $p(y_{T+1}|\boldsymbol{y}_T)$ es la "verdadera" densidad predictiva para Y_{T+1}, que en realidad desconocemos. Bajo la perspectiva $\mathcal{M}$-cerrada esta densidad es el resultado de una mezcla predictiva entre todos los modelos de la clase $\mathcal{M}$. Para efectos prácticos podemos implementar esta mezcla a través del método de Monte Carlo vía cadenas de Markov con salto reversible (MCCMSR) como describimos más adelante.

En el caso la perspectiva $\mathcal{M}$-abierta la densidad predictiva de referencia para calcular la utilidad esperada (3.44) debe pertenecer a una familia de modelos más flexibles que englobe como caso particular a los modelos dentro de la clase $\mathcal{M}$. Dado que los modelos autorregresivos modelan los cambios en la parte sistemática del proceso respecto a su nivel (vea la primera sección del capítulo 4), podemos considerar utilizar la distribución predictiva del modelo definido por la relación

$$y_t = f(y_{t-1}, ..., y_{t-p_{\text{máx}}}) + \varepsilon_t, \tag{3.45}$$

donde $f : \Re^{p_{\text{máx}}} \to \Re$ denota una función suave de autorregresión desconocida, y $\{\varepsilon_t\}$ es un proceso de ruido blanco Gaussiano con precisión τ desconocida. Para ejemplificar este criterio elegimos aproximar la función de regresión (autorregresión) mediante un polinomio de segundo orden, el cual da origen a una clase particular de modelos autorregresivos no lineales en las variables (NAR) (vea Fitzgerald *et al.* (1999)), aunque lineal en los parámetros, lo que simplifica su análisis significativamente al representarlo como un modelo de regresión lineal.

Modelo Autorregresivo no Lineal (NAR)

En esencia, el criterio de selección de modelos predictivo que aquí adoptamos consiste en proponer un modelo flexible que no imponga las restricciones estructurales de los modelos bajo comparación. Un requisito mínimo consiste en proponer un modelo que contenga al menos todas las estructuras de los modelos bajo comparación, aunque en este caso la optimalidad de la selección estará limitada respecto a un modelo aún restrictivo.

Dado que los modelos autorregresivos modelan en esencia los cambios en nivel de la serie, podemos proponer comparar los modelos respecto a la distribución predictiva un modelo autorregresivo no lineal, que es una aproximación a (3.45), de la forma

$$y_t = \sum_{i=1}^{p_{\text{máx}}} \phi_i y_{t-i} + \sum_{j=1}^{p_1} \sum_{k=j}^{p_2} \phi_{j,k} y_{t-j} y_{t-k} + \varepsilon_t, \tag{3.46}$$

donde p_1 y p_2 con los coeficientes del componente de segundo orden de autorregresión, con p_1 y p_2 menores o iguales a $p_{\text{máx}}$, y $\{\varepsilon_t\}$ es un proceso de ruido blanco Gaussiano con una precisión τ desconocida.

De manera semejante al modelo autorregresivo lineal, el modelo (3.46) puede reexpresarse y analizarse como un modelo de regresión lineal. En este caso los parámetros del modelo son $(\boldsymbol{\phi}, \tau)$, donde $\boldsymbol{\phi} = (\phi_1, .., \phi_{p_{\text{máx}}}, \phi_{1,1}, ..., \phi_{p_1,p_1}, \phi_{1,2}, ..., \phi_{p_1,p_2})'$. Si utilizamos una distribución inicial de referencia para $(\boldsymbol{\phi}, \tau)$, podemos obtener una distribución predictiva para Y_{T+1} de manera analítica cerrada como

$$p(y_{T+1}|\boldsymbol{y}_T) = St(y_{T+1}|\boldsymbol{y}'_{T-p_{\text{máx}}:T}\boldsymbol{f}, \frac{b}{a}(1 + \boldsymbol{y}'_{T-p_{\text{máx}}:T}\boldsymbol{B}^{-1}\boldsymbol{y}_{T-p_{\text{máx}}:T}), \eta), \tag{3.47}$$

con $\boldsymbol{f}$, $\boldsymbol{B}$, a y b definidos de manera semejante al caso del modelo autorregresivo lineal. Esta densidad nos sirve como la densidad de referencia para comparar los modelos a través de puntaje logarítmico esperado (3.44).

Modelo AR vía MCCM con Salto Reversible

En la perspectiva $\mathcal{M}$-cerrada, la selección de los modelos se basa en calcular el puntaje logarítmico esperado (3.44) respecto a la densidad

$$\begin{aligned} p(y_{T+1}|\boldsymbol{y}_T) &= \sum_{k=p_{\text{mín}}}^{p_{\text{máx}}} p(k|\boldsymbol{y}_T) p_k(y_{T+1}|\boldsymbol{y}_T), \\ &= \sum_{k=p_{\text{mín}}}^{p_{\text{máx}}} p(k|\boldsymbol{y}_T) \int p_k(y_{T+1}|\boldsymbol{y}_T, \boldsymbol{\phi}_k, \tau_k) p_k(\boldsymbol{\phi}_k, \tau_k|\boldsymbol{y}_T) d\boldsymbol{\phi}_k d\tau_k, \end{aligned} \tag{3.48}$$

donde $p(k|\boldsymbol{y}_T)$ denota la masa de probabilidad final del modelo M_k, y $p_k(y_{T+1}|\boldsymbol{y}_T)$ denota la densidad predictiva del modelo M_k un tiempo adelante al tiempo T. Esta densidad es difícil de calcular analíticamente. Sin embargo, podemos aproximar el puntaje logarítmico

(3.44) mediante su estimador de Monte Carlo a través de una muestra $\{y_{T+1}^{(i)} : i = 1, ..., N\}$ de la densidad predictiva (3.48). La muestra y la aproximación a (3.44) se obtienen mediante un proceso de marginalización de una muestra de tamaño N de la distribución final conjunta $p(y_{T+1}, k, \boldsymbol{\phi}_k, \tau_k|\boldsymbol{y}_T)$, denotada por $\{(y_{T+1}^{(i)}, k^{(i)}, \boldsymbol{\phi}_k^{(i)}, \tau_k^{(i)}) : i = 1, ..., N\}$. Esta muestra la podemos obtener usando el algoritmo de Monte Carlo vía Cadenas de Markov con saltos reversibles (MCCMSR), que describiremos brevemente a continuación.

Si asignamos una distribución inicial conjugada, preferentemente difusa, sobre los parámetros $(\boldsymbol{\phi}_k, \tau_k)$ en cada modelo M_k, podemos obtener su correspondiente distribución final de manera cerrada como (vea el apéndice B.1):

$$p_k(\boldsymbol{\phi}_k, \tau_k|\boldsymbol{y}_T) = N_k(\boldsymbol{\phi}_k|\boldsymbol{f}_k, \tau_k^{-1}\boldsymbol{B}_k^{-1})Ga(\tau_k|a_k/2, b_k/2), \tag{3.49}$$

donde $\boldsymbol{f}_k$ es un vector de dimensión k, $\boldsymbol{B}_k$ es una matriz simétrica positivo definida de dimensiones $k \times k$, a_k y b_k son valores reales positivos, todos éstos definidos como en el caso del modelo lineal de regresión (vea el apéndice B.1).

Para la clase de modelos $\mathcal{M}$ podemos asignar una distribución uniforme discreta

$$p(k) = \frac{1}{p_{\text{máx}} - p_{\text{mín}} + 1}\mathbf{1}_{\{p_{\text{mín}}, ..., p_{\text{máx}}\}}(k). \tag{3.50}$$

Con el algoritmo de MCCMSR podemos obtener una muestra simulada de la distribución final conjunta

$$p(y_{T+1}, k, \boldsymbol{\phi}_k, \tau_k|\boldsymbol{y}_T) = p(k|\boldsymbol{y}_T)p_k(\boldsymbol{\phi}_k, \tau_k|\boldsymbol{y}_T)p_k(y_{T+1}|\boldsymbol{\phi}_k, \tau_k, \boldsymbol{y}_T). \tag{3.51}$$

Dado que en este caso es posible obtener la distribución final de los parámetros de cada modelo y también es posible obtener muestras simuladas de ésta de manera directa, se tiene que la probabilidad de aceptación de los movimientos entre diferentes modelos y espacios parametrales puede simplificarse como (2.22) de la forma

$$\alpha(k^{(i)}, k') = \text{mín}(1, Q), \tag{3.52}$$

con

$$Q = \frac{p(\boldsymbol{y}_T|k')}{p(\boldsymbol{y}_T|k^{(i)})} \times \frac{J(k', k^{(i)})}{J(k^{(i)}, k')}, \tag{3.53}$$

donde el primer término corresponde al Factor de Bayes (FB) entre los modelos $M_{k'}$ y $M_{k^{(i)}}$, el cual puede calcularse analíticamente (vea el apéndice B.1), y $J(k,k')$ denota la probabilidad de transición del modelo M_k al modelo $M_{k'}$. En este caso particular elegimos una probabilidad de transición simétrica definida solamente en la vecindad del k-ésimo modelo para cada iteración, i.e. $J(k,k') = c_1\mathbf{1}(k'=k-1)+c_2\mathbf{1}(k'=k+1)+c_3\mathbf{1}(k'=k)$, con los c_i's números reales positivos tales que $c_1+c_2+c_3=1$ y $c_1=c_2$. La descripción del algoritmo de MCCMSR para la clase de modelos AR lineales es la siguiente.

Algoritmo 1: MCCMSR - Modelo Autorregresivo Lineal

1. Determinamos los valores iniciales de la cadena $(k, \boldsymbol{\phi}_k, \tau_k)$.

2. Iniciamos la cadena de Markov con salto reversible

   (a) Proponemos el tipo de movimietno de la cadena en el modelo de k a k'

      i. NACIMIENTO con probabilidad c_1 (si el movimiento es admisible), $k' = k+1$

      ii. MUERTE con probabilidad c_2 (si el movimiento es admisible), $k' = k-1$

      iii. ESTÁTICO con probabilidad c_3, $k' = k$

   (b) Condicional en k' y τ_k, obtenemos una muestra de $\boldsymbol{\phi}'_{k'} \sim N_{k'}(\boldsymbol{\phi}_k|\boldsymbol{f}_{k'}, \tau_k^{-1}\boldsymbol{B}_{k'})$.

   (c) Condicional en k' y $\boldsymbol{\phi}'_{k'}$, obtenemos una muestra de $\tau'_{k'} \sim Ga(\tau_{k'}|a/2, b/2)$, con b calculada usando $\boldsymbol{\phi}'_{k'}$.

   (d) Evaluamos la probabilidad de aceptación del movimiento. Generamos $u \sim U(0,1)$.

      i. Aceptamos el movimiento si $u < Q$, con Q dada en (3.53)

      ii. Rechazamos en otro caso

3. Obtenemos una muestra de la densidad predictiva un paso adelante al tiempo T, $y_{T+1} \sim N(y_{T+1}|y'_{T-k:T}\boldsymbol{\phi}_k, \tau_k^{-1})$.

4. Continuamos hasta obtener la convergencia de la cadena.

En el caso en que $k' = k^{(j)}$ durante el paso 2, se tiene que los pasos subsecuentes corresponden al algoritmo de Metropolis-Hastings tradicional, con las distribuciones propuestas independientes. Los cálculos realizados en esta sección se obtuvieron mediante las funciones descritas en el apéndice D.

A continuación presentamos los resultados de los criterios predictivos que mencionamos para el caso de una serie de tiempo simulada y una serie de tiempo real. Los resultados de la selección se comparan con los criterios AIC y BIC, que son ampliamente usados en la práctica.

Los criterios AIC y BIC son criterios de selección de modelos que se basan en la estimación por máxima verosimilitud de los parámetros asociados asociados a un modelo, considerando además un factor de penalización por sobreparametrización. Conceptualmente tienen su origen en la teoría de la información, y generalmente son consistentes entre sí. El lector interesado puede referirse a Akaike (1973) y Schwarz (1978). Los puntajes calculados para cada modelo en ambos criterios pueden expresarse de manera genérica como:

$$f(M_k) \approx -2\log\left(p_k(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_k)\right) + h(M_k), \tag{3.54}$$

donde $\hat{\boldsymbol{\theta}}_k$ es el estimador de máxima verosimilitud del vector de parámetros asociados al modelo M_k, y $h(M_k)$ es un factor de penalización sobre el número de parámetros asociados al modelo. Con $h(M_k) = 2q(k)$ se obtiene el criterio de información de Akaike (AIC), donde $q(k)$ denota al número de parámetros en el modelo M_k, y con $h(M_k) = \log(n)q(k)$ se obtiene el criterio de información de Schwarz (BIC), donde n es el número de observaciones utilizadas para estimar el modelo. El modelo óptimo para cada criterio es el que minimiza el puntaje (3.54). Generalmente la penalización sobre el número de parámetros es más fuerte en el criterio BIC, por lo que presumiblemente selecciona un modelo con más parsimonia que el criterio AIC. Para aplicar estos criterios en la comparación de modelos autorregresivos lineales Gaussianos, -2 veces la log-verosimilitud en (3.54) se sustituye por $n\log(\hat{\sigma^2})$, eliminando términos irrelevantes, donde $\hat{\sigma^2}$ denota el estimador de máxima verosimilitud de la varianza de los errores aleatorios σ^2.

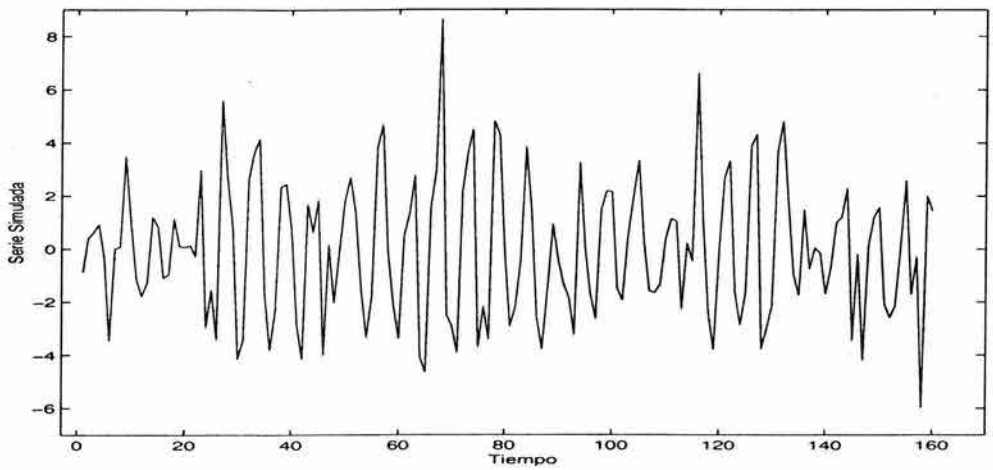


Figura 3.1: Proceso AR(4) simulado.

3.6.2 Proceso Simulado

Para ejemplificar el procedimiento de selección consideraremos un proceso autorregresivo lineal Gaussiano simulado de orden cuatro con una varianza en los errores igual a 5 y los coeficientes de autorregresión ϕ =(0.02,0.02969,-0.68,0.0068)'. En la figura 3.1 presentamos la trayectoria de la serie simulada de 120 datos.

Para modelar la serie consideramos en la clase $\mathcal{M}$ propuestà a todos los modelos AR lineales hasta el orden máximo 20. En el enfoque predictivo de selección $\mathcal{M}$-abierta consideramos el modelo flexible NAR(20,3,1), después de realizar una inspección sobre el comportamiento predictivo del modelo NAR con diferentes parámetros usando las medidas del apéndice C. En la figura 3.2 graficamos las utilidades esperadas (puntajes logarítmicos esperados) para los modelos de orden de autorregresión $p_{\text{mín}} = 1$ a $p_{\text{máx}} = 20$. Observamos que el modelo con $p = 4$ maximiza la utilidad esperada en el espacio de acciones $\mathcal{A}$, pero también observamos que la diferencia respecto al modelo con $p = 3$ es marginalmente igual a cero (vea también el cuadro 3.1), i.e. bajo nuestro esquema de comparación estos dos modelos son similares en términos de predicción.

Para aproximar la distribución predictiva "verdadera" en el criterio predictivo $\mathcal{M}$-cerrada usamos el algoritmo 1 (MCCMSR), considerando un periodo inicial de calentamiento de la cadena de 32,000 iteraciones y una muestra simulada de 10,000 observaciones con la probabilidad de transición entre modelos dada por $c_1 = c_2 = 0.3$ y $c_3 = 0.4$.

Elegimos un vector nulo k-dimensional y una matriz $\boldsymbol{B}_0 = 100\boldsymbol{I}_k$ como hiperparámetros de la distribución inicial sobre $\boldsymbol{\phi}_k$ y $a_0 = b_0 = 0.001$ para la distribución inicial sobre τ_k, en cada modelo M_k. En la figura 3.3 presentamos la traza (trayectoria) de la evolución de la cadena para el orden del modelo y los valores predictivos para el tiempo $T + 1$. Los puntajes logarítmicos esperados en el caso $\mathcal{M}$-cerrada corresponden a los estimadores de Monte Carlo calculados con la muestra de tamaño 10,000 de la distribución predictiva para el tiempo $T + 1$ marginalizada de la muestra final de MCCMSR (vea la figura 3.3 (b)). El resultado que obtenemos en el caso cerrado es similar al de la perspectiva abierta, el modelo óptimo es el que tiene orden $p = 4$ y también existe una gran similitud, en términos de predicción, con el modelo $p = 3$ (vea el cuadro 3.1).

Considerando el criterio Bayesiano de máxima probabilidad, el modelo óptimo corresponde al modelo AR(3) (vea la figura 3.3 (a)). El modelo AR(3) resulta ser el modelo óptimo también usando los criterios AIC y BIC (vea el cuadro 3.1). Recordemos que los criterios AIC y BIC penalizan de alguna forma la sobreparametrización del modelo, y es destacable que el criterio Bayesiano de máxima probabilidad coincida con estos criterios, estableciendo una especie de navaja de Ockham [1] a posteriori de manera natural. Este no es un resultado que se encuentre en la práctica de manera general usando el paradigma Bayesiano para la selección de modelos. Rasmussen y Ghahramani (2001) realizaron un estudio donde el resultado de obtener una navaja de Ockham usando el criterio Bayesiano de máxima probabilidad depende fundamentalmente de la asignación de la distribución de probabilidades definida sobre el espacio de los modelos contendientes. Existe también la posibilidad de que modelos más sofisticados o complejos se ajusten adecuadamente a un conjunto de datos, en tal caso puede ser que este comportamiento se vea reflejado en el procedimiento de selección de modelos.

[1] La navaja de Ockham es un principio metodológico que ha tenido una gran aceptación en las ciencias en general, y en la estadística en particular. Se atribuye al filósofo William de Ockham (1287-1347), aunque él no lo haya expresado como hoy en día se entiende. El principio consiste esencialmente en modelar o interpretar un fenómeno de interés usando el modelo más simple que represente eficientemente al fenómeno después de haber observado la evidencia de la naturaleza, y los demás modelos más sofisticados quedan deshechados. Este principio se conoce también como el principio de *parsimonia*.

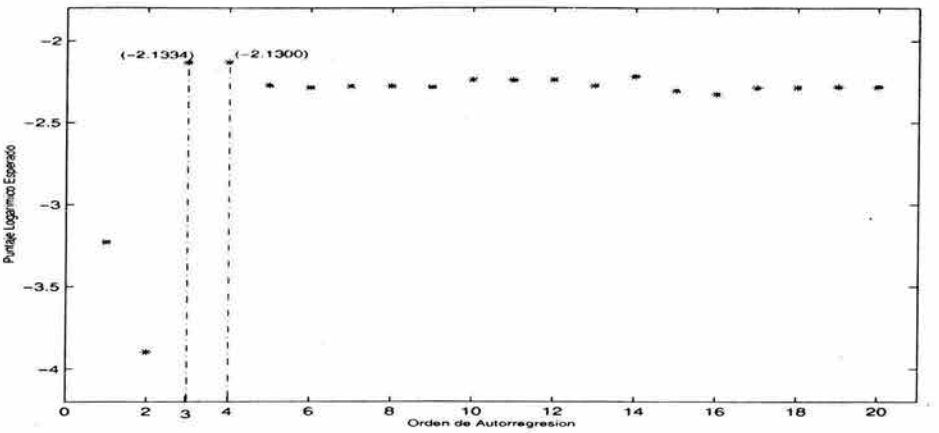
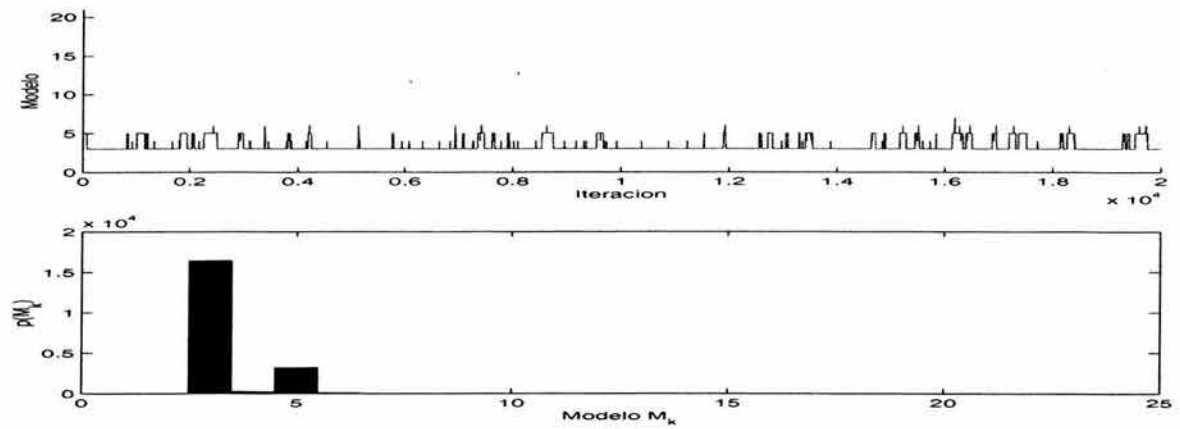
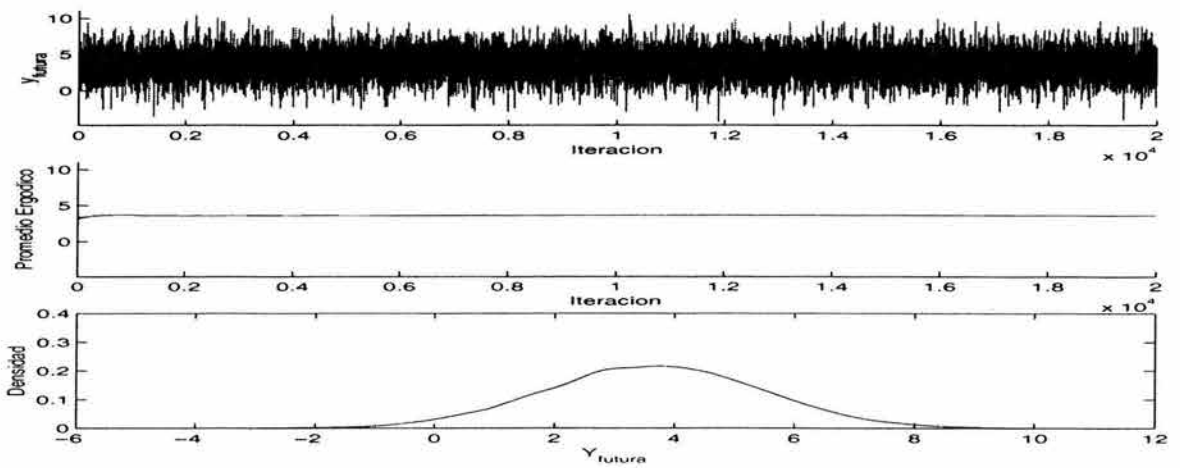


Figura 3.2: Puntaje logarítmico esperado para los modelos AR usando la densidad de referencia del modelo NAR(20,3,1) en la serie simulada.

El resultado de parsimonia a posteriori que encontramos en el criterio Bayesiano de máxima probabilidad no se extiende a los criterios predictivos que utilizamos para seleccionar los modelos. Para el caso $\mathcal{M}$-abierto, la similitud en términos de predicción entre los modelos AR(3) y AR(4) puede inducirnos a considerar la posibilidad de elegir de manera subjetiva uno de estos dos modelos, el más sencillo digamos, pues en términos de predicción son prácticamente indistintos. Es decir, a través de del puntaje logarítmico esperado encontramos una medida de 'cercanía' de la distribución predictiva de un modelo paramétrico respecto a la del modelo semiparamétrico que es interpretado como el modelo verdadero. Si tenemos evidencias cuantitativas de diferentes modelos que tienen el mismo comportamiento predictivo, o si no el mismo bastante semejante, y éstos representan las mejores alternativas para modelar el fenómeno de interés entonces en la búsqueda de simplificar el análisis estaríamos tentados por elegir el modelo más simple dentro de las 'mejores' alternativas. En el siguiente ejemplo real encontraremos un escenario en el que introducir un elemento subjetivo de parsimonia que mencionamos después de calcular las utilidades esperadas puede tener más sentido.



(a) Modelo



(b) Y_{futura}

Figura 3.3: Muestras de la mezcla Bayesiana de modelos AR(p), con $p = 1, \ldots, 20$.

AR(p)	AIC	BIC	$\mathcal{M}$-cerrada	$\mathcal{M}$-abierta
1	2.2946	2.3009	-2.9296	-3.2282
2	2.3233	2.3544	-3.4811	-3.8963
3	**1.6245**	**1.6808**	**-2.0279**	-2.1334
4	1.6602	1.7419	**-2.0284**	**-2.1300**
5	1.6322	1.7395	-2.0655	-2.2727
6	1.6601	1.7934	-2.0729	-2.2859
7	1.6980	1.8576	-2.0700	-2.2790
8	1.7291	1.9154	-2.0700	-2.2773
9	1.7519	1.9651	-2.0742	-2.2842
10	1.7914	2.0319	-2.0532	-2.2392

Cuadro 3.1: Criterios de selección de modelos AR para la serie simulada.

3.6.3 Serie Real

En este caso analizamos el Índice Nacional de Crecimiento del Nivel de Producción de la Industria Manufacturera[2] (INPIM) en México. Básicamente este índice mide el incremento porcentual de producción nacional en el sector manufacturero con base en el agregado ponderado nacional. La serie observada está formada por 260 observaciones comprendidas entre marzo de 1981 y octubre de 2002, en la figura 3.4 podemos observar su evolución. De nuevo, consideramos comparar e identificar el modelo AR, entre los órdenes $p_{\text{mín}} = 1$ y $p_{\text{máx}} = 25$, con el mejor poder predictivo para el tiempo $T + 1$, i.e. noviembre de 2002. Mediante una inspección gráfica inicial, observamos que la serie presenta un fuerte comportamiento estacional. Para realizar la comparación de los modelos mediante el criterio predictivo $\mathcal{M}$-abierta elegimos emplear el modelo NAR(25,5,2) como el modelo *flexible*, después de haber realizado una inspección predictiva del ajuste con base en las medidas descritas en el apéndice C. En la figura 3.5 presentamos las utilidades (puntajes logarítmicos) esperados de los modelos en la clase $\mathcal{M}$ con la perspectiva abierta. Como mencionamos antes, este criterio no penaliza la sobreparametrización (o complejidad) de

[2] Fuente: Indicadores Económicos, Banco de México. (`http://www.banxico.org.mx`).

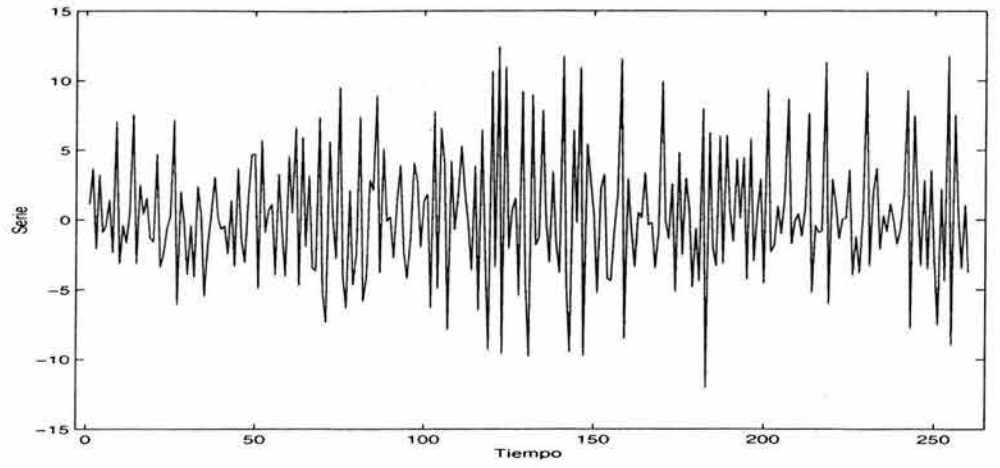


Figura 3.4: Índice Nacional de Crecimiento del Nivel de la Industria Manufacturera (INPIM) en México.

los modelos, y el modelo óptimo bajo este criterio es el AR(18), sin embargo existe una evidente similitud de otros modelos postulados respecto al modelo óptimo. Los modelos con orden de autorregresión mayor o igual a 12 presentan prácticamente el mismo comportamiento predictivo (vea la figura 3.5). Este un ejemplo donde modelos más complicados presentan el mismo poder predictivo que otro modelo más simples. Es notoria la cercanía de los modelos que tienen un orden de autorregresión mayor al del modelo óptimo.

En el cuadro 3.2 presentamos también los puntajes logarítmicos esperados considerando la perspectiva $\mathcal{M}$-cerrada. Para obtener una muestra simulada de la mezcla de densidades de la clase $\mathcal{A}$ empleamos el algoritmo de MCCMSR con los parámetros $c_1 = c_2 = 0.3$ y $c_3 = 0.4$, un periodo inicial de calentamiento de 3,000 iteraciones y un total de 10,000 observaciones. Los hiperparámetros que elegimos para las distribuciones iniciales en cada modelo M_k son un vector de medias nulo k-dimensional y una matriz de covarianzas $100\boldsymbol{I}_k$ para el vector $\boldsymbol{\phi}_k$, y $a = b = 0.001$ para la distribución inicial sobre τ_k. La traza de la evolución del índice del modelo (orden de autorregresión) y los valores predictivos al tiempo $T + 1$ se muestran en la figura 3.6.

De acuerdo con este criterio el modelo óptimo es el modelo AR(19), aunque de igual forma que con el criterio $\mathcal{M}$-abierta, la diferencia entre los modelos a partir del orden $p = 12$ es casi indistinguible (vea el cuadro 3.2). En el mismo cuadro presentamos los

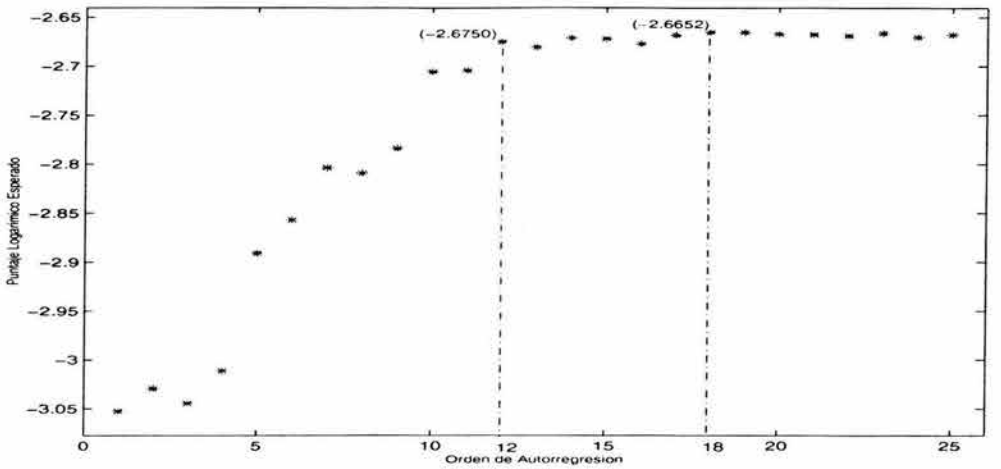


Figura 3.5: Utilidades esperadas de los modelos AR aplicados a la serie del Índice Nacional de Crecimiento del Nivel de la Industria Manufacturera (INPIM) en México.

puntajes usando los criterios AIC y BIC, en donde los modelos óptimos que se obtienen son el AR(16) y el AR(13) respectivamente. En estos criterios observamos que los modelos en una vecindad del modelo óptimo poseen puntajes semejantes, que en este caso están expresados en términos de la verosilimilitud penalizada de los modelos.

Bajo la perspectiva cerrada, si eligiéramos seleccionar el modelo con base en el criterio Bayesiano de máxima probabilidad, el modelo AR(13) resulta ser el óptimo (vea la figura 3.6(a)). Al igual que en el ejemplo anterior, en este criterio encontramos una resultado de parsimonia a posteriori.

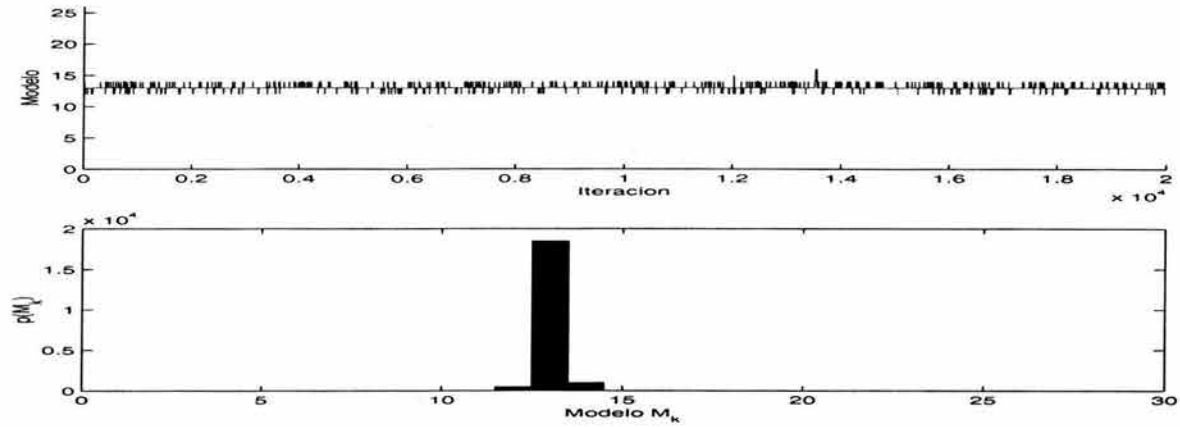
En torno al problema de selección de modelos bajo el enfoque predictivo, hemos encontrado que existe una colección amplia de modelos que generan predicciones muy similares.

Finalmente, en estos ejemplos hemos encontrado que los resultados de los criterios predictivos de selección son consistentes, con el modelo usado para generar la serie en el caso de la serie simulada y con los resultados obtenidos usando las dos perspectivas para la clase $\mathcal{M}$.

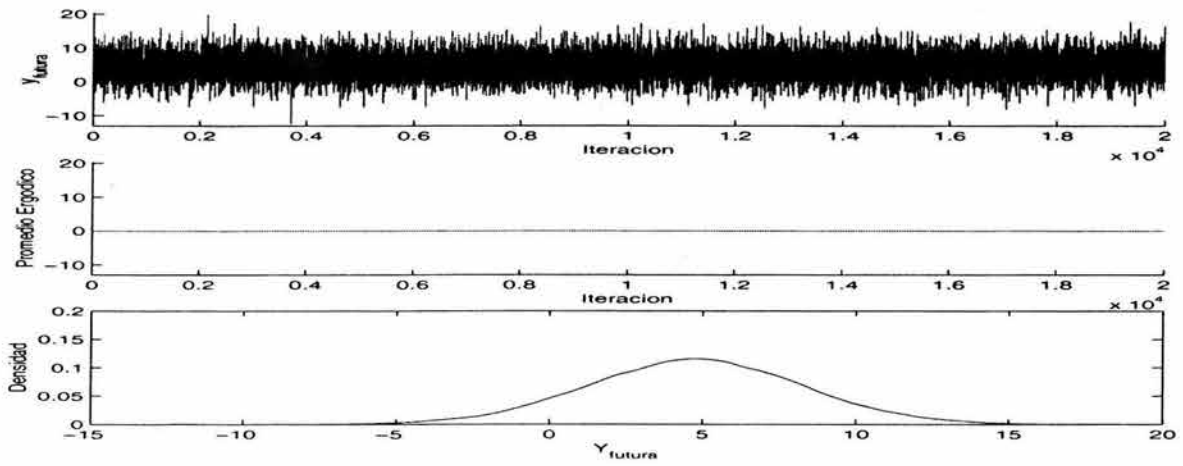
Las diferencias respecto a los resultados usando los criterios de AIC y BIC no son significativos, aún cuando éstos últimos contienen distintivamente un componente de penalización sobre la complejidad del modelo, entendida en este caso simplemente como la

Retraso	AIC	BIC	$\mathcal{M}$-cerrada (PBM)	$\mathcal{M}$-abierta (NAR)
1	2.8484	2.8544	-2.9390	-3.0525
2	2.8161	2.8359	-2.9173	-3.0296
3	2.8313	2.8649	-2.9299	-3.0446
4	2.8237	2.8713	-2.9020	-3.0115
5	2.8249	2.8865	-2.8087	-2.8909
6	2.8375	2.9132	-2.7843	-2.8570
7	2.8500	2.9398	-2.7486	-2.8035
8	2.8620	2.9662	-2.7522	-2.8090
9	2.8637	2.9821	-2.7360	-2.7839
10	2.8588	2.9917	-2.7040	-2.7055
11	2.8023	2.9496	-2.7521	-2.7042
12	2.6258	2.7877	-2.6797	-2.6750
13	2.5659	**2.7425**	-2.6738	-2.6807
14	2.5544	2.7458	-2.6767	-2.6713
15	2.5707	2.7769	-2.6763	-2.6721
16	**2.5361**	2.7572	-2.6735	-2.6771
17	2.5500	2.7862	-2.6794	-2.6681
18	2.5483	2.7996	-2.6922	**-2.6652**
19	2.5650	2.8314	**-2.6897**	**-2.6652**
20	2.5792	2.8610	-2.7030	-2.6669

Cuadro 3.2: Criterios de selección de modelos AR para la serie del INPIM.



(a) Modelo



(b) Y_{futura}

Figura 3.6: Muestras de la mezcla Bayesiana de modelos AR(p), con $p = 1, \ldots, 25$, para la serie del INPIM.

dimensión del espacio parametral asociado a cada modelo. La complejidad del modelo no está expresada en nuestro planteamiento del problema como un elemento del problema de decisión, sin embargo los datos reflejan que la contribución marginal que representa la incorporación de más información, i.e. mayores retrasos, es marginalmente insignificante en términos de la capacidad predictiva el modelo. Esto se ve reflejado en el puntaje logarítmico esperado final, que no presenta variaciones significativas a partir del modelo óptimo. Con esto podemos observar que aún cuando el criterio predictivo que consideramos no incorpora explícitamente un factor de penalización sobre el aumento o extensión de los modelos, que en este caso particular se ve reflejado como modelos con mayores retrasos, existe una determinación por los modelos mismos en la que la incorporación de información adicional no contribuye significativamente para obtener mejores predicciones de la serie, induciendo en este caso un resultado con parsimonia.

En este capítulo, para el caso muy particular de comparación de modelos AR lineales, empleamos un modelo en la clase de modelos AR no lineales (NAR) como el modelo flexible que toma la posición de juez de los modelos AR lineales postulados. Desde luego este modelo continua teniendo elementos restrictivos respecto a su forma estructural, y no puede ser considerada para comparar modelos con formas estructurales más variadas o sofisticadas pues no es lo deseablemente flexible. Para poder emplear el criterio predictivo de selección $\mathcal{M}$-abierta es necesario definir un modelo dentro de una clase de modelos que relajen lo más posible sus supuestos estructurales. En el siguiente capítulo presentamos dos clases de modelos semiparamétricos para series de tiempo donde los supuestos sobre sus formas estructurales son relajados para cubrir formas muy generales. La presentación que realizamos no es exhaustiva, ni tiene la pretensión de serlo, solamente nos enfocamos en dos clases de modelos Bayesianos semiparamétricos y no paramétricos que permiten definir modelos autorregresivos semiparamétricos flexibles. En éstos además es posible incorporar otras posibles variables exógenas para definir la estructura de dependencia condicional de la serie de interés.

Capítulo 4

Análisis Semiparamétrico de Series de Tiempo

4.1 Antecedentes

Al final del capítulo 3 presentamos el criterio de selección de modelos $\mathcal{M}$-semiabierta. Para implementar este criterio debe hacerse uso de un modelo *flexible*, en el cual se deben de tener menos supuestos estructurales que la gama de modelo paramétricos contendientes en el problema de selección. En ese mismo capítulo ejemplificamos este criterio para seleccionar modelos autorregresivos lineales considerando como modelo *flexible* a uno dentro de la clase de modelos autorregresivos no lineales, sin embargo la elección de este modelo flexible es *ad-hoc* y no es lo suficientemente general y lo suficientemente flexible como para poder incorporar a la clase de modelos contendientes otros con formas estructurales más generales. Recordemos que en primera instancia basta con que este modelos sea más flexible que los modelos paramétricos propuestos, sin embargo, debemos recordar que en esencia este modelo es la apuesta que el analista asume como la 'mejor' aproximación al "verdadero" modelo del estado de la naturaleza, por lo que preferentemente debemos especificar un modelo mucho más flexible que el que empleamos en el ejemplo citado.

Por otro lado sabemos que las aproximaciones más flexibles a ese modelo de la na-

turaleza pertenecen a la clase de modelos no paramétricos, pero para fenómenos como el que es de nuestro interés en este trabajo no conocemos que a la fecha exista un modelo lo suficientemente flexible cómo para ser considerado no paramétrico. Sin embargo, la mejor alternativa posible para nuestras capacidades consiste en definir un modelo semiparamétrico, en el que se combinen elementos paramétricos y no paramétricos simultáneamente.

Dentro del contexto de series de tiempo existe una extensa gama de modelos semiparamétricos que sirven para analizar diferentes características de interés de un proceso, e inclusive para realizar predicciones. Algunos de éstos se encuentran descritos en la revisión realizada por Härdle *et al.* (1997). En este trabajo nos concentraremos solamente en modelos enfocados en el ajuste y predicción de procesos con dominio en el tiempo. Para tal efecto, supongamos que un proceso temporal $\{Y_t\}$ puede ser representado mediante un proceso de localización y escala temporal soportado por un proceso de ruido, de la forma

$$Y_t = f(t) + g(t)\varepsilon_t, \tag{4.1}$$

donde $\{\varepsilon_t\}$ es un proceso de ruido aleatorio, y f y g son funciones reales desconocidas que caracterizan el nivel medio y la variabilidad del proceso al tiempo t respectivamente. Estas funciones pueden tener diferentes argumentos, y en modelos paramétricos tienen especificada previamente una forma estructural fija. Por ejemplo, en el caso de un modelo autorregresivo lineal $f(t) = f(t, \theta_1, \ldots, \theta_p, \boldsymbol{y}_t) = \sum_{j=1}^{p} \theta_j y_{t-j}$, donde $\boldsymbol{y}_t = (y_{t-1}, \ldots, y_{t-p})'$ denota los últimos p valores de la serie antes del tiempo t y $g(t) = \sigma$ es un valor constante desconocido pata todo t, donde $\sigma > 0$. De manera conjunta o individualmente, las funciones f y g denotan en conjunto al componente sistemático del proceso $\{Y_t\}$.

Otro componente relevante del modelo (4.1) se atribuye a la forma distribucional del proceso $\{\varepsilon_t\}$. Bajo el supuesto de que este proceso es ruido blanco, la sucesión de valores $\{\varepsilon_t\}$ es i.i.d. con una función de distribución F desconocida. En modelos paramétricos la forma funcional de F se restringe a una familia de distribuciones paramétricas previamente especificada, en cuyo caso nuestra incertidumbre se reduce al valor de los parámetros que la caracterizan. El modelo paramétrico más simple es el proceso de ruido blanco Gaussiano, donde F es una distribución Normal (Gaussiana) con media cero y varianza

unitaria. En modelos no paramétricos F es completamente desconocida y aleatoria (más adelante describimos cómo se puede realizar la inferencia más general sobre la distribución F mediante el enfoque Bayesiano). En términos generales consideramos que el proceso $\{\varepsilon_t\}$ representa el componente aleatorio del proceso de interés $\{Y_t\}$, y será modelado de acuerdo a la forma de su distribución.

En las siguientes subsecciones expondremos algunos modelos utilizados para modelar semiparamétricamente algún componente de un proceso con la forma (4.1). Como mencionamos, en nuestro conocimiento, aún no existe un modelo que sea completamente no paramétrico, i.e. donde para el caso de la representación (4.1) ambos componentes sean modelados de manera no paramétrica o semiparamétrica simultáneamente, o inclusive un modelo completamente no paramétrico que involucre la información de covariables como en el caso más simple de regresión.

4.2 Modelando el Componente Sistemático

En esta sección revisaremos algunos métodos de inferencia donde se modela semiparamétricamente el componente sistemático en la representación (4.1). Por simplicidad, consideramos que la función $g(t)$ es constante y positiva en t, restringiendo nuestro análisis semiparamétrico sobre el nivel medio o tendencia de la serie. Existen dos alternativas naturales para definir un estructura de dependencia en t de la función f. La primera consiste en suponer que f depende exclusivamente del tiempo t, o de una variable reescalada de ésta. Esta alternativa permite realizar un análisis flexible del problema, aunque puede conllevar ciertos problemas en cuanto a la predicción. Otra alternativa, que en la práctica resulta tener también una mejor interpretación, consiste en suponer que la función f depende de la trayectoria pasada de la serie observada hasta el tiempo t, i.e. $f(t) = f(y_{t-1}, y_{t-2}, ...)$, aunque en la práctica es preferible trabajar en términos de una dependencia local a través de la definición $f(t) = f(y_{t-1}, ..., y_{t-p})$, reduciendo nuestro problema al caso de un modelo autorregresivo lineal o de manera más general no lineal. Este es un procedimiento *ad-hoc* ya que es necesario definir el orden de autorregresión del modelo. En ciertas áreas como economía es deseable incorporar variables exógenas en

f correspondientes a realizaciones de otras series de tiempo de interés. Muchos de estos modelos pueden expresarse como un modelo de regresión, donde las covariables son las observaciones de un segmento de retraso inmediato del proceso. En adelante trabajaremos con esta notación más general, y después retomaremos el problema dentro del contexto de series de tiempo.

El problema de regresión semiparamétrica respecto al nivel medio puede expresarse en términos generales de la forma

$$y_t = f(\boldsymbol{x}_t) + \xi_t, \tag{4.2}$$

donde $f \in \mathcal{P}$ que es la clase de todas las posibles funciones de regresión, $\boldsymbol{x}_t$ son las variables explicativas (contiene una trayectoria pasada de la serie o variables exógenas), y ξ_t es la perturbación aleatoria i.i.d. con media cero y varianza $\gamma^2 > 0$ desconocida.

Así, el problema de regresión consiste en encontrar o aproximar la "verdadera" función de regresión f. Siendo capaces de encontrar una aproximación $\widehat{f}$ de f, el problema de regresión puede simplificarse de la forma

$$y_t = \widehat{f}(\boldsymbol{x}_t) + \varepsilon_t, \tag{4.3}$$

donde, a diferencia de (4.2), $\widehat{f}$ es una función conocida, que posiblemente depende de algunos parámetros desconocidos, y ε_t son variables aleatorias i.i.d. con media cero y varianza $\sigma^2 > 0$ desconocida.

Una forma de determinar la función $\widehat{f}$ se obtiene de suponer que la relación entre y_t y $\boldsymbol{x}_t$ es lineal, i.e. $\widehat{f}(\boldsymbol{x}_t) = \boldsymbol{\beta}'\boldsymbol{x}_t$ o para tener más precisión, que $\widehat{f}(\boldsymbol{x}_t) = \boldsymbol{\alpha}'h(\boldsymbol{x}_t)$ donde $h(\cdot)$ es una función polinomial que mapea de $\Re^p$ a $\Re^q$, con $\boldsymbol{\beta}$ y $\boldsymbol{\alpha}$ vectores de coeficientes de regresión de dimensión p y q respectivamente, e.g. en el ejemplo del capítulo 3 especificamos $h(\boldsymbol{x}_t)$ como un polinomio de segundo orden donde $\boldsymbol{x}_t$ fue considerado como un vector de retrasos del mismo proceso de interés. Bajo estos supuestos se obtiene la representación general del modelo de regresión lineal. Sin embargo este modelo es restrictivo en el sentido que se supone una forma funcional fija y en ocasiones poco flexible de $\widehat{f}$.

Si el problema central es el desconocimiento de la función f, podemos tomar una posición semiparamétrica y tratar de expresarla mediante una expansión de funciones bases de un espacio funcional conveniente, o inclusive definiendo una distribución inicial sobre ésta siguiendo el enfoque Bayesiano. Gutiérrez-Peña (1997) empleó este último enfoque definiendo un proceso Gaussiano sobre el espacio de las posibles funciones de regresión, aunque esta alternativa conlleva ciertas complicaciones respecto a la definición de los componentes del proceso Gaussiano. En este trabajo nosotros consideramos la alternativa de modelar semiparamétricamente a f mediante expansiones de ciertas funciones bases. Al respecto existe una gran variedad de alternativas, nosotros consideramos una en particular que permite modelar la función de regresión considerando interacciones entre las diferentes covariables del proceso en la cual se aproxima f mediante combinaciones lineales de superficies de respuesta definidas sobre el espacio generado por las covariables relacionadas al proceso. La alternativa que elegimos consiste en realizar esta aproximación mediante la expansión de funciones base radiales de f. Dentro de esta clase semiparamétrica muy particular se tiene dos enfoques para la construcción (definición) de dichas funciones bases. A continuación presentamos dos exposiciones distintas en las que se consideran funciones bases distintas.

4.2.1 Redes Neuronales

Este modelo consiste en aproximar la función f mediante la adición de K funciones base para el espacio de las funciones reales continuas definidas sobre $\Re^p$. La representación de (4.2) en este caso es

$$y_t = \sum_{j=1}^{K} \lambda_j \phi_j(\boldsymbol{x}_t) + \varepsilon_t, \tag{4.4}$$

donde $\{\lambda_j\}$ son un conjunto de coeficientes escalares, $\phi_j(\cdot)$ es una función base real conocida, y ε_t es un error aleatorio con media cero y varianza constante. Existen diferentes representaciones de modelos de la forma (4.4), por ejemplo: podemos utilizar funciones *kernel* o densidades (vea el apéndice B.3), *Projection Pursuit Regression*, donde $\phi_j(\boldsymbol{x}_t) = \psi_j(\boldsymbol{u}_j'\boldsymbol{x}_t + v_j)$, con $\boldsymbol{u}_j$ un vector p-dimesional de proyección de $\boldsymbol{x}_t$ en un hiperplano

indexado por j, v_j es un parámetro de desplazamiento y $\psi_j(\cdot)$ es una función real conocida, para cada $j = 1, ..., K$; *MARS* (*Multivariate Adaptive Regression Splines*) donde $\phi_j(\boldsymbol{x}_t) = \prod_{k=1}^{K_j} \psi_{jk}(x_{t\nu(j,k)})_+$, i.e. la j-ésima función base está dada por el producto de K_j funciones *splines*, las cuales dependen sólo de una de las variables explicativas $x_{t\nu(j,k)}$ cuyo índice es función de la función base y del correspondiente *spline* $\nu(j,k)$; entre otras posibles representaciones. De manera alternativa las funciones base se pueden definir de manera que cada una de estas dependa sólo de una variable del vector de covariables, en cuyo caso la representación del proceso (4.4) da origen a los modelo aditivos (Hastie y Tibshirani, 1992). Esta última alternativa es bastante viable, incluso a pesar de que no considera interacciones en las covariables, i.e. indirectamente supone independencia entre las covariables. A pesar de que este último supuesto puede ser considerado poco alentador, en la práctica existen resultados donde se tienen resultados más que aceptables con estos modelos.

Como ya mencionamos, nosotros elegimos modelar también de manera explícita la interacción de todas las covariables, construyendo las funciones base con estas interacciones. Definir estas interacciones en una sola función no es una tarea simple, por lo que adoptamos una forma bastante simple suponiendo cierta simetría en el comportamiento de las covariables de manera que cada función $\phi_j(\cdot)$ dependa de la distancia del vector $\boldsymbol{x}$ respecto a un cierto centroide $\boldsymbol{\mu}$, i.e. $\phi_j(\boldsymbol{x}_t) = \varphi_j(d(\boldsymbol{x}_t, \boldsymbol{\mu}_j))$ donde $d(\boldsymbol{x}_t, \boldsymbol{\mu}_j)$ denota una distancia de $\boldsymbol{x}_t$ respecto al centroide $\boldsymbol{\mu}_j$ para $j = 1, .., K$. Esta clase particular de bases se conocen como *bases radiales*, donde cada función ϕ da una noción de simetría respecto a cada centroide. La distancia puede definirse de diferentes formas, por ejemplo a través de la distancia inducida por la métrica Euclidiana donde $d(\boldsymbol{x}_t, \boldsymbol{\mu}_j) = \|\boldsymbol{x}_t - \boldsymbol{\mu}_j\| =$ que es la opción que consideramos en este trabajo. También se puede considerar la distancia de Mahalanobis donde $d(\boldsymbol{x}_t, \boldsymbol{\mu}_j) = (\boldsymbol{x}_t - \boldsymbol{\mu}_j)'\boldsymbol{\Sigma}_j(\boldsymbol{x}_t - \boldsymbol{\mu}_j)$, sin embargo, trabajar con esta distancia induciría posibles problemas de sobreparametrización del modelo.

El modelo (4.4) es una extensión del modelo propuesto por Powell (1987) para interpolar cualquier función multivariada mediante funciones base radiales sin considerar la perturbación aleatoria ε_t. En el modelo original Powell (1987) consideró que el número

$\varphi_1(u) = \exp\left\{-\frac{u^2}{2\sigma^2}\right\}$, con $\sigma^2 > 0$.
$\varphi_2(u) = u^2 \log u$.
$\varphi_3(u) = (u^2 + \sigma^2)^{-\alpha}$, con $\alpha > 0$.
$\varphi_4(u) = (u^2 + \sigma^2)^{\beta}$, con $0 < \beta < 1$.

Cuadro 4.1: Algunas funciones φ's utilizadas para construir funciones bases radiales.

de funciones radiales base es igual al número de datos observados, y los centros de las funciones base radiales corresponden a los puntos observados en la muestra, $\boldsymbol{x}_t$. Computacionalmente el modelo (4.4) es más sencillo que la propuesta original de Powell (1987), ya que el número de funciones base radiales, K, generalmente es menor que el tamaño de la muestra. Generalmente se supone que la forma funcional de la base radial $\phi_j(\cdot)$, a través de la especificación de $\varphi_j(\cdot)$ y de $d(\cdot,\cdot)$, es la misma para cada $j = 1, ..., K$. El modelo puede extenderse aún más si se considera un componente lineal adicional Holmes y Mallick (1999), de manera que la representación alternativa para (4.4) puede extenderse como

$$y_t = \sum_{j=1}^{K} \lambda_j \varphi_j(d(\boldsymbol{x}_t, \boldsymbol{\mu}_j)) + \boldsymbol{\alpha}'\boldsymbol{x}_t + \varepsilon_t, \tag{4.5}$$

donde $\boldsymbol{\alpha}$ es un vector p-dimensional de coeficientes desconocidos, con los demás componentes definidos como antes. La determinación de los centroides $\{\boldsymbol{\mu}_j\}$ es un problema delicado. Diferentes autores han sugerido determinarlos como una muestra aleatoria de los puntos observados, de acuerdo a la propuesta original de Powell (1987). La idea de este procedimiento es que la aproximación esté soportada dentro de la región de exploración. Este es un problema que aún está abierto.

En el cuadro 4.1 se enlistan algunas funciones utilizadas para construir bases radiales en la práctica y que tiene una alta sensibilidad de ajuste. Particularmente la función Gaussiana ha tenido una gran aceptación debido a que reproduce, bajo ciertas condiciones el estimador no paramétrico de Nadaraya-Watson del nivel medio de Y (Härdle *et al.*, 1997). Este modelo que brevemente hemos descrito puede ser visto también como un caso particular de los modelos de redes neuronales (e.g. Cheng y Titterington (1994)). En los modelos con funciones base radiales, cada una de estas funciones corresponde a un *nodo*

de la red neuronal.

Alternativamente podemos suponer que la función f pertenece a un espacio funcional particular, donde la construcción de las bases está dada por la teoría de onduletas. A continuación presentamos una breve introducción a esta teoría, la cual es una alternativa viable y flexible para encontrar bases de funciones donde ahora el único supuesto respecto a la forma funcional de f es que esta pertenezca al espacio de todas las funciones cuadrado integrables. Los resultados que empleamos corresponden básicamente a la transformada continua de onduletas, que fue la primera transformación de onduletas estudiada en la historia.

4.2.2 Onduletas

Las *onduletas* consisten básicamente en la descomposición y reconstrucción de funciones mediante combinaciones lineales de traslaciones y dilataciones (cambio de escala) de una función conocida como *onduleta.* El término de onduleta se refiere en un sentido coloquial a una onda pequeña que tiende a disiparse rápidamente. En el sentido matemático, una onduleta es una función que está definida básicamente en un intervalo compacto y que se desvanece rápidamente fuera de éste. Esta teoría de descomposición y reconstrucción de funciones en el espacio $L^2(\Re)$ (espacio de todas las funciones medibles cuadrado integrables definidas en $\Re$) se puede extender también a su contraparte discreta mediante el análisis de multirresolución.

Como mencionamos, la idea detrás de la teoría de onduletas consiste en la descomposición de una función $f \in L^2(\Re)$ mediante combinaciones lineales de traslaciones y dilataciones de una función, ψ, de manera que preferentemente exista una clase numerable formada por estas traslaciones y dilataciones que sea densa en el espacio $L^2(\Re)$, más adelante describimos una manera de construir tales bases. Suponiendo que podemos construir una base con estas características, la familia formada por todas las combinaciones lineales finitas de éstas también será densa en $L^2(\Re)$. Es precisamente aquí donde la teoría de onduletas, a través de la *transformada continua de onduletas* (TCO), nos ayuda a exhibir la función ψ que satisfaga esta condición. Para efectos prácticos de este tra-

bajo, sólo consideramos las funciones reales en el espacio $L^2(\Re)$, aunque la teoría también se desarrolla para funciones complejas (vea Daubechies (1992)). La construcción de la función ψ se obtiene de considerar una función $\psi : \Re \to \Re$ tal que satisfaga inicialmente las siguientes condiciones: i) $\int \psi(x)dx = 0$, y ii) $\int \psi^2(x)dx = 1$. La primera restricción da el sentido para interpretar a ψ como una función oscilante. Además de estas dos condiciones debe satisfacer la *condición de admisibilidad* (Daubechies, 1992; Vidakovic, 1999), que se define como

$$C_\psi = \int \frac{|\Psi(\omega)|^2}{|\omega|} d\omega < \infty, \tag{4.6}$$

donde $\Psi(\omega)$ denota la *transformada de Fourier* de la función $\psi(x)$, a su vez definida como $\int \psi(x)e^{-i\omega x}dx$. La función que satisface estas condiciones se conoce como *onduleta madre.* La clase formada por la traslación y dilatación de funciones que satisfacen esta condición es densa en $L^2(\Re)$, y garantiza la descomposición de la función f mediante combinaciones lineales de elementos en esta clase, de la forma

$$\psi_{a,b}(x) = |a|^{-1/2} \psi\left(\frac{x-b}{a}\right), \tag{4.7}$$

donde $a \in \Re \setminus \{0\}$ es el parámetro de dilatación y $b \in \Re$ es el parámetro de traslación.

La *transformada continua de onduletas* de la función f respecto a la onduleta $\psi_{a,b}$ se define como

$$\mathcal{W}_f(a,b) = \int f(x)\psi_{a,b}(x)dx. \tag{4.8}$$

Cuando ψ satisface la condición de admisibilidad (4.6), podemos reconstruir a f por medio de la *transformada continua inversa de onduletas* (TCIO). Esta relación es conocida como la *resolución de identidad*, con la que se tiene que la función f puede ser expresada como

$$f(x) = \frac{1}{C_\psi} \int_\Re \int_{\Re\setminus\{0\}} \mathcal{W}_f(a,b)a^{-1/2}\psi_{a,b}(x)dadb. \tag{4.9}$$

La clase formada por las traslaciones y dilataciones de la forma (4.7) es demasiado grande, para especificar una clase numerable con estas características se consideran los valores específicos de $a = 2^j$ y $b = k2^j$, para $j, k \in \mathbb{Z}$, en cuyo caso la clase formada por las combinaciones lineales de traslaciones y dilataciones de la onduleta de esta clase forma una base numerable del espacio $L^2(\Re)$, dada por

$$\left\{\psi_{j,k}(x) = 2^{-j/2}\psi(2^{-j}x - k), \text{ con } j,k \in \mathbb{Z}\right\}, \tag{4.10}$$

además forma una base ortonormal de $L^2(\Re)$. De esta forma, para cualquier función f en $L^2(\Re)$, se obtiene la representación discretizada de (4.9), a través de

$$f(x) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} c_{j,k}\psi_{j,k}(x), \tag{4.11}$$

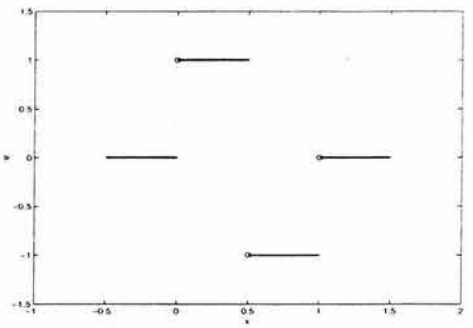
donde los coeficientes $c_{j,k}$ se definen a través de la relación (4.8). Adicionalmente se sabe que la clase formada por las combinaciones lineales finitas de (4.10) es densa en $L^2(\Re)$. En la figura 4.1 mostramos la forma de seis tipos de onduleta madre. La más simple es la onduleta de Harr, que se define sobre el intervalo $[0,1)$ como $\psi(x) = 1\ \mathbf{1}_{[0,1/2)}(x) + (-1)\ \mathbf{1}_{[1/2,1)}(x)$, aunque es deficiente entre otras cosas por no poder representar funciones constantes en el intervalo $[0,1)$. Esta deficiencia puede corregirse mediante la incorporación de una *onduleta padre* en la descomposición de la función a través del análisis de multirresolución, que describiremos más adelante. Otra onduleta de nuestro interés es la onduleta de Marr (figura 4.1(e)), que se define como la primera derivada de la distribución Gaussiana estándar (vea Vidakovic (1999)), i.e.

$$\psi(x) = (1 - x^2)\exp\{-x^2/2\}. \tag{4.12}$$

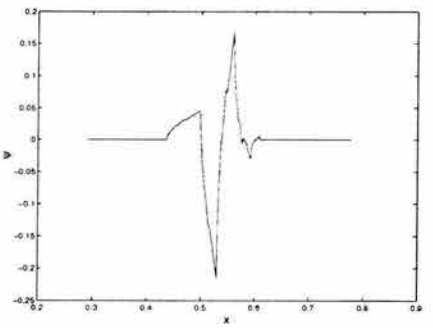
Las onduletas D(4,6,12) fueron propuestas por Daubechies (1992), donde podemos encontrar una completa descripción de ellas. Recordemos que nuestro problema central consiste en estimar estadísticamente una función de regresión f definida en $\Re^p$, con $p \geq 2$, por lo que describiremos la extensión de la descomposición de onduletas a la clase $L^2(\Re^p)$.

Caso Multivariado

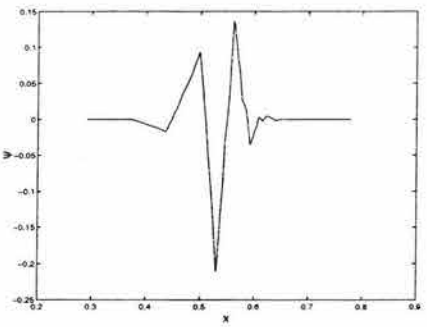
En la sección previa describimos descomposición de funciones $f \in L^2(\Re)$ mediante la transformada continua de onduletas. Ahora describiremos brevemente la extensión de esta teoría al espacio $L^2(\Re^p)$, donde buscamos exhibir una función $\psi : \Re^p \to \Re$ que satisfaga que en principio exista una familia numerable de funciones formadas por traslaciones y dilataciones de la misma que sea densa en $L^2(\Re^p)$. Zhang y Benveniste (1992) propusieron definir ψ como el producto de p onduletas univariadas, pertenecientes a la misma familia de onduletas, i.e. $\psi(\boldsymbol{x}) = \psi^*(x_1) \times \cdots \times \psi^*(x_p)$, donde $\psi^* : \Re \to \Re$ es una onduleta



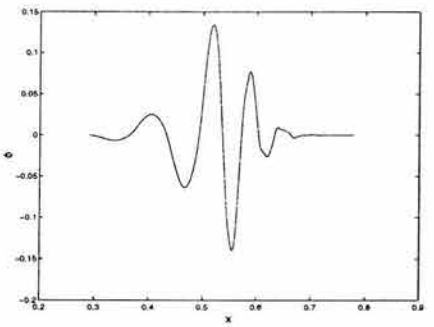
(a) Harr



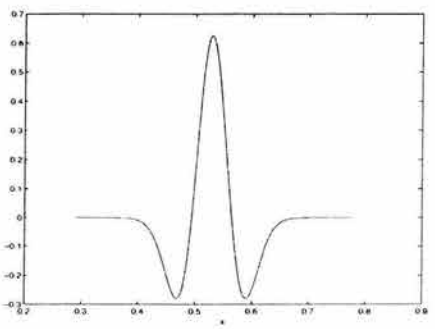
(b) D(4)



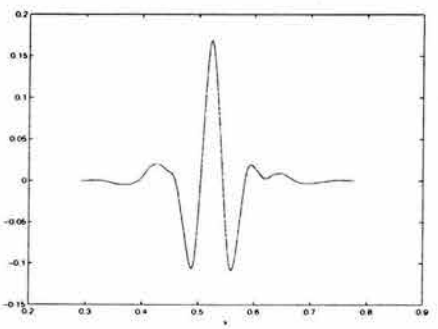
(c) D(6)



(d) D(12)



(e) Marr



(f) Symmlet(8)

Figura 4.1: Seis tipos de onduleta madre ψ.

que satisface las condiciones de admisibilidad (4.6). Esta es una extensión natural de la descomposición univariada de onduletas en la que existe la posibilidad de incorporar onduletas de diferentes familias preservando la condición de admisibilidad en $L^2(\Re^p)$ (vea Daubechies (1992)).

Por otro lado, la propuesta original de Zhang y Benveniste (1992) consistió en tomar una sola forma funcional de sólo una función $\psi : \Re^p \to \Re$, con traslaciones y dilataciones de la forma

$$\psi_{\boldsymbol{a},\boldsymbol{b}}(\boldsymbol{x}) = \det(\mathrm{diag}(\boldsymbol{a}))^{1/2}\psi\left(\mathrm{diag}(\boldsymbol{a})(\boldsymbol{x}-\boldsymbol{b})\right), \tag{4.13}$$

donde $\boldsymbol{a}$ es un vector real p-dimensional de dilatación con elementos estrictamente positivos, y $\boldsymbol{b}$ es un vector real de dimensión p que representa el vector de traslación de la onduleta. Zhang y Benveniste (1992) demostraron que la familia formada por las combinaciones lineales finitas de funciones de la forma (4.13) es densa en $L^2(\Re^p)$, en cuyo caso $C_\psi = (C_{\psi^\star})^p$ satisface la condición de admisibilidad. En esta representación es válido trabajar con vectores de dilatación cuyas entradas tengan el mismo valor escalar, que es la alternativa que consideramos en este trabajo.

Otra alternativa para construir la función ψ, utilizada por Zhang (1997), consiste en determinar una función que sea del tipo radial, como las que describimos en la sección anterior, pero que adicionalmente satisfaga la condición de admisibilidad

$$C_\psi = (2\pi)^p \int_0^\infty \frac{\eta(\omega)}{\omega} d\omega < \infty,$$

con $\eta(\cdot)$ una función real que satisface que $\eta(\omega) = \eta(\|\boldsymbol{\omega}\|) = \Psi(\boldsymbol{\omega})$, donde este último término denota la transformada de Fourier de ψ p-dimensional, a su vez definida como $\Psi(\boldsymbol{\omega}) = \int \cdots \int \psi(\boldsymbol{x}) e^{-i\boldsymbol{x}'\boldsymbol{\omega}} d\boldsymbol{x}$. Satisfaciendo esta condición, la transformada continua de onduletas de f respecto a la onduleta $\psi_{a,\boldsymbol{b}}$, para un parámetro de dilatación a estrictamente positivo y un vector de traslación p-dimensional, $\boldsymbol{b}$, se define como

$$\mathcal{W}_f(a,\boldsymbol{b}) = \int \cdots \int f(\boldsymbol{x}) a^{-p/2} \psi\left(\frac{1}{a}(\boldsymbol{x}-\boldsymbol{b})\right) d\boldsymbol{x},$$

y la reconstrucción de la función f se obtiene a partir de la transformada continua inversa de onduletas

$$f(\boldsymbol{x}) = \frac{1}{C_\psi} \int_0^\infty a^{-(p+1)} \int \cdots \int \mathcal{W}_f(a,\boldsymbol{b}) a^{-p/2} \psi\left(\frac{1}{a}(\boldsymbol{x}-\boldsymbol{b})\right) d\boldsymbol{b} da. \tag{4.14}$$

De esta forma se tiene, que la familia formada por las combinaciones lineales finitas de la forma

$$\sum_{i=1}^{K} w_i a_i^{-p/2} \psi\left(\frac{1}{a}(\boldsymbol{x}-\boldsymbol{b})\right) \tag{4.15}$$

es densa en $L^2(\Re^p)$. De nuevo, en la práctica, se deben estimar en conjunto los coeficientes lineales y los parámetros de traslación y dilatación. La onduleta radial que usó Zhang (1997) es una extensión de la onduleta de Marr para el espacio Euclidiano p-dimensional, dada por

$$\psi(\boldsymbol{x}) = (p - \|\boldsymbol{x}\|^2)\exp\{-\|\boldsymbol{x}\|^2/2\}, \tag{4.16}$$

donde $\|\boldsymbol{x}\|^2 = \boldsymbol{x}'\boldsymbol{x}$. Esta onduleta radial fue usada posteriormente por Holmes y Mallick (1999) para aproximar semiparamétricamente funciones de regresión con un enfoque Bayesiano. Con este modelo podemos aproximar funciones de regresión mediante combinaciones lineales de superficies de respuesta.

Por completez, presentamos una breve introducción al análisis de multirresolución propuesto por Mallat (una descripción completa se encuentra en Daubechies (1992)). A través de éste podemos considerar transformaciones discretas de onduletas, cuya teoría contribuye significativamente al análisis estadístico semiparamétrico (vea por ejemplo Vidakovic (1999)). Particularmente en el problema de regresión, podemos aproximar funciones de regresión mediante procedimientos como regresión por umbrales (*thresholding*) para funciones definidas en el intervalo [0,1), en el caso de series de tiempo y regresión podemos reindizar la serie en t de manera que los nuevos índices de la serie $\{y_s\}$ sea una sucesión finita en el intervalo [0,1). esta técnica es empleada para reproducir la serie de interés y eliminar el ruido que la afecta, sin embargo en nuestro caso no resulta de utilidad pues no es posible generar predicciones fuera de la región de exploración en s.

Análisis de Multirresolución

El concepto del *análisis de resolución* (AMR) fue introducido por Mallat para estudiar funciones en diferentes escalas. Este concepto es fundamental para el entendimiento de la *transformada discreta de onduletas* que describiremos más adelante.

El análisis de multirresolución se define como una sucesión de subespacios anidados $\{V_j : j \in \mathbb{Z}\}$ en $L^2(\Re)$ tal que satisface

i) $V_j \subset V_{j+1} \subset L^2(\Re)$

ii) $\bigcup_{j\in\mathbb{Z}} V_j$ es denso en $L^2(\Re)$

iii) $\bigcap_{j\in\mathbb{Z}} V_j = \{0\}$

iv) $f(x) \in V_j$ si y sólo si $f(2x) \in V_{j+1}$

v) $f(x) \in V_0$ si y sólo si $f(x-k) \in V_0$, con $k \in \mathbb{Z}$

vi) Existe una función única $\phi \in V_0$ tal que la sucesión $\{\phi(x-k) : k \in \mathbb{Z}\}$ forma una base ortonormal de V_0

Supondremos que $\phi \in V_0$ es tal que $\int \phi(x)dx \neq 0$. De las condiciones anteriores se sigue que $\{\phi_{1,k} : k \in \mathbb{Z}\}$ forma una base ortogonal de V_1, entonces

$$\phi(x) = \sum_{k=-\infty}^{\infty} h_k \phi_{1,k}(x), \tag{4.17}$$

donde $\boldsymbol{h} = \{h_k\}$ es conocido como *filtro de onduleta*, y la función ϕ es conocida como *onduleta padre* o *función de escalamiento*. El filtro $\boldsymbol{h}$ satisface que $\sum_{k=-\infty}^{\infty} h_k = \sqrt{2}$ y $\sum_{k=-\infty}^{\infty} h_k h_{k-2l} = \delta_l$, para todo $l \in \mathbb{Z}$, donde δ_l denota la función delta de Kronecker[1].

Cuando una sucesión de subespacios en $L^2(\Re)$ satisface un AMR, existe una base ortonormal en este espacio

$$\left\{\psi_{j,k}(x) = 2^{j/2}(2^j x - k) : j,k \in \mathbb{Z}\right\},$$

tal que

$$\left\{\psi_{j,k}(x) = 2^{j/2}(2^j x - k) : j \text{ fija y } k \in \mathbb{Z}\right\},$$

forma una base ortonormal del *espacio de detalle* W_j definido como el complemento ortogonal de V_j en V_{j+1}[2]. La función $\psi = \psi_{0,0}$ es conocida como la *onduleta madre* o *función de onduleta*. Ésta puede derivarse a partir de ϕ como

$$\psi(x) = \sum_{k=-\infty}^{\infty} g_k \phi_{1,k}(x),$$

[1] $\delta_l = 1$ si $l = 0$ ó 0 si $l \neq 0$, i.e. satisface la condición de ortonormalidad.

[2] La sucesión de espacios de detalle $\{W_j\}$ satisfacen: *i)* $W_j$es ortogonal con V_j, *ii)* $W_j$es ortogonal con $W_{j'}$ si $j \neq j'$, y *iii)* $L^2(\Re)$ es la suma directa de $\{W_j\}$.

en donde $\boldsymbol{g} = \{g_k : g_k = (-1)^k h_{1-k}\}$, que define el filtro conjugado de $\boldsymbol{h}$.

En general, una función f en $L^2(\Re)$ puede aproximarse mediante combinaciones lineales de $\phi_{j,k}$ y $\psi_{j,k}$ como

$$f(x) \approx \sum_{k=-\infty}^{\infty} c_{0,k}\phi_{0,k}(x) + \sum_{j=0}^{J-1}\sum_{k=-\infty}^{\infty} d_{j,k}\psi_{j,k}(x), \tag{4.18}$$

donde J denota el número de niveles en la aproximación, y k el índice sobre el número de componentes en cada nivel. Los términos $\{c_{0,k}\}$ denotan los coeficientes de la onduleta padre y $\{d_{j,k}\}$ denotan los coeficientes de onduletas en el nivel j-ésimo.

4.2.3 Regresión semiparamétrica vía onduletas o redes neuronales

En esta sección revisaremos un modelo semiparamétrico basado en onduletas (o bases radiales), que permite incorporar algunos retrasos de la serie, así como la información adicional sobre variables o series de tiempo exógenas. El modelo que describiremos a continuación puede ser visto como un caso particular de los modelos de regresión basados en redes neuronales (vea la sección 4.2.1), con la diferencia que en este caso existe una posible justificación sustentada en la expansión truncada en onduletas de la función de interés, para funciones definidas en el espacio $\Re^p$ para alguna $p > 1$. En este caso las predicciones sobre la variable de interés se obtienen de manera directa con dominio en el tiempo.

Supongamos que tenemos una variable aleatoria Y, y un conjunto de posibles variables explicativas $\boldsymbol{x} = (x_1, ..., x_p)'$. Para un conjunto de datos, suponemos que el nivel medio de Y está determinado por el modelo de regresión

$$y_t = f(\boldsymbol{x}_t) + \xi_t, \tag{4.19}$$

donde $\{\xi_t\}$ es una sucesión de ruido blanco con media cero y varianza $\gamma^2 > 0$ desconocida, y $f(\cdot)$ es una función real desconocida definida sobre $\Re^p$. Como mencionamos, en modelos paramétricos de regresión se considera que la forma funcional de f es conocida, ejerciendo supuestos restrictivos sobre ella. En los modelos semiparamétricos de regresión

los supuestos sobre f son más generales y menos restrictivos, considerando a f dentro de una familia funcional general, por ejemplo las funciones reales continuas. Nosotros supondremos que f pertenece al espacio $L^2(\Re^p)$ y que podemos aproximarla mediante una expansión truncada de onduletas tal que, eliminando los índices de la expansión, es de la forma

$$f(\boldsymbol{x}) \approx \sum_{j=1}^{J} \omega_j \psi_j(\boldsymbol{x}), \tag{4.20}$$

donde $\psi(\cdot)$ es una función de onduletas, que en este caso consideremos radial, y el índice j indica la traslación y dilatación de ψ. En la práctica, nosotros desconocemos los coeficientes de esta expansión, por lo que debemos estimarlas para un número fijo de bases. Al considerar onduletas (o bases) radiales, si definimos el vector de regresión definido por trayectorias pasadas finitas de las serie, estamos definiendo un modelo autorregresivo semiparamétrico *flexible*. También desconocemos los coeficientes de traslación y dilatación, así que debemos estimarlos en conjunto con los coeficientes de la expansión lineal. La determinación de los parámetros de traslación puede hacerse de la manera tradicional, definiendo una muestra aleatoria de los vectores observados en la muestra, como en el caso de los modelos de bases radiales usuales. La especificación de los parámetros de dilatación es un poco más complicada, más adelante presentaremos una alternativa propuesta por Holmes y Mallick (2000).

La función de regresión f también puede aproximarse empleando bases radiales, como las mostradas en la tabla 4.1, aunque en este caso sólo suponemos que la función de regresión f pertenece al espacio de las funciones continuas en $\Re^p$.

Análisis Bayesiano

A continuación describimos el modelo Bayesiano propuesto por Holmes y Mallick (2000) basado en esta aproximación a f, con el que es posible obtener predicciones o aproximar la densidad predictiva de Y de manera simple. El modelo que analizaremos está descrito como

$$y_t = \hat{f}(\boldsymbol{x}_t) + \varepsilon_t, \tag{4.21}$$

donde

$$\hat{f}(\boldsymbol{x}_t) = \alpha_{k,0} + \boldsymbol{x}_t'\boldsymbol{\beta}_{k,1} + \sum_{j=1}^{k} \omega_{k,j}\psi_j(\boldsymbol{x}_t),$$

con k el número de bases de onduletas consideradas en el modelo, y

$$\psi_j(\boldsymbol{x}_t) = \varphi\left(D_{k,j}\left\|\boldsymbol{x}_t - \boldsymbol{\mu}_{k,j}\right\|\right),$$

para ciertos valores de $D_{k,j} > 0$ y $\boldsymbol{\mu}_{k,j} \in \Re^p$ que representan los parámetros de dilatación y traslación de la j-ésima onduleta (o base radial). En el modelo los coeficientes $\alpha_{k,0}$ y $\boldsymbol{\beta}_{k,1} \in \Re^p$ y $\omega_{k,j}$, para $j = 1, ..., k$, son desconocidos. Supondremos que $\{\epsilon_t\}$ es una sucesión de ruido blanco con distribución $N(\epsilon_t|0, \tau^{-1})$, donde $\tau = 1/\sigma^2$ denota la precisión desconocida de los errores aleatorios. Nuestro objetivo es obtener la densidad predictiva de y_{T+1} con base en los datos observados $\{\boldsymbol{y}_T, \boldsymbol{x}_1, ..., \boldsymbol{x}_T\}$ y $\boldsymbol{x}_{T+1}$, con $\boldsymbol{y}_T = (y_1, ..., y_T)'$. El componente lineal en las variables $\boldsymbol{x}_t$ en el modelo es opcional. Su incorporación busca una representación parsimoniosa y un mejor ajuste del modelo.

En la definición de (4.21) existe una gran cantidad de elementos de incertidumbre. El primer elemento de incertidumbre se relaciona con la aproximación de la función de regresión f. Denotemos por

$$I(k) = \left\{k, \left(D_{k,j}, \boldsymbol{\mu}_{k,j}\right), j = 1, ..., k\right\}, \tag{4.22}$$

a los elementos que determinan la estructura de la aproximación a f por medio de onduletas (o bases radiales), i.e. el número de bases y sus correspondientes parámetros de dilatación y traslación. En esta caso, consideramos que la forma funcional de ψ es determinada de manera subjetiva al inicio del análisis. Holmes y Mallick (2000), haciendo referencia los trabajos de Zhang y Benveniste (1992) y Zhang (1997), utilizan la onduleta de Marr p-dimensional. Habiendo determinado la onduleta por usar, la calidad de la aproximación a f depende del número de bases de onduleta usadas y de los parámetros de dilatación y traslación. Conforme aumentemos el número de bases de onduletas, la aproximación tenderá a ser mejor, sin embargo existe una restricción de recursos computacionales que restringe el número de bases disponibles para realizar la aproximación. En este caso es esencial modelar nuestra incertidumbre sobre el número de

bases y sus parámetros de caracterización. Modelar esta incertidumbre de manera separada implicaría ciertos problemas analíticos para determinar las distribuciones finales de estos elementos, debido a que tenemos en este caso una estructura notoriamente no lineal en los parámetros. Holmes y Mallick (2000) propusieron modelar de manera conjunta la incertidumbre sobre el número de onduletas, sus centroides y parámetros de dilatación, i.e. $I(k)$, sobre el cual ahondaremos más adelante.

Condicional en $I(k)$, el modelo (4.21) corresponde al caso especial de un modelo de regresión lienal, con la representación matricial

$$\boldsymbol{y}_T = \boldsymbol{B}_k \boldsymbol{\omega}_k + \boldsymbol{\varepsilon}, \tag{4.23}$$

donde

$$\boldsymbol{B}_k = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,p} & \psi\left(D_{k,1}\left\|\boldsymbol{x}_1 - \boldsymbol{\mu}_{k,1}\right\|\right) & \dots & \psi\left(D_{k,k}\left\|\boldsymbol{x}_1 - \boldsymbol{\mu}_{k,k}\right\|\right) \\ 1 & x_{2,1} & \dots & x_{2,p} & \psi\left(D_{k,1}\left\|\boldsymbol{x}_2 - \boldsymbol{\mu}_{k,1}\right\|\right) & \dots & \psi\left(D_{k,k}\left\|\boldsymbol{x}_2 - \boldsymbol{\mu}_{k,k}\right\|\right) \\ 1 & x_{3,1} & \dots & x_{3,p} & \psi\left(D_{k,1}\left\|\boldsymbol{x}_3 - \boldsymbol{\mu}_{k,1}\right\|\right) & \dots & \psi\left(D_{k,k}\left\|\boldsymbol{x}_3 - \boldsymbol{\mu}_{k,k}\right\|\right) \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & \dots & x_{n,p} & \psi\left(D_{k,1}\left\|\boldsymbol{x}_n - \boldsymbol{\mu}_{k,1}\right\|\right) & \dots & \psi\left(D_{k,k}\left\|\boldsymbol{x}_n - \boldsymbol{\mu}_{k,k}\right\|\right) \end{pmatrix}, \tag{4.24}$$

$$\boldsymbol{\omega}_k = \left(\alpha_{k,0}, \beta_{k,1}, ..., \beta_{k,p}, \omega_{k,1}, ..., \omega_{k,k}\right)',$$

y

$$\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_T)',$$

con $\boldsymbol{\varepsilon} \sim N_T\left(\boldsymbol{\varepsilon}|\mathbf{0}, \tau^{-1}\boldsymbol{I}_T\right)$.

Aprovechando la estructura lineal del modelo, condicional en $I(k)$, podemos definir una estructura jerarquizada de (4.21) como

$$p\left(k, \left\{(D_{k,j}, \boldsymbol{\mu}_{k,j}), j = 1, .., k\right\}, \boldsymbol{\omega}_k, \tau\right) = p(I(k))p(\boldsymbol{\omega}_k, \tau|I(k)). \tag{4.25}$$

Eliminando de la notación la dependencia en los datos observados hasta T, podemos obtener la densidad predictiva final de y_{T+1} como

$$\begin{aligned} p(y_{T+1}|\boldsymbol{x}_{T+1}) \;=\; & \sum_k \int\int\int\int p(y_{T+1}|\boldsymbol{x}_{T+1}, I(k), \boldsymbol{\omega}_k, \tau)p(\boldsymbol{\omega}_k, \tau|I(k)) \\ & \times p(\boldsymbol{D}_k, \boldsymbol{M}_k|k)d\boldsymbol{\omega}_k d\tau d\boldsymbol{D}_k d\boldsymbol{M}_k p(k), \end{aligned} \tag{4.26}$$

donde

$$\begin{aligned} \boldsymbol{D}_k &= (D_{k,1}, ..., D_{k,k}), \\ \boldsymbol{M}_k &= (\boldsymbol{\mu}_{k,1}, ..., \boldsymbol{\mu}_{k,k}), \\ p(y_{T+1}|\boldsymbol{x}_{T+1}, I(k), \boldsymbol{\omega}_k, \tau) &= N(y_{T+1}|\boldsymbol{B}_k(\boldsymbol{x}_{T+1})\boldsymbol{\omega}_k, \tau^{-1}), \end{aligned}$$

con

$$\boldsymbol{B}_k(\boldsymbol{x}_{T+1}) = (1, x_{T+1,1}, ..., x_{T+1,p}, \psi_1(\boldsymbol{x}_{T+1}), ..., \psi_k(\boldsymbol{x}_{T+1})).$$

Es imposible resolver analíticamente la integral (4.26). Su solución depende de integrales iteradas de diferentes dimensiones. Inclusive su solución numérica no es una tarea trivial. Sin embargo, podemos aproximarla usando el algoritmo de MCCMSR (vea la sección 2.6), utilizando el algoritmo propuesto por Holmes y Mallick (2000). Como ya mencionamos, podemos caracterizar cada modelo mediante $I(k)$, reduciendo así de manera significativa el problema de determinar distribuciones iniciales sobre los parámetros $\{(D_{j,k}, \boldsymbol{\mu}_{j,k}), j = 1, ..., k\}$ de manera independiente.

Condicional en $I(k)$ y usando la representación (4.23) del modelo bajo el supuesto de normalidad sobre $\boldsymbol{\epsilon}$, podemos asignar una distribución inicial conjugada Normal/Gamma para $(\boldsymbol{\omega}_k, \tau)$, de la forma (Bernardo y Smith, 1994)

$$\begin{aligned} p(\boldsymbol{\omega}_k, \tau|I(k)) &= p(\boldsymbol{\omega}_k|I(k), \tau)p(\tau|I(k)) \\ &= NGa(\boldsymbol{\omega}_k, \tau|\delta_1, \delta_2, \mathbf{0}, \lambda^{-1}\boldsymbol{I}_{k+p+1}), \end{aligned} \tag{4.27}$$

con λ la precisión de los coeficientes de regresión, $\mathbf{0}$ el vector nulo de dimensión $(k + n + 1) \times 1$, y δ_1 y δ_2 constantes conocidas mayores a cero.

Como mencionamos, Holmes y Mallick (2000) propusieron modelar de manera agrupada los componentes de $I(k)$. Para este efecto, se concentraron en la capacidad de ajuste del modelos a los datos observados. Con el propósito de hacer comparable esta característica entre diferentes especificaciones de $I(k)$, Holmes y Mallick (2000) sugieren utilizar una medida global de ajuste del modelo en términos de $\hat{\boldsymbol{f}} = \boldsymbol{S}_k\boldsymbol{y}_T$, con $\boldsymbol{S}_k = \boldsymbol{B}_k(\boldsymbol{B}_k'\boldsymbol{B}_k + \boldsymbol{\lambda}\boldsymbol{I}_{k+p+1})^{-1}\boldsymbol{B}_k$, donde $\boldsymbol{B}_k$ está definida como antes y λ es el hiperparámetro de precisión de la distribución inicial sobre los coeficientes de regresión. La

medida para determinar el grado de ajuste del modelo que usaron Holmes y Mallick (2000) corresponde a los *grados de libertad* (g.l.) del modelo, que en este caso se puede definir como la traza de la matriz de suavizamiento $\boldsymbol{S}_k$, i.e. $g.l.(I(k)) = tr(\boldsymbol{S}_k)$, que es la suma de los eigenvalores de la matriz $\boldsymbol{S}_k$. Existen otras definiciones respecto a los grados de libertad de un modelo, vea por ejemplo Hastie y Tibshirani (1992). Intuitivamente los grados de libertad pueden interpretarse como el número efectivo de parámetros de un modelo, esta interpretación es motivada por su analogía con el modelo de regresión lineal. La idea de esta medida es que conforme el ajuste del modelos sea mejor los grados de libertad se reducirán.

Para modelar la incertidumbre respecto a los grados de libertad, Holmes y Mallick (2000) propusieron asignar una distribución inicial $Gamma(\gamma_1, \gamma_2)$, de manera que la distribución tenga una esperanza relativamente pequeña para reflejar nuestra impresión de que el modelo tiene una buena capacidad de ajuste.

Condicional en $I(k)$, se tiene que la distribución final para $\boldsymbol{\omega}_k$ y τ_k es (vea el apéndice B.1):

$$p(\boldsymbol{\omega}_k, \tau | I(k), \boldsymbol{y}) = N_{k+p+1}\left(\boldsymbol{\omega}_k | \boldsymbol{m}_k, \tau^{-1}\boldsymbol{V}_k\right) Ga\left(\tau | \delta_1 + n/2, \delta_2 + 1/2(\boldsymbol{y}'\boldsymbol{y} - \boldsymbol{m}_k'\boldsymbol{V}_k^{-1}\boldsymbol{m}_k)\right), \tag{4.28}$$

con

$$\begin{aligned} \boldsymbol{V}_k &= (\boldsymbol{B}_k'\boldsymbol{B}_k + \lambda \boldsymbol{I}_{k+p+1})^{-1}, \\ \boldsymbol{m}_k &= \boldsymbol{V}_k\boldsymbol{B}_k'\boldsymbol{y}, . \end{aligned}$$

Dado que condicionalmente en $I(k)$ podemos encontrar la distribución final de $(\boldsymbol{\omega}_k, \tau)$, entonces podemos utilizarla como la distribución de transición de la cadena en el espacio de los parámetros. En esta caso la probabilidad de aceptación de la transición del modelo $M_{I(k)}$ al modelo $M_{I(k')}$ estará dada por

$$\alpha = \text{mín}(1, Q), \tag{4.29}$$

con

$$\begin{aligned} Q &= \frac{p(\boldsymbol{y}|I(k'))}{p(\boldsymbol{y}|I(k))} \times \frac{p(I(k'))}{p(I(k))} \times \frac{J(M_{I(k')}, M_{I(k)})}{J(M_{I(k)}, M_{I(k')})} \\ &= \underset{\text{Bayes}}{\text{Factor de}} \times \underset{\text{Prob. Iniciales}}{\text{Cociente de}} \times \underset{\text{Transición}}{\text{Cociente de Prob.}} , \end{aligned} \tag{4.30}$$

donde $J(M', M)$ denota la probabilidad de transición del modelo M' al modelo M.

El factor de Bayes en (4.30) tiene una solución analítica dada por (vea el apéndice B.1):

$$FB = \frac{|\boldsymbol{V}_{k'}|^{1/2} |\lambda^{-1}\boldsymbol{I}_{k+p+1}|^{1/2}}{|\boldsymbol{V}_{k}|^{1/2} |\lambda^{-1}\boldsymbol{I}_{k'+p+1}|^{1/2}} \left(\frac{\delta_{2,k'}}{\delta_{2,k}}\right)^{\delta_1+n/2} , \tag{4.31}$$

donde $\delta_{2,k} = \delta_2 + 1/2(\boldsymbol{y}'\boldsymbol{y} - \boldsymbol{m}_k'\boldsymbol{V}_k\boldsymbol{m}_k)$, y n es el número de datos observados

El cociente de probabilidades iniciales se obtiene de manera directa mediante el cociente de la densidad evaluada en la traza de la matriz de ajuste de cada modelo. Se consideran cuatro posibles movimientos en la evolución de la cadena. Dos de ellos se relacionan con un cambio de dimensión del espacio parametral. El movimiento denotado por NACIMIENTO consiste en añadir una base de onduletas caracterizada con diferentes parámetros de traslación y dilatación. El movimiento denotado por MUERTE consiste en la eliminación aleatoria de una base de onduletas dentro de las bases empleadas en la iteración anterior. Se consideran además dos movimientos adicionales que no implican un cambio en la dimensión del espacio parametral. El movimiento DILATACIÓN consiste en alterar aleatoriamente un parámetro de dilatación de una base de onduleta existente en la iteración anterior. Holmes y Mallick (2000) propusieron definir esta modificación a través de la relación

$$D'_{k,i} = \exp\{\log(D_{k,i}) + \eta\} ,$$

para un $\eta \sim N(\eta|0, \sigma_\eta^2)$. Alternativamente podemos pensar en alterar el coeficiente de dilatación seleccionado con la adición de una variable aleatoria con distribución centrada en cero y truncada en $-D_{k,i}$, por ejemplo una Normal. El movimiento TRASLACIÓN consiste en la sustitución del parámetro de traslación de una base de onduletas, seleccionada aleatoriamente.

Dada la muestra de la distribución final de los modelos y parámetros asociados extraída de la cadena antes descrita, $\{k^{(l)}, (\boldsymbol{\omega}_{k^{(l)}}^{(l)}, \tau^{(l)}), l = 1, ..., N\}$, considerando que éstos se encuentran en el periodo estacionario de la cadena, y habiendo eliminado posibles efectos de correlación entre ellos, podemos aproximar (4.26) mediante

$$\begin{aligned} \hat{p}(y_{T+1}|\boldsymbol{x}_{T+1}, \boldsymbol{y}_T) &= \frac{1}{N}\sum_{l=1}^{N} p(y_{T+1}|\boldsymbol{x}_{T+1}, I(k^{(l)}), \boldsymbol{\omega}^{(l)}k^{(l)}, \tau^{(l)}) \\ &= \frac{1}{N}\sum_{l=1}^{N} N(y_{T+1}|\boldsymbol{B}_{k^{(l)}}(\boldsymbol{x}_{T+1})\boldsymbol{\omega}_{k^{(l)}}^{(l)}, (\tau^{(l)})^{-1}). \end{aligned} \tag{4.32}$$

A continuación presentamos de manera esquematizada el algoritmo de MCCMSR para el modelo de onduletas radiales. Este esquema puede implementarse también en funciones base radiales y no sólo de onduletas radiales simplemente eliminando los pasos concernientes a la actualización de los parámetros de dilatación, y preservando los pasos relacionados con la actualización de los parámetros de localización.

Algoritmo 2: MCCMSR - Redes Neuronales de Onduletas y Bases Radiales

Determinamos un valor inicial de k, y definimos los valores iniciales para los coeficientes de traslación y dilatación de las bases radiales o de onduletas, $\{(D_{j,k}, \boldsymbol{\mu}_{j,k}), j = 1, ..., k\}$.

1. Determinamos una muestra de $\tau \sim Ga(\tau|\delta_1, \delta_2)$, y otra de $\boldsymbol{\omega}_k \sim N_{k+p+1}(\boldsymbol{\omega}_k|\boldsymbol{m}_k, \tau^{-1}\boldsymbol{V}_k)$.

2. Iniciamos la cadena de Markov con salto reversible

   (a) Proponemos el tipo de movimiento de la cadena $I(k')$

      i. NACIMIENTO con probabilidad a (si el movimiento es admisible), $k' = k + 1$

      ii. MUERTE con probabilidad b (si el movimiento es admisible), $k' = k - 1$

      iii. DILATACIÓN con probabilidad c

      iv. TRASLACIÓN con probabilidad d

   (b) Obtenemos una muestra actualizada de $\boldsymbol{\omega}_{k'} \sim N_{k'+p+1}(\boldsymbol{\omega}_{k'}|\boldsymbol{m}_{k'}, \tau^{-1}\boldsymbol{V}_{k'})$

   (c) Evaluamos la probabilidad de aceptación del modelo. Generamos $u \sim U(0, 1)$

      i. Aceptamos el movimiento si $u < Q$

ii. Rechazamos en caso contrario

(d) Obtenemos una muestra de τ a través de la condicional completa en el movimiento aceptado $Ga\left(\tau|\delta_1 + n/2, \delta_2 + 1/2(\boldsymbol{y}'\boldsymbol{y} - \boldsymbol{m}_k'\boldsymbol{V}_k^{-1}\boldsymbol{m}_k)\right)$

3. Obtenemos una muestra de la densidad predictiva un paso adelante al tiempo T, $y_{T+1} \sim N(y_{T+1}|B_k(\boldsymbol{x}_{T+1})\boldsymbol{\omega}_k, \tau^{-1})$

4. Continuamos hasta obtener la convergencia de la cadena

En la siguiente sección describimos un modelo Bayesiano que permite modelar el componente aleatorio del modelo de manera semiparamétrica, a través de la incorporación de un método Bayesiano no paramétrico de inferencia sobre la distribución del ruido aleatorio.

4.3 Modelando el Componente Aleatorio

En la sección anterior describimos una manera de modelar semiparamétricamente el componente sistemático de un proceso, en esta sección revisaremos la forma de modelar semiparamétricamente el componente aleatorio de un proceso de la forma (4.1). Antes de continuar, realizaremos una breve introducción al problema de inferencia Bayesiana no paramétrica, donde básicamente el problema consiste en asignar una medida de probabilidad sobre el conjunto de todas las funciones de distribución definidas sobre un espacio de interés. Con este propósito, nos concentraremos en la descripción de la clase conocida como árboles de Pólya (Ferguson, 1974; Mauldin *et al.*, 1992; Lavine, 1992, 1994). La extensión a nuestro problema particular es relativamente simple, como se muestra en Gelfand (1998) para el caso particular de regresión.

4.3.1 Inferencia Bayesiana no Paramétrica

En esta sección discutiremos algunas alternativas para modelar el componente aleatorio de manera semiparamétrica considerando que el componente sistemático del modelo está previamente determinado de manera completamente paramétrica. Esta alternativa de modelación es relativamente reciente, y corresponde al caso general de inferencia Bayesiana no

paramétrica para variables aleatorias independientes. La descripción que hacemos en este trabajo se basa en el trabajo de Ferguson (1974) y Lavine (1992, 1994).

Para introducir el problema de inferencia Bayesiana no paramétrica es pertinente revisar el problema Bayesiano de inferencia. En la primera instancia debemos asignar una distribución inicial sobre el conjunto de todos los elementos de incertidumbre en el modelo, y después actualizar este conocimiento a la luz de nuevas observaciones mediante el Teorema de Bayes.

En términos generales podemos suponer que un conjunto de datos, denotados por $X_1, ..., X_n$, corresponden a realizaciones de una variable aleatoria X asociada a una función de distribución (o medida de probabilidad) F desconocida, y sobre la cual deseamos obtener inferencias a partir de la muestra. En los problemas paramétricos se supone que esta función pertenece a una familia de distribuciones paramétricas previamente determinada, que está indexada por un conjunto de parámetros desconocidos de dimensión finita, denotados por $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Nuestra incertidumbre se reduce al desconocimiento sobre el verdadero valor de $\boldsymbol{\theta}$. Usando el paradigma Bayesiano es necesario asignar una distribución inicial que represente, en la medida de lo posible, nuestro conocimiento respecto a $\boldsymbol{\theta}$. Después la inferencia y posibles predicciones de la variable aleatoria se realizan mediante los procedimientos descritos en el capítulo 2.

El problema de inferencia Bayesiana no paramétrica consiste en relajar, o reducir a un nivel mínimo, los supuestos sobre la forma funcional de la distribución F. En términos generales podemos suponer que la variable X tiene un soporte en el espacio $\mathcal{X}$, así podemos suponer que la función F es una medida de probabilidad definida sobre $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. Para efectos prácticos en este trabajo consideramos que $\mathcal{X}$ corresponde a la recta real. Ahora nuestro espacio de incertidumbre no corresponde al espacio parametral $\boldsymbol{\Theta}$, como en el caso paramétrico, sino al conjunto de todas las funciones de distribución o medidas de probabilidad definidas sobre $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, i.e.

$$\mathcal{F} = \{F : F \text{ es una medida de probabilidad sobre } (\mathcal{X}, \mathcal{B}(\mathcal{X}))\}. \tag{4.33}$$

Nuestro desconocimiento sobre F se refleja mediante una apreciación de aleatoriedad sobre su comportamiento o forma funcional, por lo que todas aquellas cantidades derivadas

de ésta, como son esperanzas, $\int x dF(x)$, y probabilidades $F(B)$, con $B \in \mathcal{B}(\mathcal{X})$, son también aleatorias. Bajo el paradigma Bayesiano nuestra incertidumbre respecto a F se ve reflejada mediante una medida de probabilidad inicial sobre el conjunto $\mathcal{F}$ y un σ-álgebra de subconjuntos de $\mathcal{F}$ adecuado, que denotamos por $\mathcal{B}(\mathcal{F})$. Esta medida de probabilidad inicial, definida sobre $(\mathcal{F}, \mathcal{B}(\mathcal{F}))$, la denotamos por $\mathcal{K}$ y, entre otras características, debe satisfacer que para una medida de probabilidad aleatoria $F \in \mathcal{F}$, la distribución final de F dado un conjunto de realizaciones de la variable de interés, $X_1, ..., X_n$, sea manejable analíticamente (Ferguson, 1974). Dada esta medida de probabilidad $\mathcal{K}$, podemos calcular la 'verdadera' probabilidad de que X pertenezca a un conjunto $B \in \mathcal{B}(\mathcal{X})$, mediante un proceso de marginalización respecto a $\mathcal{K}$, de la forma

$$Pr(X \in B) = \int F(B) d\mathcal{K}(F). \tag{4.34}$$

Existen diferentes alternativas para especificar la distribución inicial sobre $(\mathcal{F}, \mathcal{B}(\mathcal{F}))$. Una opción ampliamente conocida consiste en definir $\mathcal{K}$ mediante un *proceso Dirichlet* (Ferguson, 1974, sección 1). Esta alternativa resulta simple de trabajar analíticamente debido a que es conjugada[3] ante actualizaciones, sin embargo presenta el mayor inconveniente o limitación de generar medidas de probabilidad discretas con probabilidad uno. Una clase más flexible de medidas sobre $(\mathcal{F}, \mathcal{B}(\mathcal{F}))$ que corrige esta deficiencia del proceso Dirichlet consiste en definir un *proceso de colas libres* (vea por ejemplo Ferguson (1974), sección 2, o Schervish (1995), sección 1.6.2), el cual es un proceso que consiste básicamente en definir probabilidades aleatorias sobre los elementos de un árbol infinito de particiones medibles de $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, con el supuesto que $\mathcal{B}(\mathcal{X})$ sea generado por el árbol infinito de particiones. Como caso particular tiene a los procesos Dirichlet, y también a unos procesos estocásticos más sencillos de manejar en la práctica, los *árboles de Pólya.* Estos últimos son muy flexibles, bajo ciertas condiciones permiten muestrear medidas de probabilidad continuas e inclusive absolutamente continuas con probabilidad uno (Ferguson, 1974; Mauldin *et al.*, 1992; Lavine, 1992), además de mantener la propiedad conjugacional deseada que antes mencionamos. Los árboles de Pólya pueden ser una alternativa relativa-

[3] En este texto la propiedad conjugacional se refiere al hecho que la familia de distribuciones es cerrada ante incorporaciones de nuevas observaciones.

mente simple para incorporar un elemento no paramétrico en modelos de series de tiempo escalares o modelos de regresión en general. En la siguiente subsección realizaremos una breve descripción de estos procesos.

4.3.2 Árboles de Pólya

Un *árbol de Pólya* consiste en la asignación de probabilidades aleatorias sobre particiones sucesivas del espacio de interés, $\mathcal{X}$, que definen una árbol binario de particiones con un número infinito de niveles. A diferencia de los procesos de colas libres, donde las probabilidades de pertenencia a cada elemento de la partición son modeladas como independientes sólo dentro de cada nivel, en los arboles de Pólya éstas son independientes entre y dentro de los niveles.

La forma de particionar este espacio consiste en un refinamiento de un árbol binario de particiones. Sea $E = \{0, 1\}$ y $E_m = E \times \cdots \times E$ el conjunto de todas las posibles sucesiones binarias de longitud m, para todo m entero positivo. En el nivel inicial del árbol definimos la partición trivial del espacio $\mathcal{X}$ en los conjuntos B_0 y B_1, en el siguiente nivel los conjuntos B_0 y B_1 son particionados cada uno en dos subconjuntos B_{00} y B_{01} el primero, y B_{10} y B_{11} el segundo. Continuando sucesivamente este procedimiento para un índice de partición $\epsilon \in \cup_{m=1}^{\infty} E_m$, cada conjunto B_ϵ es particionado en los subconjuntos $B_{\epsilon 0}$ y $B_{\epsilon 1}$. En el nivel m del árbol de particiones, el conjunto Δ_m denotará la partición de $\mathcal{X}$ derivada de la partición Δ_{m-1} del nivel anterior, conteniendo 2^m elementos. Continuando sucesivamente hasta un nivel infinito de particiones obtenemos un árbol infinito de particiones jerarquizado de $\mathcal{X}$, denotado por $\Delta = \{\Delta_m : m = 0, 1, 2, \ldots\}$, con $\Delta_m = \{B_\epsilon : \epsilon \in E_m\}$ y $\Delta_0 = \{\mathcal{X}\}$. Supongamos además que $\cup_{m=0}^{\infty} \Delta_m$ genera al σ-álgebra de Borel en $\mathcal{X}$, denotado por $\mathcal{B}(\mathcal{X})$.

Se dice que una medida de probabilidad F sobre $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ tiene un *árbol de Pólya* por distribución inicial con parámetros (Δ, Γ) (Lavine, 1992), denotado como $F \sim PT(\Delta, \Gamma)$, si existe una colección de números no negativos $\Gamma = \{\gamma_\epsilon : \epsilon \in \cup_{m=1}^{\infty} E_m\}$ y un conjunto de variables aleatorias $\mathcal{C} = \{C_\epsilon : \epsilon \in \cup_{m=1}^{\infty} E_m\}$ tales que,

i) Todas las variables en $\mathcal{C}$ son independientes.

ii) Para todo $\epsilon \in \cup_{m=1}^{\infty} E_m$ la variable aleatoria C_ϵ tiene una distribución $Be(\gamma_{\epsilon 0}, \gamma_{\epsilon 1})$.

iii) Para todo $m = 1, 2, \ldots$ y para todo $\epsilon = \epsilon_1 \cdots \epsilon_m$ en $\cup_{m=1}^{\infty} E_m$ se tiene que,

$$F(B_{\epsilon_1 \cdots \epsilon_m}) = \prod_{j=1}^{m} \left[C_{\epsilon_1 \cdots \epsilon_{j-1}}\right]^{1-\epsilon_j} \left[1 - C_{\epsilon_1 \cdots \epsilon_{j-1}}\right]^{\epsilon_j}, \tag{4.35}$$

donde $F(B)$ denota la probabilidad (aleatoria) de que el valor de la variable aleatoria X pertenezca al conjunto B.

Las variables $C_{\epsilon 0}$ y $(1 - C_{\epsilon 0})$, con $\epsilon = \epsilon_1 \cdots \epsilon_{m-1}$, se interpretan como la probabilidad de que la variable aleatoria X pertenezca a los conjuntos $B_{\epsilon 0}$ y $B_{\epsilon 1}$ respectivamente, dado que pertenece al conjunto B_ϵ, para todo m. Así, la probabilidad inicial de que la variable aleatoria X pertenezca a un conjunto B_ϵ en el árbol de partición Δ, con $\epsilon = \epsilon_1 \cdots \epsilon_m$, se puede calcular mediante el proceso de marginalización (4.34),

$$\begin{aligned} Pr(X \in B_\epsilon) &= \mathbb{E}_{PT}\left[F(B_\epsilon)\right] \\ &= \prod_{j=1}^{m} \mathbb{E}\left[\left[C_{\epsilon_1 \cdots \epsilon_{j-1}}\right]^{1-\epsilon_j} \left[1 - C_{\epsilon_1 \cdots \epsilon_{j-1}}\right]^{\epsilon_j}\right] \\ &= \frac{\gamma_{\epsilon_1}}{\gamma_0 + \gamma_1} \frac{\gamma_{\epsilon_1 \epsilon_2}}{\gamma_{\epsilon_1 0} + \gamma_{\epsilon_1 1}} \cdots \frac{\gamma_{\epsilon_1 \cdots \epsilon_{m-1} \epsilon_m}}{\gamma_{\epsilon_1 \cdots \epsilon_{m-1} 0} + \gamma_{\epsilon_1 \cdots \epsilon_{m-1} 1}}. \end{aligned} \tag{4.36}$$

La probabilidad (4.36) define una medida de probabilidad sobre $\cup_{m=0}^{\infty} \Delta_m$, pero dado que esta unión genera al σ-álgebra de Borel en $\mathcal{X}$, ésta se puede extender de manera única a $\mathcal{B}(\mathcal{X})$ (Lavine, 1992). La definición de un árbol de Pólya depende fundamentalmente de la definición de las particiones del espacio $\mathcal{X}$ y de los parámetros asociados Γ. Lavine (1992) demostró que es posible centrar el árbol de Pólya en una función de distribución G fija y arbitraria, tal que $\mathbb{E}_{PT}\left[F(B_\epsilon)\right] = G(B_\epsilon)$, para todo $\epsilon \in \cup_{m=1}^{\infty} E_m$. Esto se obtiene haciendo coincidir la partición de $\mathcal{X}$ con ciertos cuantiles de la distribución G. Si $\mathcal{X}$ coincide con la recta real, como es nuestro caso, podemos definir el primer nivel del árbol de particiones como $B_0 = (-\infty, G^{-1}(1/2))$ y $B_1 = [G^{-1}(1/2), \infty)$. Subsecuentemente, en cada nivel $m \geq 2$, podemos definir cada j-ésimo elemento de la partición Δ_m como el conjunto $B_j = [G^{-1}((j-1)/2^m), G^{-1}(j/2^m))$ con $G^{-1}(0) = -\infty$ y $G^{-1}(1) = \infty$, para $j = 1, \ldots, 2^m$.

Además del papel fundamental que juega en el proceso la definición del árbol de particiones Δ, la definición de los parámetros en Γ también puede determinar diferentes ca-

racterísticas del árbol de Pólya. Como explica Lavine (1992, página 1227), los parámetros γ_ϵ's pueden determinar el tipo de distribuciones que modelamos con el árbol de Pólya. Por ejemplo, si deseamos modelar distribuciones continuas, esperamos que en los valores altos de m, i.e. en niveles bajos del árbol de particiones, exista un efecto presente de cercanía entre las probabilidades de los conjuntos contiguos de este nivel, ya que conforme m aumenta la partición de $\mathcal{X}$ es más fina. Si deseamos que la función sea continua esperamos heurísticamente que no exista una variación significativa entre las probabilidades de los elementos contiguos de Δ_m en los niveles bajos del árbol. Este efecto se logra si elegimos valores "grandes" de los parámetros γ's de la distribución Beta de ese nivel, y que $\gamma_{\epsilon 0}$ y $\gamma_{\epsilon 1}$ sean semejantes también. Por otro lado también es deseable que exista una mayor libertad del proceso en los niveles altos del árbol, para flexibilizar el comportamiento de la distribución que estamos modelando.

Los parámetros en Γ pueden determinar el comportamiento del árbol de Pólya dentro de una clase de medidas de probabilidad específicas (Ferguson, 1974, páginas 620-621). Si definimos para cada $\epsilon = \epsilon_1 \cdots \epsilon_m$ que los parámetros de $C_\epsilon \sim Be(\gamma_{\epsilon 0}, \gamma_{\epsilon 1})$ satisfagan que $\gamma_\epsilon = \gamma_{\epsilon 0} + \gamma_{\epsilon 1}$, particularmente si $\gamma_\epsilon = 2^{-m}$, entonces el árbol de Pólya $PT(\Delta, \Gamma)$ será un proceso Dirichlet y F será discreta con probabilidad uno (Lavine, 1992). Para flexibilizar el modelo, es posible elegir ciertos valores en Γ de manera que F sea continua e inclusive absolutamente continua con probabilidad uno. Por ejemplo, si definimos $\gamma_{\epsilon_1 \cdots \epsilon_m} = m^2$ para todo m (e.g., Ferguson (1974), página 621, o Lavine (1992)), generaremos distribuciones absolutamente continuas con probabilidad uno. La condición general para que un árbol de Pólya muestree distribuciones absolutamente contínuas con probabilidad uno es que $\sum_{m=1}^{\infty} \gamma_m^{-1} < \infty$ (Kraft, 1964), por lo que si elegimos $\gamma_m = \gamma_{\epsilon_1 \cdots \epsilon_m} = Cm^2$ para todo m, con $C > 0$, se satisface esta condición.

Como mencionamos, los árboles de Pólya preservan la propiedad conjugacional ante actualizaciones con nuevas observaciones de la variable de interés. Sean $X_1, \ldots, X_n$ un conjunto de variables aleatorias tales que condicional en F son independientes, con $X_i \sim F$ y $F \sim PT(\Delta, \Gamma)$, entonces la distribución final de F dado el conjunto de datos observado es también un árbol de Pólya con parámetros Δ y Γ' (Lavine, 1992), denotado como

$F|X_1,\ldots,X_n \sim PT(\Delta,\Gamma')$, donde los valores de $\Gamma' = \{\gamma'_\epsilon :$ para todo $\epsilon\}$ son tales que

$$\gamma'_\epsilon = \gamma_\epsilon + \sum_{i=1}^{n} \delta_{B_\epsilon}(X_i), \tag{4.37}$$

donde $\delta_{B_\epsilon}(x)$ denota la función indicadora de que $x \in B_\epsilon$, i.e. $C_\epsilon|X_1,\ldots,X_n \sim Be(\gamma'_{\epsilon 0}, \gamma'_{\epsilon 1})$.

Haciendo uso de la propiedad conjugacional del árbol de Pólya y de la relación (4.36), podemos calcular probabilidades predictivas finales de una variable futura X_f mediante el cálculo de

$$\begin{aligned} Pr(X_f \in B_\epsilon|X_1,\ldots,X_n) &= \frac{\gamma_{\epsilon_1}+n_{\epsilon_1}}{\gamma_0+\gamma_1+n}\frac{\gamma_{\epsilon_1\epsilon_2}+n_{\epsilon_1\epsilon_2}}{\gamma_{\epsilon_1 0}+\gamma_{\epsilon_1 1}+n_{\epsilon_1}} \\ &\quad \times\cdots\times \frac{\gamma_{\epsilon_1\cdots\epsilon_{m-1}\epsilon_m}+n_{\epsilon_1\cdots\epsilon_{m-1}\epsilon_m}}{\gamma_{\epsilon_1\cdots\epsilon_{m-1}0}+\gamma_{\epsilon_1\cdots\epsilon_{m-1}1}+n_{\epsilon_1\cdots\epsilon_{m-1}}}, \end{aligned} \tag{4.38}$$

donde $n_\epsilon = \sum_{i=1}^{n} \delta_{B_\epsilon}(X_i)$, para todo B_ϵ.

Como vimos en (4.37), el cálculo de la distribución final de F es simple, pero en casos donde consideremos que los datos de la variable X son observados de manera continua es necesario actualizar un número infinito de parámetros en el árbol de Pólya, lo que imposibilita su implementación práctica. Lavine (1992, 1994, sección 3) demostró que podemos detener la actualización de los parámetros hasta un nivel finito predeterminado del árbol de particiones, digamos L. En el caso de modelar distribuciones absolutamente continuas, debemos considerar la estructura de los parámetros en Γ en la especificación del nivel máximo de actualizaciones. Como mencionamos, si deseamos modelar distribuciones continuas, es deseable que a partir de un cierto nivel las probabilidades de los elementos contiguos de los niveles bajos de particiones sean semejantes entre sí, y que los parámetros asociados a éstos sean suficientemente grandes, para lograr este efecto ante actualizaciones con nuevas observaciones, de manera que tengamos un efecto de "contigüidad" entre las probabilidades en los niveles bajos del árbol de particiones.

Cuando consideramos la actualización hasta un nivel finito del árbol de particiones se conocen como *árboles de Pólya finitos o parcialmente especificados.* Lavine (1994) brinda un argumento para determinar el nivel mínimo M de manera que aproximemos la densidad predictiva final con una precisión deseada, podemos definir un factor $\delta > 0$ pequeño y actualizar el árbol de Pólya hasta el nivel M, de manera que $\log(\delta) \geq (n/2)\sum_{m=L}^{\infty} m^{-2}$,

donde n es el tamaño de la muestra. En la práctica, ciertos autores han especificado el nivel mínimo igual a 8 (Walker *et al.*, 1999).

Un árbol de Pólya parcialmente especificado hasta un nivel predeterminado L se denota por $PT(\Delta_L, \Gamma_L)$, donde Δ_L y Γ_L están definidos como antes hasta el nivel L. Para efectos prácticos, los cálculos relacionados con el árbol de Pólya parcialmente especificado están relacionados con el último nivel de la partición asociadas a los elementos B_j, para $j = 1, \ldots, 2^L$, de la misma. La idea de considerar un nivel de actualizaciones finito del árbol de Pólya se relaciona con el hecho de concentrar nuestras inferencias en un refinamiento pequeño de una zona de alta probabilidad en F. El cálculo de las probabilidades condicionales en F, y de las probabilidades marginales, se obtiene de manera directa de las derivaciones (4.35) y (4.36) respectivamente, todas éstas realizadas hasta el nivel L del árbol de particiones. En algunas ocasiones, para obtener inferencias respecto a F es necesario generar muestras de F. Éstas se obtienen de manera directa mediante la definición de árbol de Pólya, a través de la simulación sucesiva de variables aleatorias independientes Beta con los parámetros respectivos, y aplicando la definición (4.35). En el caso de los árboles de Pólya parcialmente especificados, de la muestra de F sólo son de interés las masas de probabilidad asociadas a los elementos de la partición en el último nivel L de actualización. Estas últimas las denotamos por $\boldsymbol{\rho}_{PT} = \{\rho_j : j = 1, \ldots, 2^L\}$, donde $\rho_j = F(B_j)$ para todo B_j en Δ_L.

Walker *et al.* (1999) realizan una revisión de éste y otros modelos Bayesianos no paramétricos relacionados. Los detalles teóricos sobre los árboles de Pólya se encuentran en Lavine (1992, 1994) y Mauldin *et al.* (1992). Resultados más generales para los procesos de colas libres se encuentran en Schervish (1995, sección 1.6). A continuación describimos la incorporación de los árboles de Pólya al modelo de regresión lineal descrito por Gelfand (1998).

4.3.3 Regresión Semiparamétrica vía Arboles de Pólya

De vuelta a nuestro problema principal, podemos considerar un modelo de regresión lineal en el que la parte aleatoria tenga una distribución no paramétrica. Gelfand (1998) realiza

una descripción de éste y otros modelos semiparamétricos relacionados. Supongamos que $\{y_t\}$ es una serie de tiempo y $\{\boldsymbol{x}_t\}$ es una serie de vectores de regresión de dimensión p, en la que posiblemente se incluyan retrasos de la misma serie de interés. El modelo de regresión semiparamétrica es de la forma

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + \varepsilon_t, \tag{4.39}$$

donde $\boldsymbol{\beta}$ son los coeficientes de regresión, y $\{\varepsilon_t\}$ es una sucesión de variables independientes -o intercambiables-, que representan las perturbaciones o ruido del modelo. Estas perturbaciones tienen asociada una distribución F sobre la recta real, y es tal que condicional en F la sucesión de perturbaciones aleatorias $\{\varepsilon_t\}$ son independientes e idénticamente distribuidas con distribución F desconocida. Supongamos inicialmente que $\boldsymbol{\beta}$ y F son independientes. Sobre los coeficientes de regresión asignaremos una distribución inicial Normal multivariada vaga o poco informativa, mientras que a F le asignaremos inicialmente un árbol de Pólya parcialmente especificado, con $\gamma_{\epsilon_1 \cdots \epsilon_m} = Cm^2$, donde $C > 0$ es un valor que facilita el efecto de contigüidad de F en los niveles bajos del árbol de particiones. Algunos autores sugieren considerar $C = 0.1$ (Walker *et al.*, 1999). Los procesos de ruido tradicionales restringen a F de manera que tenga media cero, en los árboles de Pólya esta restricción es difícil de implementar, sin embargo debido a la naturaleza de las particiones, podemos sustituir esta restricción utilizando a la mediana como el componente de localización del proceso de ruido. En este caso F se restringe de manera que tenga mediana cero, que es simple de obtener definiendo el primer nivel de la partición del árbol de Pólya como lo mencionamos antes, de manera que $F(B_0) = F(B_1) = 1/2$. Para obtener cierto control en el proceso de inferencia, centraremos el árbol de Pólya en una distribución Normal con media (mediana) cero y una varianza conocida. Podemos ampliar una jerarquía en el modelo, y asignar una distribución sobre la varianza de la distribución de centralidad del árbol de Pólya, aunque no consideramos esta alternativa por razones de simplificación computacional del modelo. Algunos autores sugieren determinar una varianza grande en la distribución de centralidad de manera que el modelo sea más robusto en las actualizaciones, y no se concentre en actualizaciones sobre las colas de la distribución.

Ahora, supongamos que $\boldsymbol{y}_T = (y_1, ..., y_T)'$ es la trayectoria observada del proceso hasta el tiempo T. Para realizar inferencias o predicciones es necesario obtener la distribución final conjunta de $(\boldsymbol{\beta}, F)$ dado $\boldsymbol{y}_T$, que analíticamente es imposible de obtener. Sin embargo podemos obtener una muestra de la distribución final conjunta mediante MCCM. De hecho es posible implementar un algoritmo híbrido (Tierney, 1994; Müller, 1993) del muestreador de Gibbs con un paso de Metropolis-Hastings para $\boldsymbol{\beta}$. Obtener una muestra de la distribución final $p(F|\boldsymbol{\beta}, \boldsymbol{y}_T)$ es relativamente simple siguiendo la definición y construcción del árbol de Pólya. Condicional en $\boldsymbol{\beta}$, los valores $\varepsilon_t = y_t - \boldsymbol{x}_t'\boldsymbol{\beta}$, $t = 1, \ldots, T$, constituyen una muestra aleatoria independiente e idénticamente distribuida de F, así que la actualización de la distribución de F se obtiene de manera directa mediante la regla de actualización (4.37) hasta el nivel L. La muestra de $F \sim PT(\Delta_L, \Gamma_L')$ se obtiene mediante la generación de muestras sucesivas de variables aleatorias Beta de acuerdo a la definición del árbol de Pólya, de la que sólo se conservan las masas de probabilidad para los conjuntos de la partición Δ_L, denotadas por $\boldsymbol{\rho}_{PT} = \{\rho_j : j = 1, \ldots, 2^L\}$.

Por otro lado, la distribución final condicional completa de $\boldsymbol{\beta}$ es de la forma

$$p(\boldsymbol{\beta}|F, \boldsymbol{y}_T) \propto p(\boldsymbol{y}_T|F, \boldsymbol{\beta})\pi(\boldsymbol{\beta}), \tag{4.40}$$

donde $\pi(\boldsymbol{\beta})$ denota la distribución inicial de $\boldsymbol{\beta}$, y $p(\boldsymbol{y}_T|F, \boldsymbol{\beta})$ denota la verosimilitud del modelo condicional en F, que en el caso de los árboles de Pólya ésta es proporcional al producto $\prod_{t=1}^{T} F(B_{\epsilon_1,\ldots,\epsilon_L}(\varepsilon_t))$, donde $B_{\epsilon_1,\ldots,\epsilon_L}(\varepsilon_t)$ denota al elemento de la partición del árbol de Pólya en el nivel L que contiene a ε_t, y las probabilidades son calculadas como (4.35) con $F \sim PT(\Delta_M, \Gamma_M)$. En este caso la distribución final condicional completa de $\boldsymbol{\beta}$ no es conocida analíticamente, sin embargo, podemos generar una muestra de ésta mediante un paso de M-H. De acuerdo a Walker *et al.* (1999) podemos implementar una caminata aleatoria usando una distribución instrumental Normal multivariada con una matriz de varianzas y covarianzas constante conocida. La probabilidad de aceptación del movimiento para un nuevo estado $\boldsymbol{\beta}'$ propuesto, condicional en una muestra de F, es

$$\alpha(\boldsymbol{\beta}, \boldsymbol{\beta}') = \text{mín}\left(1, \frac{p(\boldsymbol{y}_T|F, \boldsymbol{\beta})\pi(\boldsymbol{\beta})}{p(\boldsymbol{y}_T|F, \boldsymbol{\beta}')\pi(\boldsymbol{\beta}')}\right). \tag{4.41}$$

A continuación describimos el esquema general de muestreo para el modelo semi-

paramétrico de regresión mediante árboles de Pólya.

Algoritmo 3: Regresión Semiparamétrica mediante Arboles de Pólya

Definimos los valores iniciales de la cadena para $\boldsymbol{\beta}$, con su estimador de mínimos cuadrados, y generamos una muestra de F con los valores iniciales (Δ_L, Γ_L), de la cual sólo conservamos las masas de probabilidad $\boldsymbol{\rho}_{PT} = \{\rho_j : j = 1, \ldots, 2^L\}$.

1. Suponiendo que el estado actual de la cadena es $(\boldsymbol{\beta}, F)$, obtenemos una muestra de $\boldsymbol{\beta}' \sim N(\boldsymbol{\beta}'|\boldsymbol{\beta}, \boldsymbol{V})$, con $\boldsymbol{V}$ conocida.

   (a) Aceptamos $\boldsymbol{\beta}'$ como nuevo estado de la cadena con probabilidad $\alpha(\boldsymbol{\beta}, \boldsymbol{\beta}')$ dada por (4.41), calculada a partir del estado actual de F.

2. Condicional en $\boldsymbol{\beta}$ (después del paso de M-H), calculamos Γ'_L de acuerdo a la relación (4.37) y generamos una muestra de $F \sim PT(\Delta_L, \Gamma'_L)$ en el nivel L de particiones del árbol de Pólya.

3. Condicional en el estado actual de F generamos una muestra de ε, y condicional en el estado actual de $\boldsymbol{\beta}$ obtenemos una muestra predictiva futura de $y_{T+1} = \boldsymbol{x}'_{T+1}\boldsymbol{\beta} + \varepsilon$.

4. Continuamos hasta obtener la convergencia de la cadena.

En la siguiente subsección describiremos un modelo más flexible, que permite modelar comportamientos heteroscedásticos de la serie de interés.

4.3.4 Regresión Semiparamétrica con Errores GARCH

En ocasiones el proceso $\{y_t\}$ presenta un evidente comportamiento heteroscedástico o de cambio en la volatilidad. Para modelar este comportamiento necesitamos incorporar en el modelo un componente o estructura relacionada con la evolución la varianza del proceso. Con este objetivo existen diferentes alternativas para modelar evoluciones de este tipo. Los modelos más representativos son los modelos ARCH y generalizaciones, donde se supone que la varianza de la serie evoluciona de acuerdo a un proceso lineal determinista relacionado con valores pasados de la misma varianza o de valores pasados del proceso

como en los modelos GARCH. Otra alternativa consiste en suponer que la varianza del proceso sigue un proceso aleatorio latente, independiente de sus propios valores pasados o de los errores, esta clase de modelos se conoce como modelos de volatilidad estocástica. En general, no existe evidencia alguna sobre el dominio de alguno de estos modelos respecto al otro en términos de su capacidad predictiva. En Ghysels *et al.* (1996) y Shephard (1996) podemos encontrar una revisión general sobre estos dos tipos de modelos y sus generalizaciones. En este trabajo consideramos sólo el caso de los modelos GARCH, siguiendo el trabajo de Denison y Mallick (1999), quienes incorporan un elemento no paramétrico mediante el uso de árboles de Pólya en la distribución del proceso de ruido en un modelo GARCH(1,1). Aquí consideramos solamente la extensión al modelo de regresión lineal con esta estructura en los errores, además de un esquema de muestreo distinto.

Consideremos que $\{y_t\}$ es una serie de tiempo escalar y $\{\boldsymbol{x}_t\}$ es un conjunto de vectores de regresión de dimensión p, con la posibilidad de incluir retrasos de la misma serie de interés. La extensión flexible al modelo (4.39) está dada por

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + \sigma_t\varepsilon_t, \tag{4.42}$$

donde $\boldsymbol{\beta}$ es el vector de coeficientes de regresión desconocidos, $\{\varepsilon_t\}$ es un proceso de ruido aleatorio con función de distribución F desconocida, y $\{\sigma_t\}$ es el proceso de volatilidad de la serie que evoluciona de acuerdo a un proceso GARCH(1,1) dado por

$$\sigma_t^2 = \alpha_0 + \alpha_1\varepsilon_{t-1}^2 + \alpha_2\sigma_{t-1}^2, \tag{4.43}$$

donde $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2)'$ son valores reales positivos desconocidos. Modelando no paramétricamente F, tenemos que los elementos de interés desconocidos son $(\boldsymbol{\beta}, \boldsymbol{\alpha}, F)$. Suponemos inicialmente que estos componentes son independientes. Al igual que en el modelo de regresión (4.39), asignamos una distribución inicial Normal multivariada difusa para $\boldsymbol{\beta}$, centrada en el vector p-dimensional nulo y matriz de covarianzas conocida. A F le asignamos inicialmente un árbol de Pólya parcialmente especificado, con parámetros dados como el caso de regresión semiparamétrica. Respecto a $\boldsymbol{\alpha}$ existen diferentes elementos por considerar, en el caso de α_0 es conveniente trabajar con la reparametrización $a_0 =$

$\log \alpha_0$, sobre la cual asignamos una distribución inicial Normal centrada en su estimador máximo verosímil, bajo el supuesto de ruido blanco Gaussiano, y varianza constante. Para garantizar que el proceso (4.43) sea estacionario, restringimos que $\alpha_1 + \alpha_2 < 1$, de manera que estamos trabajando en el simplex euclidiano bidimensional.

Supongamos que $\{\boldsymbol{y}_T\}$ denota la trayectoria de la serie hasta el tiempo T. Para inferir y realizar predicciones necesitamos generar muestras de la distribución final conjunta de $(\boldsymbol{\beta}, F, a_0, \alpha_1, \alpha_2)$, que puede resultar problemático, inclusive en el caso paramétrico usual donde $\{\varepsilon_t\}$ es un proceso de ruido blanco Gaussiano. Podemos obtener muestras de la distribución final de estos elementos mediante MCCM, en particular adaptaremos el esquema de muestreo híbrido de Gibbs/Metropolis-Hastings que Müller y Pole (1998) propusieron para el caso del modelo GARCH con ruido blanco Gaussiano.

La distribución final condicional completa de $\boldsymbol{\beta}$, dado $(\boldsymbol{\alpha}, F, \boldsymbol{y}_T)$, es analíticamente desconocida. Para poder obtener una muestra de ésta instrumentaremos un paso de M-H independiente proponiendo un estado nuevo, $\boldsymbol{\beta}'$, mediante la distribución final instrumental obtenida del modelo alternativo

$$\boldsymbol{y}_T = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{H}\boldsymbol{\varepsilon}_T, \tag{4.44}$$

donde $\boldsymbol{X} = (\boldsymbol{x}_1', \ldots, \boldsymbol{x}_T')'$, $\boldsymbol{H} = \text{diag}(\sigma_1, \ldots, \sigma_T)$ y $\boldsymbol{\varepsilon}_T = (\varepsilon_1, ..., \varepsilon_T)'$, estos últimos calculados recursivamente con los estados actuales de la cadena $(\boldsymbol{\beta}, F, \boldsymbol{\alpha})$, suponiendo que $\boldsymbol{\varepsilon}_T$ tiene una distribución Normal multivariada estándar. En este caso, la distribución final instrumental corresponde a la distribución final del modelo de regresión lineal heteroscedástico con matriz de varianzas-covarianzas, $\boldsymbol{H}$, conocida. Así el nuevo estado propuesto es $\boldsymbol{\beta}' \sim N(\boldsymbol{\beta}'|\boldsymbol{m}, \boldsymbol{V})$, donde $\boldsymbol{V} = (\boldsymbol{V}_0^{-1} + \boldsymbol{X}'(\boldsymbol{H}\boldsymbol{H}')^{-1}\boldsymbol{X})^{-1}$ y $\boldsymbol{m} = \boldsymbol{V}\boldsymbol{X}'(\boldsymbol{H}\boldsymbol{H}')^{-1}\boldsymbol{y}_T$, con $\boldsymbol{V}_0$ la matriz de varianzas-covarianzas inicial de $\boldsymbol{\beta}$, y la probabilidad de aceptación del nuevo estado está dada por

$$\delta_1(\boldsymbol{\beta}, \boldsymbol{\beta}') = \text{mín}\left(1, \frac{p(\boldsymbol{y}_T|\boldsymbol{\beta}', F, \boldsymbol{\alpha})\pi(\boldsymbol{\beta}')/N(\boldsymbol{\beta}'|\boldsymbol{m}, \boldsymbol{V})}{l(\boldsymbol{y}_T|\boldsymbol{\beta}, F, \boldsymbol{\alpha})\pi(\boldsymbol{\beta})/N(\boldsymbol{\beta}|\boldsymbol{m}, \boldsymbol{V})}\right), \tag{4.45}$$

donde en este caso,

$$p(\boldsymbol{y}_T|\boldsymbol{\beta}, F, \boldsymbol{\alpha}) \propto \prod_{t=1}^{T} 1/\sigma_t F(B_{\epsilon_1 \cdots \epsilon_L}(\varepsilon_t)), \tag{4.46}$$

con $\varepsilon_t = (y_t - \boldsymbol{x}_t'\boldsymbol{\beta})/\sigma_t$, para $t = 1, \ldots, T$.

Muestrear de la distribución final condicional completa de $(a_0, \alpha_1, \alpha_2)$ es bastante más complicado. Empíricamente encontramos que existe una probabilidad de aceptación muy pequeña para nuevos movimientos cuando utilizamos una distribución inicial uniforme en el simplex para los coeficientes del componente GARCH, i.e. $\pi(\alpha_1, \alpha_2) \propto \mathbf{1}(\alpha_1 > 0, \alpha_2 > 0, \alpha_1 + \alpha_2 < 1)$. Para eliminar este problema consideramos trabajar con la reparametrización obtenida mediante la transformación logística del simplex bidimensional $a_1 = \log(\alpha_1/(1-\alpha_1-\alpha_2))$ y $a_2 = \log(\alpha_2/(1-\alpha_1-\alpha_2))$, y asignamos una distribución inicial informativa sobre (a_1, a_2) Normal multivariada centrada en el vector $\boldsymbol{b} = (b_1, b_2)'$ con una matriz de covarianzas $\boldsymbol{B}_1$.

En este caso utilizamos como distribución instrumental para el paso de M-H a una distribución Normal sobre $\boldsymbol{a} = (a_0, a_1, a_2)'$ centrada en los estimadores de máxima verosimilitud de $\boldsymbol{a}$, que es el vetor $\boldsymbol{b} = (b_0, b_1, b_2)'$, donde b_0 es el el logaritmo natural del estimador de máxima verosimilitud de α_0, y b_1 y b_2 son la transformación logística del estimador de máxima verosimilitud de α_1 y α_2, en el caso de que los errores sean ruido blanco Gaussiano. La matriz de covarianzas de la distribución instrumental es $\boldsymbol{B} = \text{diagbloq}(B_0, \boldsymbol{B}_1)$, donde B_0 es la varianza previamente especificada para a_0 y $\boldsymbol{B}_1$ es una matriz de covarianzas estimada con los puntos de las curvas de nivel de la verosimilitud para (a_1, a_2). Suponiendo que el estado actual de la cadena es $(\boldsymbol{\beta}, F, \boldsymbol{a})$, proponemos $\boldsymbol{a}' \sim N(\boldsymbol{a}'|\boldsymbol{b}, \boldsymbol{B})$, y aceptamos este valor como el nuevo estado de la cadena con probabilidad

$$\delta_2(\boldsymbol{a}, \boldsymbol{a}') = \text{mín}\left(1, \frac{p(\boldsymbol{y}_T|\boldsymbol{\beta}, F, \boldsymbol{a}')\pi(\boldsymbol{a}')/N(\boldsymbol{a}'|\boldsymbol{b}, \boldsymbol{B})}{p(\boldsymbol{y}_T|\boldsymbol{\beta}, F, \boldsymbol{a})\pi(\boldsymbol{a})/N(\boldsymbol{a}|\boldsymbol{b}, \boldsymbol{B})}\right), \tag{4.47}$$

donde $p(\boldsymbol{y}_T|\boldsymbol{\beta}, F, \boldsymbol{a})$ esta dada por (4.46), con los valores de la sucesión $\{\sigma_t\}$ que son calculados recursivamente mediante los valores obtenidos de la transformación antilogística de (a_1, a_2).

Condicional en $(\boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{y}_T)$ la distribución final de F es un árbol de Pólya $PT(\Delta_L, \Gamma_L')$, donde Γ_L' es el conjunto de parámetros actualizados de acuerdo a (4.37) con base en la sucesión de valores i.i.d. $\varepsilon_t = (y_t - \boldsymbol{x}_t'\boldsymbol{\beta})/\sigma_t$, para $t = 1, \ldots, T$, con $\{\sigma_t\}$ calculado recursivamente por medio de los estados actuales de la cadena de $(\boldsymbol{\beta}, \boldsymbol{a})$. Una muestra de $PT(\Delta_L, \Gamma_L')$ se obtiene de manera directa siguiendo la definición del árbol de Pólya (4.35),

en la que de nuevo sólo se conservan las masas de probabilidad $\boldsymbol{\rho}_{PT} = \{\rho_j : j = 1, \ldots, 2^L\}$ asociados a los elementos de la partición en el nivel L.

Bajo este esquema de muestreo existe la posibilidad de que la cadena se estanque en regiones donde la probabilidad de aceptación sea cercana a cero, particularmente en la probabilidad de aceptación de los coeficientes del componente GARCH (4.47) cuando la cadena alcanza puntos donde la densidad instrumental tiene significativamente una menor densidad que la distribución final de interés. Para mitigar este posible comportamiento corremos de manera simultánea R cadenas de Markov con distinta distribución propuesta e implementamos un intercambio de estados en cada iteración mediante un paso de *Metropolis acoplado* (Gilks y Roberts, 1996). Cada cadena utiliza una distribución instrumental Normal para el componente $\boldsymbol{a}$ centrada en $\boldsymbol{b}$ como mencionamos anteriormente, pero con una matriz de covarianzas aumentada por un factor k_i, para $i = 2, ..., R$, de manera que todas las cadenas tengan la misma distribución estacionaria, con $k_1 = 1$ y $k_1 > 1$ para $i = 2, ..., R$. En cada cadena, las probabilidades de aceptación de los movimientos están dadas como antes, particularmente para el componente los coeficientes del componente GARCH, la probabilidad de aceptación del movimiento está dada por

$$\delta_{2,i}(\boldsymbol{a}_{(i)}, \boldsymbol{a}'_{(i)}) = \text{mín}\left(1, \frac{p(\boldsymbol{y}_T|\boldsymbol{\beta}_{(i)}, F_{(i)}, \boldsymbol{a}'_{(i)})\pi(\boldsymbol{a}'_{(i)})}{p(\boldsymbol{y}_T|\boldsymbol{\beta}_{(i)}, F_{(i)}, \boldsymbol{a}_{(i)})\pi(\boldsymbol{a}_{(i)})}\right), \tag{4.48}$$

donde $(\boldsymbol{\beta}_{(i)}, F_{(i)})$ denota el estado actual de la i-ésima cadena en el componente de regresión y del árbol de Pólya, $\boldsymbol{a}_{(i)}$ denota el estado actual del componente GARCH y $\boldsymbol{a}'_{(i)}$ denota el estado propuesto, para $i = 1, \ldots, R$. Así en cada iteración, suponiendo que el estado actual de las R cadenas es $(\boldsymbol{\beta}_{(i)}, F_{(i)}, \boldsymbol{a}_{(i)})$, para $i = 1 \ldots, K$, proponemos el intercambio de estados entre la cadena i y j con una probabilidad de aceptación

$$\eta(\boldsymbol{\theta}_{(i)}, \boldsymbol{\theta}_{(j)}) = \text{mín}\left(1, \frac{\pi_i(\boldsymbol{\theta}_{(j)}|\boldsymbol{y}_T)}{\pi_i(\boldsymbol{\theta}_{(i)}|\boldsymbol{y}_T)} \frac{\pi_j(\boldsymbol{\theta}_{(i)}|\boldsymbol{y}_T)}{\pi_j(\boldsymbol{\theta}_{(j)}|\boldsymbol{y}_T)}\right), \tag{4.49}$$

donde $\pi_i(\boldsymbol{\theta}|\boldsymbol{y}_T) \propto l(\boldsymbol{\theta}|\boldsymbol{y}_T)\pi_i(\boldsymbol{\theta})$, con $\boldsymbol{\theta} = (\boldsymbol{\beta}, F, \boldsymbol{a})$. Al final se eliminan los registros de las $R - 1$ cadenas auxiliares y se conservan las muestras de la primera cadena, que es la que tiene a la distribución final de $(\boldsymbol{\beta}, \boldsymbol{a}, F)$ como distribución invariante.

Algoritmo 4: Regresión Semiparamétrica con Errores GARCH(1,1)

Determinamos los valores iniciales de $(\boldsymbol{\beta}, F, \boldsymbol{a})$ para las R cadenas.

1. En cada cadena calculamos $\rho_{L,(i)}$ que sirve de base para calcular la verosimilitud (4.46).
2. Para cada cadena proponemos un estado $\beta'_{(i)}$ de (4.44) y aceptamos con probabilidad (4.45).
3. Condicional en $(F_{(i)}, \boldsymbol{\beta}_{(i)})$ proponemos $\boldsymbol{a}'_{(i)} \sim N(\boldsymbol{a}'_{(i)}|\boldsymbol{b}, k_i\boldsymbol{B})$. Aceptamos como un estado nuevo de la cadena con probabilidad (4.48).
4. Condicional en $(\boldsymbol{\beta}_{(i)}, \boldsymbol{a}_{(i)})$ actualizamos $\Gamma'_{L,(i)}$ y obtenemos una muestra de $F_{(i)}$ de la que sólo conservamos $\boldsymbol{\rho}_{L,(i)}$.
5. En cada iteración intercambiamos los estados actuales entre las cadenas 1 y j, con $j \geq 2$, con probabilidad (4.49).
6. Continuamos hasta obtener la convergencia de la cadena.

Con el propósito de obtener un modelo más flexible, consideramos mezclar el modelo que acabamos de describir junto con el modelo modelo semiparamétrico de la sección 4.1. A continuación presentamos el protocolo de mezcla.

4.4 Mezcla de Modelos Semiparamétricos

En un intento por determinar un modelo más flexible consideramos mezclar los dos modelos semiparamétricos descritos en las secciones 4.2 y 4.3, considerando que cada uno se concentra en modelar semiparamétricamente diferentes componentes del proceso representado por (4.1). El primer modelo considerado es,

M-I:

$$\begin{aligned} y_t &= \alpha_{k,0} + \boldsymbol{x}'_t\boldsymbol{\beta}_{k,1} + \sum_{j=1}^{k} \omega_{k,j}\psi_j(\boldsymbol{x}_t) + \varepsilon_t, \\ \varepsilon_t &\overset{iid}{\sim} N(\varepsilon_t|0, \sigma^2), \end{aligned} \tag{4.50}$$

donde $\{\varepsilon_t\}$ es un proceso de ruido blanco Gaussiano con varianza σ^2 desconocida, y el número de bases k es desconocido en el conjunto $\{k_{\text{mín}}, \ldots, k_{\text{máx}}\}$, junto con sus correspondientes parámetros de traslación y dilatación. En este caso asignamos una distribución

inicial uniforme sobre el número de bases, i.e.

$$p(k) = \frac{1}{k_{\text{máx}} - k_{\text{mín}} + 1} \mathbf{1}_{\{k_{\text{mín}},\ldots,k_{max}\}}(k),$$

mientras que en las demás cantidades de interés asignamos distribuciones iniciales propias, pero difusas, como describimos en la sección 4.2.3. El segundo modelo de la mezcla es,

M-II:

$$\begin{aligned} y_t &= \boldsymbol{x}_t'\boldsymbol{\beta} + \sigma_t\varepsilon_t, \\ \sigma_t^2 &= \alpha_0 + \alpha_1\varepsilon_{t-1}^2 + \alpha_2\sigma_{t-1}^2, \\ \varepsilon_t|F &\overset{iid}{\sim} F, \end{aligned} \tag{4.51}$$

con las distribuciones iniciales sobre $(F, \boldsymbol{\beta}, \boldsymbol{a})$ dadas como describimos en la sección 4.3.4.

La mezcla de los modelos M-I y M-II la efectuamos usando el paradigma Bayesiano siguiendo la idea de Draper (1995) de mezclar modelos con características y formas estructurales distintas. Asignamos inicialmente un peso p(M-I) y p(M-II) a cada modelo respectivamente, de manera que p(M-I) y p(M-II) sean mayores a cero y p(M-I)+p(M-II) = 1. Denotamos éstos como pesos y no como probabilidades, como en el esquema usual de mezcla Bayesiana de modelos, ya que consideramos que ninguno de estos modelos corresponde al modelo verdadero, i.e. usamos el enfoque $\mathcal{M}$-abierto, aunque usamos este esquema para determinar los pesos en la mezcla de manera que éstos dependan de la naturaleza de los datos, y así formar un *super-modelo* semiparamétrico. Suponiendo que $\boldsymbol{y}_T$ es la trayectoria observada del proceso, los pesos finales de cada modelo están dados por la regla de Bayes como

$$p(\text{M-I}|\boldsymbol{y}_T) \propto p(\boldsymbol{y}_T|\text{M-I})p(\text{M-I}) \tag{4.52}$$

$$p(\text{M-II}|\boldsymbol{y}_T) \propto p(\boldsymbol{y}_T|\text{M-II})p(\text{M-II}) \tag{4.53}$$

donde $p(\boldsymbol{y}_T|\text{M})$ denota la verosimilitud integrada del modelo M, y la constante de proporcionalidad dada por $p(\boldsymbol{y}_T) = p(\boldsymbol{y}_T|\text{M-I})p(\text{M-I}) + p(\boldsymbol{y}_T|\text{M-II})p(\text{M-II})$.

Utilizando (4.52) y (4.53) obtenemos la distribución predictiva final para $\boldsymbol{Y}_f'$, para

$k \geq 1$, como

$$p(\boldsymbol{y}_{T+1:T+k}|\boldsymbol{y}_T) = p(\boldsymbol{y}_{T+1:T+k}|\text{M-I}, \boldsymbol{y}_T)p(\text{M-I}|\boldsymbol{y}_Y) + p(\boldsymbol{y}_{T+1:T+k}|\text{M-II}, \boldsymbol{y}_T)p(\text{M-II}|\boldsymbol{y}_T). \quad (4.54)$$

Es imposible obtener una expresión cerrada para esta distribución, pero podemos generar muestras de ésta mediante aproximaciones numéricas de Monte Carlo para las verosimilitudes integradas de cada modelo, y muestras de la distribución predictiva de cada modelo a través del método de Monte Carlo vía cadenas de Markov de las distribuciones finales de cada modelo descritos en los algoritmos 2 y 4, de este capítulo.

En ambos modelos hemos asignado distribuciones iniciales propias y difusas para garantizar que las verosimilitudes integradas de cada modelo existan. éstas son imposibles de calcular de manera cerrada, e inclusive en algunos casos, los estimadores de Monte Carlo directos pueden ser inestables, ya que al muestrear de la distribución inicial directamente, podemos dejar de explorar regiones del espacio parametral donde la verosimilitud tiene más peso. Para eliminar este problema, en el caso del modelo M-I, elegimos aproximar su verosimilitud integrada mediante el estimador por importancia (vea la sección 2.3.1), eligiendo el *kernel* de la distribución final del modelo como la función de importancia, i.e. $p(\boldsymbol{y}_T|I(k), \boldsymbol{\omega}_k, \tau)\pi(I(k), \boldsymbol{\omega}_k, \tau)$. Así, el estimador por importancia de la verosimilitud integrada, para un tamaño de muestra N, es

$$p(\boldsymbol{y}_T|\text{M-I}) \approx \left[\frac{1}{N}\sum_{m=1}^{N} \frac{1}{p(\boldsymbol{y}_T|I(k^{(m)}), \boldsymbol{\omega}_{k^{(m)}}^{(m)}, \tau^{(m)})}\right]^{-1}. \quad (4.55)$$

Por otro lado, tenemos que para el modelo M-II la verosimilitud integrada puede expresarse como

$$p(\boldsymbol{y}_T|\text{M-II}) = \int\int \left\{\prod_{t=1}^{T} 1/\sigma_t Pr\left(\varepsilon_t|\boldsymbol{\beta}, \boldsymbol{a}\right)\right\} \pi(\boldsymbol{\beta})\pi(\boldsymbol{a})d\boldsymbol{\beta}d\boldsymbol{a}, \quad (4.56)$$

donde $\varepsilon_t = (y_t - \boldsymbol{x}_t'\boldsymbol{\beta})/\sigma_t$ para $t = 1, \ldots, T$, y

$$Pr\left(\varepsilon_t|\boldsymbol{\beta}, \boldsymbol{a}\right) = \int F\left(\varepsilon_t|\boldsymbol{\beta}, \boldsymbol{a}\right) dPT(F), \quad (4.57)$$

con la sucesión $\{\sigma_t\}_{t=1}^{T}$ en (4.56) calculada recursivamente mediante los valores fijos de $(\boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{y}_T)$, y $\pi(\boldsymbol{\beta})$ y $\pi(\boldsymbol{a})$ las distribuciones iniciales para $(\boldsymbol{\beta}, \boldsymbol{a})$. La probabilidad en (4.57)

no es más que la probabilidad de que ε_t pertenezca a un elemento del árbol de partición en el nivel L del árbol de Pólya, que puede calcularse de manera cerrada a partir los pesos $\boldsymbol{\rho}_{PT} = \{\rho_j, j = 1, \ldots, 2^L\}$ calculadas con base en (4.36).

Así, una aproximación de Monte Carlo para (4.56) se obtiene de una muestra de tamaño N de $\pi(\boldsymbol{\beta})$ y $\pi(\boldsymbol{a})$, denotadas por $\{(\boldsymbol{\beta}^{(n)}, \boldsymbol{a}^{(n)}), n = 1, \ldots, N\}$, como

$$p(\boldsymbol{y}_T|\text{M-II}) \approx \frac{1}{N}\sum_{n=1}^{N}\left\{\prod_{t=1}^{T} 1/\sigma_t Pr(\varepsilon_t|\boldsymbol{\beta}^{(n)}, \boldsymbol{a}^{(n)})\right\}. \tag{4.58}$$

El esquema general de muestreo de la distribución predictiva final del *super-modelo* semiparamétrico (mezcla) es el siguiente.

Algoritmo 5: Mezcla de Modelos Semiparamétricos

Primero corremos de manera paralela dos cadenas de Markov de los modelos M-I y M-II siguiendo los algoritmos 2 y 4 respectivamente.

1. Calculamos de manera paralela las aproximaciones de las verosimilitudes integradas de los dos modelos dadas por (4.55) y (4.58) respectivamente.

2. Obtenemos una muestra $u \sim U(0, 1)$.

   (a) Si $u <$ en (4.52), obtenemos una muestra con remplazo de la muestra predictiva del modelo M-I.

   (b) De lo contrario, obtenemos una muestra con remplazo de la muestra predictiva del modelo M-II.

La implementación de la mezcla y de los modelos semiparamétricos individuales descritos en este capítulo es computacionalmente demandante, particularmente la implementación del modelo con árboles de Pólya. Para obtener una mayor estabilidad en la cadena para el modelo basado en onduletas o bases radiales, particularmente en los parámetros de traslación y dilatación, se requiere de un periodo de calentamiento significativamente largo.

La mezcla de estos dos modelos semiparamétricos servirá como la aproximación flexible del modelo "verdadero" para la implementación del criterio predictivo de selección $\mathcal{M}$-abierto descrito al final del capítulo 3. En el siguiente capítulo aplicaremos esta mezcla de

modelos semiparamétricos para comparar y seleccionar diferentes modelos paramétricos, algunos de éstos pertenecientes a la clase de modelos dinámicos lineales, para modelar la serie del Índice Metropolitano de Calidad del Aire de la Ciudad de México. Los modelos que consideramos en ese capítulo modelan diferentes características de la serie, tanto en su parte sistemática como en su componente aleatorio, todos éstos con formas estructurales distintivas.

Capítulo 5

Aplicación

Al final del capítulo 3 presentamos un criterio de selección de modelos que denotamos $\mathcal{M}$-semiabierta. El criterio que proponemos para el análisis de series de tiempo queda completamente especificado con el modelo semiparamétrico del capítulo 4. Éste criterio puede emplearse solamente para la comparación y selección de modelos para series de tiempo escalares. Su aplicación para series de tiempo con otras características, e.g. discretas, de conteo, de proporciones, nominales, categóricas, etc, es directa salvo por la especificación del modelo semiparamétrico flexible que fungiría como modelo juez de los modelos por comparar. En este capítulo ilustramos el criterio predictivo $\mathcal{M}$-semiabierta propuesto en el contexto de la modelación del Índice Metropolitano de la Calidad del Aire (IMECA) de la Ciudad de México.

Con el IMECA las autoridades ambientales y los ciudadanos en general monitorean el comportamiento de los principales elementos contaminantes del aire. El índice se construye con base en las mediciones en cinco zonas geográficas de la ciudad (norte, sur, este, oeste y centro), y para cada una de éstas se construye un índice. Dada la naturaleza geográfica-espacial de las mediciones, existe una alta dependencia de los índices en estas zonas. En este trabajo sólo consideramos la modelación del IMECA correspondiente a la zona centro.

En la primera sección del capítulo describimos brevemente los elementos contaminantes

que son medidos a través del IMECA, así como su construcción. En la segunda sección proponemos cinco modelos completamente paramétricos para el análisis y predicción del índice en la zona centro. Cada modelo tiene elementos interpretativos característicos, además de poseer una capacidad predictiva aceptable. Al final del capítulo presentamos el planteamiento y solución del problema de selección de modelos con el criterio predictivo $\mathcal{M}$-semiabierta. Para efectos prácticos el espacio de acciones que consideramos sólo incluye a las distribuciones predictivas de los modelos contendientes un paso adelante, aunque también es posible considerar dentro de esta clase a todas las posibles mezclas de estas distribuciones con un considerable incremento en el costo computacional para su instrumentación, aunado a las complicaciones interpretativas de los modelos resultantes.

Algunos de los modelos paramétricos propuestos requieren de sutilezas computacionales para el cálculo de las distribuciones finales de algunos de sus parámetros. En estos casos extendemos la descripción del modelo con algunos pasos relevantes para el cálculo de estas distribuciones finales.

5.1 Índice Metropolitano de Calidad del Aire

Con el desarrollo industrial y tecnológico, el ser humano se ha convertido en un fuerte productor de elementos contaminantes ambientales, a los cuales también se encuentra expuesto. Aunque la generación de estos elementos es generalmente atribuida al desarrollo industrial, existen diferentes e importantes factores naturales de generación de contaminación, como son: polvo, humo y gases generados por la erupción o emanaciones volcánicas e incendios naturales, o erosión del suelo. Generalmente los habitantes de las grandes ciudades son los más expuestos a estos contaminantes, particularmente en relación a la contaminación del aire. Esta se debe a la acción conjunta de diferentes agentes. Particularmente en la Ciudad de México se observa la acción conjunta de agentes naturales y artificiales. Los principales agentes artificiales son las emanaciones generadas por la combustión de automóviles, aviones y de la industria y construcción. Los agentes naturales son principalmente la erosión y el humo generado por la quema de terrenos y emanaciones volcánicas. Por otro lado, la localización geográfica de la ciudad propicia que exista un

estancamiento del aire, lo que aunado a los altos niveles de emanación de contaminantes combustibles, propicia que se generen niveles altos de ozono suspendido en el ambiente. La exposición a estos contaminantes, principalmente en niños y personas de edad avanzada, puede generar severos problemas en el sistema inmunológico y respiratorio.

Para monitorear el comportamiento agregado de los principales contaminantes, las autoridades ambientales de la ciudad decidieron definir un índice conocido como Índice Metropolitano de la Calidad del Aire (IMECA) que monitorea el nivel agregado de los principales contaminantes: Monóxido de Carbono (CO), Ozono (O3), Bióxido de Nitrógeno (NO2), Bióxido de Azufre (SO2) y Partículas Suspendidas (PST y PM10). Los efectos dañinos para la salud de las personas expuesta a de estos contaminantes son variados; por ejemplo el CO afecta al sistema nervioso central y afecta el sistema cardiaco y pulmonar; el NO2 puede irritar los pulmones y afectar el sistema respiratorio y cardiovascular; el O3, que afecta principalmente a los niños expuestos, propicia irritación ocular y agrava las enfermedades respiratorias; el SO2 irrita el sistema ocular y el tracto respiratorio, incluso en algunos casos agrava padecimientos crónicos como la bronquitis crónica y efisema pulmonar; y las partículas suspendidas agravan los síntomas del asma y pueden almacenarse en los pulmones, lo que puede produce padecimientos severos. El IMECA se define como

$$\text{IMECA} = \max\left\{I_1 CO, I_2 O3, I_3 NO2, I_4 PST, I_5 PM10, I_6 SO2\right\},$$

donde los coeficientes I_i's denotan ponderadores para cada contaminante.

En la figura 5.1 graficamos la trayectoria diaria observada del IMECA (centro) en el periodo del 2 de agosto al 18 de diciembre de 1998. Visualmente podemos identificar un comportamiento semi-oscilatorio con algunos cambios de nivel en el mes de octubre del mismo año.

Los modelos paramétricos que proponemos son básicamente extensiones de los de los modelos autorregresivos lineales, en la modelación del nivel o la forma distribucional de los errores. En este trabajo no hemos considerado modelos que incorporen información adicional relevante para el comportamiento de la serie. En la siguiente sección presentamos una descripción general de los modelos propuestos. Para efectos ilustrativos sólo consideramos un modelo representante de cada clase, éste es elegido dentro de una gama

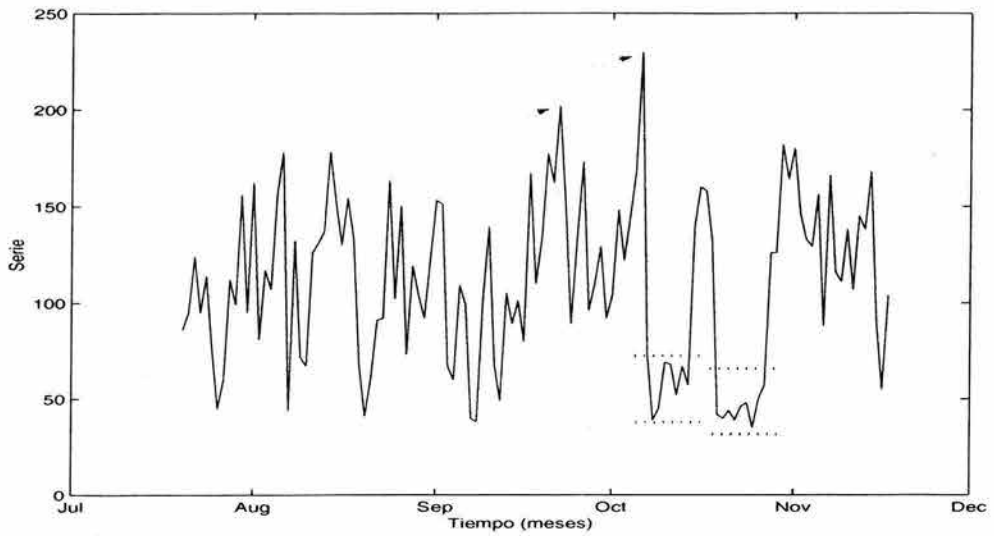


Figura 5.1: Serie del Índice Metropolitano de Calidad del Aire (Zona Centro)

representativa de modelos como el modelo que maximice alguna de las medidas de diagnóstico predictivo de modelos que describimos en el apéndice C.

5.1.1 Modelo 1

El primer modelo que consideramos forma parte de la familia de modelos autorregresivo con parámetros cambiantes en el tiempo (TVAR por sus siglas en inglés). Los modelos en esta clase no son más que modelos autorregresivos locales en el tiempo, con la localidad manifestada en términos de los parámetros. El cambio o evolución de los parámetros se modela a través de una caminata aleatoria Gaussiana, y es en estos cambios donde obtienen la flexibilidad para capturar el comportamiento de una gran variedad de procesos en la práctica. A su vez, estos modelos pueden ser vistos como un caso particular de los *modelos de espacio de estados* o *modelos dinámicos lineales* (e.g. West y Harrison (1997)).

La formulación de los modelo TVAR para una serie de tiempo escalar $\{y_t\}$ y un orden p fijo es

$$\begin{aligned} y_t &= \boldsymbol{y}'_{t-1}\boldsymbol{\phi}_t + \varepsilon_t \\ \boldsymbol{\phi}_t &= \boldsymbol{\phi}_{t-1} + \boldsymbol{\omega}_t \\ \tau_t &= \frac{\eta_t}{\beta}\tau_{t-1} \end{aligned} \tag{5.1}$$

donde $\boldsymbol{y}_{t-1} = (y_{t-1}, \ldots, y_{t-p})'$, $\{\varepsilon_t\}$ es una sucesión de ruido Gaussiano, con $\varepsilon_t \sim N(\varepsilon_t|0, \tau_t^{-1})$, y $\boldsymbol{\phi}_t = (\phi_{1,t}, \ldots, \phi_{p,t})'$ es el vector de autorregresión para el tiempo t, $\{\boldsymbol{\omega}_t\}$ es una sucesión independiente de vectores aleatorios con distribución Gaussiana, un vector de medias nulo y una matriz de covarianzas $\tau^{-1}\boldsymbol{W}_t$, con $\boldsymbol{W}_t$ conocida, $\{\eta_t\}$ es una sucesión de variables aleatorias independientes de $\{\tau_t\}$ tales que $\eta_t \sim \text{Beta}(\eta_t|\beta n_{t-1}/2, (1-\beta n_{t-1}/2))$, y $0 < \beta \leq 1$ es un factor de descuento fijo (West y Harrison, 1997, sección 10.8).

Para no sobreparametrizar el modelo, se modela la evolución en la matriz de covarianzas $\boldsymbol{W}_t$ de manera determinista usando otro factor de descuento δ, con $0 < \delta \leq 1$, fijo (West y Harrison, 1997, sección 6.3). Consideramos que junto con el orden del modelo, p, los factores de descuento, β y δ, forman parte de la especificación del modelo. Si asignamos una distribución inicial para $(\boldsymbol{\phi}_0, \tau_0)$ perteneciente a la familia conjugada Normal-Gamma, es posible definir un procedimiento secuencial conjugado de actualización que es bastante simple de instrumentar si expresamos este modelo como un caso particular de un modelo dinámico lineal. Los detalles sobre el proceso de actualización se encuentran en West y Harrison (1997, capítulo 4).

En el caso particular del IMECA, después de monitorear el poder predictivo para diferentes modelos usando las medidas de diagnóstico descritas en el apéndice C, consideramos el modelo TVAR de orden $p = 13$, con los factores de descuento $\delta = 0.99$ y $\beta = 0.985$. Los hiperparámetros de la distribución inicial sobre $(\boldsymbol{\phi}_0, \tau_0)$ son un vector de medias nulo p-dimensional, una matriz de covarianzas $10\boldsymbol{I}_{13}$, y $a_0 = b_0 = 0.01$. Realizando los cálculos correspondientes se tiene que la distribución predictiva del IMECA para el tiempo $T+1$ es

$$p_1(y_{T+1}|\boldsymbol{y}_T) = St(y_{T+1}|107.6654,\ 18.5690,\ 107). \tag{5.2}$$

5.1.2 Modelo 2

El segundo modelo que consideramos pertenece a la clase de modelos autorregresivos con saltos estructurales en nivel, que fue propuesta por McCulloch y Tsay (1993). Estos modelos se componen de una parte que modela cambios estructurales en nivel y de un componente autorregresivo estático para modelar el remanente de la serie sin tendencia.

Para una serie de tiempo $\{y_t\}$ el modelo se formula como

$$\begin{aligned} y_t &= \mu_t + x_t, \\ \mu_t &= \mu_{t-1} + \delta_t \beta_t, \\ x_t &= \boldsymbol{x}'_{t-1}\boldsymbol{\phi} + \varepsilon_t, \end{aligned} \tag{5.3}$$

donde $\{\mu_t\}$ denota el componente de tendencia de la serie y $\{x_t\}$ el residuo de la serie sin tendencia. El nivel de la serie es afectado por $\{\delta_t\}$ que es una sucesión de variables aleatorias indicadoras de la presencia de cambios estructurales en el tiempo t, y por $\{\beta_t\}$ que denota la sucesión de variables aleatorias que modelan la magnitud del cambio estructural. En cada tiempo $\delta_t \sim$ Bernoulli(ρ), con $\rho \in (0,1)$ desconocido.

Por otro lado, el componente $\{x_t\}$ sigue un proceso autorregresivo lineal Gaussiano estático, donde $\boldsymbol{x}_{t-1} = (x_{t-1}, ..., x_{t-p})'$, $\boldsymbol{\phi} = (\phi_1, ..., \phi_p)$, con p el orden del proceso, y donde $\{\varepsilon_t\}$ es un proceso de ruido aleatorio Gaussiano con precisión τ desconocida. Se supone que los procesos $\{\delta_t\}$, $\{\beta_t\}$ y $\{\varepsilon_t\}$ son interior y mutuamente independientes. La distribución inicial sobre ρ es Beta($\rho|\gamma_1, \gamma_2$), para ciertos parámetros γ_1 y γ_2 positivos.

Supongamos que $\boldsymbol{y_T}$ denota la trayectoria observada del proceso hasta el tiempo T. Las cantidades de interés desconocidas del modelo son $(\boldsymbol{\mu}, \boldsymbol{\phi}, \boldsymbol{\delta}, \boldsymbol{\beta}, \sigma^2, \rho)$, con $\boldsymbol{\mu} = (\mu_p, \mu_{p+1}, \ldots, \mu_T)'$, $\boldsymbol{\delta} = (\delta_{p+1}, \ldots, \delta_T)'$ y $\boldsymbol{\beta} = (\beta_{p+1}, \ldots, \beta_T)'$. La sucesión $\{\mu_t\}$ se calcula recursivamente condicional en $\boldsymbol{\delta}$ y $\boldsymbol{\beta}$ de acuerdo a la relación (5.3) y el valor inicial $\mu_p = y_p$. La distribución final conjunta de estos parámetros es desconocida de manera analítica. Para generar muestras de esta distribución seguimos el esquema de muestreo de Gibbs propuesto por McCulloch y Tsay (1993).

Al vector $\boldsymbol{\phi}$ le asignamos una distribución inicial Normal multivariada con un vector de medias nulo y una matriz de covarianzas $\boldsymbol{C}$ conocida de manera que esta distribución sea difusa. Por conveniencia analítica se le asigna una distribución una distribución Beta(γ_1,γ_2) a la masa de probabilidad ρ, con γ_1 y γ_2 constantes conocidas. En esta especificación debemos considerar no dejar abierta la posibilidad de que el modelo detecte cambios en nivel con gran facilidad, de esta forma podemos restringir inicialmente que γ_1 sea significativamente menor que γ_2 de manera que la probabilidad de cambio o choque aleatorio sea pequeña, y el modelo detecte sólo cambios significativos en el nivel. La

sucesión $\{\beta_t\}$ es una sucesión de variables aleatorias independientes con distribución inicial Normal centrada en cero y con varianza conocida. También por conveniencia analítica, a la varianza σ^2 le es asignada una distribución inicial Gamma-Inversa$(a/2, b/2)$, con a y b valores reales conocidos. McCulloch y Tsay (1993) desarrollaron las distribuciones finales condicionales completas para cada parámetro de manera analítica cerrada, haciendo posible instrumentar el muestreador de Gibbs de manera relativamente simple.

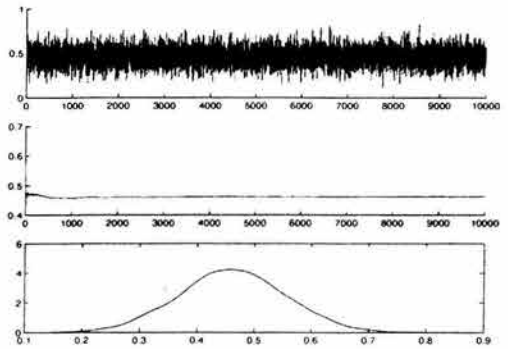
La especificación del modelo que consideramos para el IMECA está determinada de manera que maximiza la cantidad $D_{1,T}$ propuesta por Gelfand *et al.* (1992) (vea el apéndice C). Consideramos que el proceso autorregresivo $\{x_t\}$ puede estar entre los ordenes 1 y 15. Trabajamos con una matriz de covarianzas inicial para $\boldsymbol{\phi}$ igual a $100\boldsymbol{I}_3$, y para σ^2 asignamos los valores de $a = 0.01$ y $b = 0.01$. En la distribución inicial para $\rho = P(\delta_t = 1)$ asignamos los hiperparámetros de $\gamma_1 = 1$ y $\gamma_2 = 70$. Elegimos una varianza inicial para la sucesión de magnitudes de choques, $\{\beta_t\}$, igual a 20, para controlar que la magnitud de los choques no sea explosiva. El orden del proceso $\{x_t\}$ que maximiza esta cantidad es $p = 3$. Empíricamente encontramos que las distribuciones finales no son susceptibles a cambios discretos en la especificación de las distribuciones iniciales.

En la figura 5.2 graficamos las distribuciones finales marginales de los parámetros asociados al componente autorregresivo latente con 10,000 observaciones, después de un periodo inicial de calentamiento de la cadena de 5,000 iteraciones. En la gráfica superior de cada panel graficamos la traza (trayectoria) de la cadena marginal de cada parámetro, en la gráfica central presentamos los promedios ergódicos de la cadena.

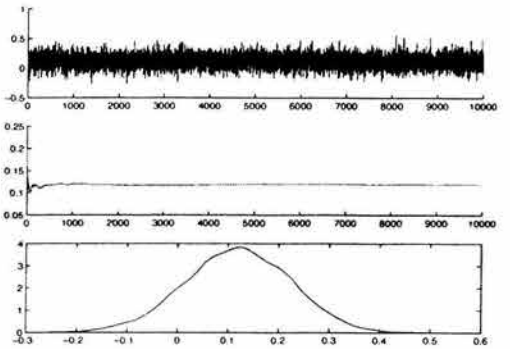
La densidad predictiva para el tiempo $T+1$ es reconstruida mediante la aproximación por *kernels* como (vea el apéndice B.3):

$$p_2(y_{T+1}|\boldsymbol{y}_T) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{b_n}K\left((y_{T+1} - y_{T+1}^{(n)})/b_n\right), \tag{5.4}$$

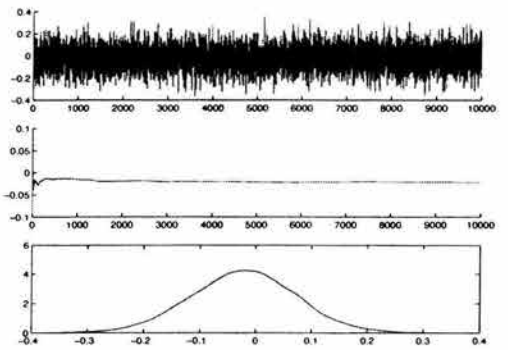
con $\{y_{T+1}^{(n)}\}$ una muestra de tamaño N de la densidad predictiva para el tiempo $T+1$ obtenida mediante el muestreador de Gibbs, $b_n = 10n^{-1/5}$ para $n = 1, 2, ..., N$ y $N = 10$ mil. En este caso elegimos reconstruir la densidad mediante el *kernel* optimal (vea el cuadro B.1).



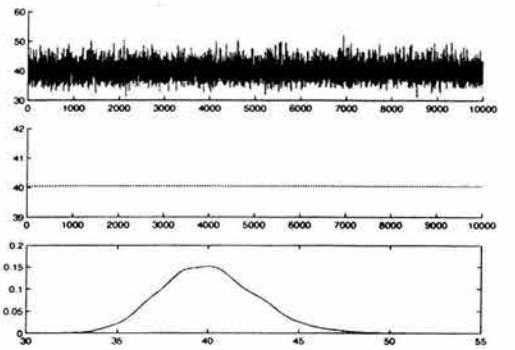
(a) ϕ_1



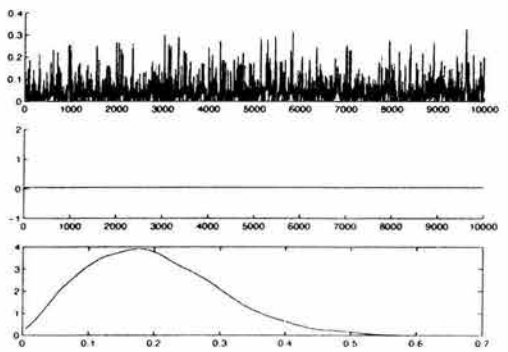
(b) ϕ_2



(c) ϕ_3



(d) σ



(e) ρ

Figura 5.2: Distribuciones finales marginales de los parámetros del componente autorregresivo latente, $\{x_t\}$, en el Modelo 2 (a-d), y la sucesión de probabilidades de aparición de choques (e).

5.1.3 Modelo 3

En este modelo consideramos un componente más robusto en la distribución del proceso. El modelo que elegimos es una extensión del modelo autorregresivo Gaussiano usual, incorporando en este caso una distribución α-estable en los errores (vea el apéndice B.2 para una definición de las distribuciones α-estables). El modelo que usamos es un caso particular del modelo lineal con errores α-estables que desarrollaron Godsill y Kuruoğlu (1999). Las distribuciones α-estables permiten modelar una amplia gama de fenómenos aleatorios, aunque su utilización es limitada debido a que salvo en casos muy particulares no se conoce la función de densidad de manera cerrada. Por otro lado, estas distribuciones tienen la ventaja que generalizan el Teorema Central de Límite, el cual sustenta la suposición inicial para modelar los errores aleatorios con distribuciones Normales.[1] Una exposición teórica respecto a esta familia de distribuciones se encuentra en Samorodnitsky y Taqqu (1994), y en el apéndice B.2 se da una definición de estas distribuciones junto con algunas propiedades generales que son útiles para este trabajo.

Las distribuciones estables son unimodales, y en general no tienen segundo momento finito, un factor adicional que ha limitado su uso en la práctica. Están caracterizadas por cuatro parámetros, un parámetro α que denota al exponente característico y determina el comportamiento de las colas de la distribución, un parámetro δ de que determina el sesgo de la distribución, y un parámetro de localización y otro de escala. Así como existe la consideración que los errores aleatorios tengan media (o mediana) cero en su distribución, también se tiene que en general la distribución sea simétrica respecto a este valor de centralidad. En este caso la especificación de las distribuciones estables simétricas se torna simple, pues pueden ser expresadas como una mezcla continua de escalas de distribuciones Normales (vea el apéndice B.2). Godsill y Kuruoğlu (1999) emplearon este resultado simplificando así el análisis Bayesiano de estas distribuciones pues es posible expresar el modelo como una extensión del modelo Gaussiano usual.

[1] Las distribuciones estables se definen como una clase de distribuciones de localización y escala que es cerrada ante convoluciones, además tienen la característica de ser una clase de distribuciones límite de sumas de variables aleatorias independientes. Para una descripción detallada véase Molina Escobar (2001).

El modelo que consideramos aquí es un caso particular del modelo lineal propuesto por Godsill y Kuruoğlu (1999). Básicamente es un modelo autorregresivo lineal con errores α-estables simétricos, y la representación alternativa es de la forma

$$y_t = \boldsymbol{y}'_{t-1}\boldsymbol{\beta} + \sigma\lambda_t^{1/2}\varepsilon_t, \tag{5.5}$$

donde $\boldsymbol{y}_{t-1} = (y_{t-1}, ..., y_{t-p})'$, $\{\varepsilon_t\}$ es una sucesión de ruido blanco Gaussiano, $\sigma > 0$ es un parámetro de escala desconocido, $\boldsymbol{\beta}$ es un vector de coeficientes de autorregresión desconocido, y $\{\lambda_t\}$ es una sucesión de ruido α-estable estrictamente positivo con exponente característico $\alpha/2$, con $\alpha \in (0, 2]$ y distribución $S_{\alpha/2,1}(0, 1)$. Suponemos que las sucesiones $\{\varepsilon_t\}$ y $\{\lambda_t\}$ son interior y mutuamente independientes. Mediante un proceso de marginalización respecto a λ_t en cada tiempo t obtenemos un proceso de ruido aleatorio α-estable simétrico, con exponente característico α.

Por simplicidad de instrumentación Godsill y Kuruoğlu (1999) consideraron al exponente característico dentro de la determinación del modelo, y no como un parámetro desconocido adicional. Realizar el análisis Bayesiano de este modelo incluyendo al parámetro α dentro de la estructura de incertidumbre puede ser problemático y computacionalmente costoso. Buckle (1995) desarrollo un sofisticado esquema de muestreo en el que se puede realizar el análisis Bayesiano conjunto de los cuatro parámetros que caracterizan a las distribuciones estable. Por simplicidad de instrumentación y exposición en este trabajo sólo consideramos la el modelo parcialmente especificado propuesto que Godsill y Kuruoğlu (1999) propusieron. Así, suponemos que la especificación del modelo incluye el orden de autorregresión p y el exponente característico α. De esta forma los parámetros desconocidos del modelo son $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\lambda}_T)$, con $\boldsymbol{\lambda}_T = (\lambda_1, \ldots, \lambda_T)'$ el vector de componentes latentes del modelo.

Bajo el supuesto de que el proceso $\{y_t\}$ es α-estable simétrico, con exponente característico $\alpha \in (0, 2]$, los parámetros de localización y escala estarían determinados por el valor de α (vea las propiedades de las distribuciones estables en el apéndice B.2). Godsill y Kuruoğlu (1999) calcularon las distribuciones finales condicionales completas para estos parámetros, con las que es posible instrumentar el muestreador de Gibbs para generar muestras de las distribuciones finales de los parámetros desconocidos.

El muestreador propuesto por Godsill y Kuruoğlu (1999) es como sigue. Supongamos que $\boldsymbol{y}_T$ denota la trayectoria del proceso hasta el tiempo T, y que empleamos una estructura independiente en la distribución inicial de los parámetros, i.e. $\pi(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\lambda}_T) = \pi(\boldsymbol{\beta}, \sigma^2)\pi(\boldsymbol{\lambda}_T)$. La distribución inicial sobre los parámetros λ_t es α-estable, para $t = 1, \ldots, T$. Condicional en $\boldsymbol{\lambda}_T$, se tiene una estructura lineal Gaussiana, entonces es conveniente analíticamente asignar una distribución inicial Gaussiana a $\boldsymbol{\beta}$ con un vector de medias nulo y una matriz de covarianzas $\boldsymbol{C}$, y a σ^2 una distribución Gamma-Inversa$(a/2, b/2)$ con parámetros a y b números reales positivos. Condicional en $\boldsymbol{\lambda}_T$ el modelo (5.5) es un caso particular de un modelo de regresión lineal heteroscedástico

$$\boldsymbol{y}_T = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{H}\boldsymbol{\varepsilon}_T, \tag{5.6}$$

donde $\boldsymbol{X}$ es la matriz de autorregresión de rango completo cuyos vectores renglón corresponden a las observaciones del proceso de autorregresión $\{\boldsymbol{x}_t\}$, $\boldsymbol{\varepsilon}_T = (\varepsilon_1, \ldots, \varepsilon_T)'$ es el vector de errores aleatorios, y la matriz $\boldsymbol{H} = \text{diag}(\lambda_1^{1/2}, \ldots, \lambda_T^{1/2})$ contiene en su diagonal a los componentes heteroscedásticos latentes del proceso. La distribución final condicional completa para $\boldsymbol{\beta}$ es $\text{N}(\boldsymbol{\beta}|\boldsymbol{\mu}, \sigma^2\boldsymbol{C}_1)$, con $\boldsymbol{C}_1 = (\boldsymbol{X}'(\boldsymbol{H}\boldsymbol{H}')^{-1}\boldsymbol{X} + \sigma^2\boldsymbol{C}^{-1})^{-1}$ y vector de medias $\boldsymbol{\mu} = \boldsymbol{C}_1\boldsymbol{X}'(\boldsymbol{H}\boldsymbol{H}')^{-1}\boldsymbol{y}_T$; mientras que para σ^2 se tiene una distribución condicional completa Gamma-Inversa con parámetros $a_1 = (a + T)/2$ y $b_1 = (b + (\boldsymbol{y}_T - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{H}\boldsymbol{H}')^{-1}(\boldsymbol{y}_T - \boldsymbol{X}\boldsymbol{\beta}))/2$. Así que en estas etapas del muestreador de Gibbs el proceso de muestreo es simple.

En la tercera etapa del muestreador de Gibbs, necesitamos obtener muestra del componente latente $\boldsymbol{\lambda}_T$. Como $\{\lambda_t\}$ es una sucesión de variables aleatorias se tiene que para cada tiempo t, el componente λ_t depende solamente de $\boldsymbol{y}_T$ y $\boldsymbol{\beta}$ a través del error estimado $\varepsilon_t = y_t - \boldsymbol{x}_t'\boldsymbol{\beta}$, y la distribución final condicional completa para cada λ_t es de la forma

$$\begin{aligned} p(\lambda_t|\boldsymbol{\beta}, \boldsymbol{y}_T, \sigma_2) &= p(\lambda_t|\varepsilon_t, \sigma^2) \\ &\propto N(\varepsilon_t|0, 2\lambda_t\sigma^2) f_{\alpha/2,1}(\lambda_t|0, 1), \end{aligned} \tag{5.7}$$

donde $N(y)$ denota la función de densidad Normal evaluada en y, y $f_{\alpha,\delta}(x|\mu, \sigma)$ denota la función de densidad de la distribución estable $S_{\alpha,\delta}(\mu, \sigma)$ evaluada en x. Generalmente es imposible obtener una expresión analítica cerrada de la relación (5.7), principalmente

porque no existe una representación analítica cerrada de la distribución estable inicial sobre λ_t, salvo en la forma de expansiones infinitas basadas en su función característica. Las muestras de la distribución condicional completa de cada λ_t se pueden obtener mediante un paso de M-H usando la distribución inicial como distribución instrumental, en cuyo caso la probabilidad de aceptación del movimiento del estado λ_t al estado λ_t' esta dada por $\alpha(\lambda_t, \lambda_t') = \text{mín}(1, N(\varepsilon_t|0, 2\lambda_t'\sigma^2)/N(\varepsilon_t|0, 2\lambda_t\sigma^2))$. Godsill y Kuruoğlu (1999) propusieron generar muestras de (5.7) mediante muestreo por rechazo, generando muestras exactas de la distribución final, aunque el proceso iterativo hace significativamente más lenta la ejecución del algoritmo. En nuestras aplicaciones consideramos el muestreo con el paso de M-H debido a que explora satisfactoriamente el espacio de interés y las probabilidades de aceptación de los movimientos son relativamente altas. Para obtener muestras de $\lambda \sim S_{\alpha/2,1}(0,1)$ utilizamos el algoritmo de Chambers-Mallows-Stuck (Chambers *et al.*, 1976; Weron, 1996).

Para la especificación del modelo consideramos que los posibles ordenes de autorregresión estén entre 1 y 15, mientras que para el exponente característico sólo consideramos los valores en el intervalo (1.8, 1) con un espaciamiento de 0.025 entre cada valor considerado. Instrumentamos el modelo usando una matriz de covarianzas inicial para $\boldsymbol{\beta}$ igual a $100\boldsymbol{I}_p$ y valores de $a = 0.01$ y $b = 0.01$ en la distribución inicial para σ^2.

El modelo representativo de esta clase es el que maximiza la cantidad $D_{2,T}$ propuesta por Gelfand *et al.* (1992) (vea el apéndice C). La determinación del modelo por este método es altamente costosa. El orden que optimiza esta cantidad es $p = 13$, con un exponente característico $\alpha = 1.875$. El exponente característico determina colas más pesadas el modelo Gaussiano usual, pero no demasiado, a su vez que brinda una mayor flexibilidad en la modelación de los coeficientes de autorregresión.

En la figura 5.3 mostramos las distribuciones finales de los parámetros del modelo considerando 10,000 observaciones después de un periodo inicial de calentamiento de 20,000 iteraciones. Dentro de cada panel, la gráfica superior muestra la traza (trayectoria) de la cadena, y la gráfica intermedia muestra el promedio ergódico en cada caso. Las densidades que se muestran en la parte inferior de cada panel corresponden a la reproducción por

kernels de la densidad final.

Nuestro interés es predecir el valor futuro del IMECA en el tiempo $T+1$. Usando la representación de distribuciones estables como mezclas de Normales, podemos expresar la densidad predictiva como

$$\begin{aligned} f_{\alpha,0}(y_{T+1}|\boldsymbol{y}_{T+1}) &= \int \cdots \int N(y_{T+1}|\boldsymbol{x}'_{T+1}\boldsymbol{\beta}, 2\lambda_{T+1}\sigma^2) f_{\alpha/2,1}(\lambda_{T+1}|0,1) d\lambda_{T+1} \\ &\quad \times \pi(\boldsymbol{\beta}, \sigma^2) d\boldsymbol{\beta} d\sigma^2, \end{aligned}$$

la cual podemos aproximar mediante la densidad predictiva Rao-Blackwellizada (Gelfand y Smith, 1990) dada por

$$p_3(y_{T+1}|\boldsymbol{y}_T) = \frac{1}{N}\sum_{i=1}^{N} N(y_{T+1}|\boldsymbol{x}'_{T+1}\boldsymbol{\beta}^{(i)}, 2\lambda_{T+1}^{(i)}(\sigma^{(i)})^2), \tag{5.8}$$

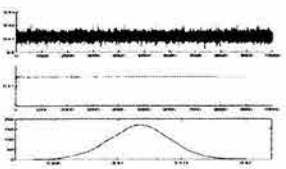
para una muestra de tamaño M de la distribución final de $(\boldsymbol{\beta}, \sigma^2)$ y los valores $\lambda_{T+1}^{(i)}$ muestras simuladas independientes de la distribución $S_{\alpha/2,1}(0,1)$, para $i = 1, \ldots, N$.

5.1.4 Modelo 4

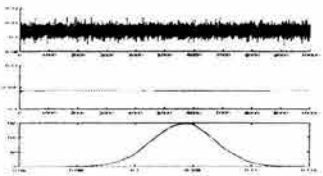
El cuarto modelo que consideramos pertenece a la clase de modelos dinámico con nivel cambiante en el tiempo y componente armónico descrito como (vea West y Harrison (1997)):

$$\begin{aligned} y_t &= \mu_t + \alpha_{1,t}\cos(2\pi t/\lambda_1) + \alpha_{2,t}\,\mathrm{sen}(2\pi t/\lambda_1) \\ &\quad + \alpha_{3,t}\cos(2\pi t/\lambda_2) + \alpha_{4,t}\,\mathrm{sen}(2\pi t/\lambda_2) + \varepsilon_t, \\ \mu_t &= \mu_{t-1} + \omega_{1,t}, \\ \boldsymbol{\alpha}_t &= \boldsymbol{\alpha}_{t-1} + \boldsymbol{\omega}_{2,t} \\ \tau_t &= \frac{\eta_t}{\beta}\tau_{t-1} \end{aligned} \tag{5.9}$$

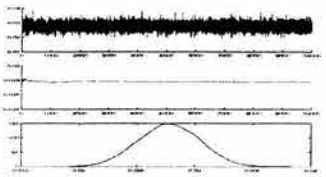
donde $\{\varepsilon_t\}$ es una sucesión de ruido Gaussiano con varianzas $\{\tau_t^{-1}\}$ cambiantes en el tiempo, $\{\mu_t\}$ es un proceso latente que mide el nivel de la serie de acuerdo a una caminata aleatoria Gaussiana, i.e. $\omega_{1,t}$ es una sucesión de variables aleatorias independientes con distribución Normal, media cero y varianza cambiante en el tiempo $\tau_{t-1}^{-1}W_{1,t}$, donde $\{W_{1,t}\}$



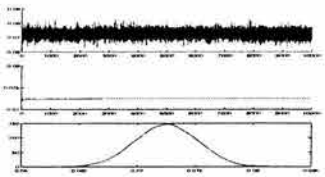
(a) β_1



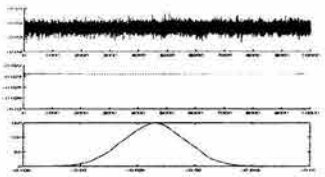
(b) β_2



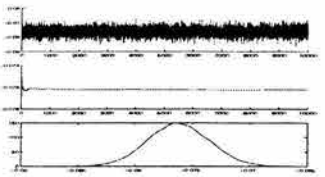
(c) β_3



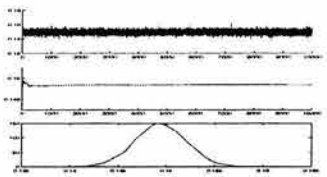
(d) β_4



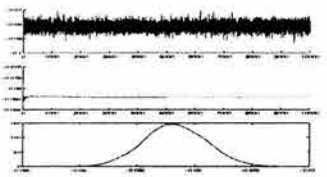
(e) β_5



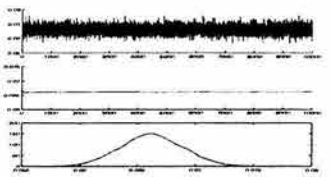
(f) β_6



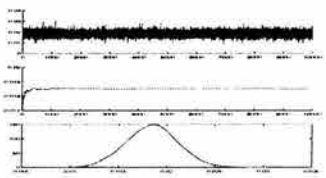
(g) β_7



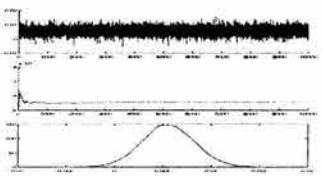
(h) β_8



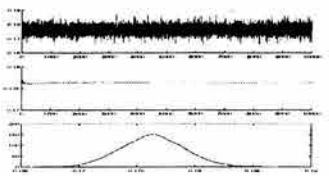
(i) β_9



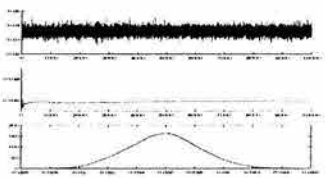
(j) β_{10}



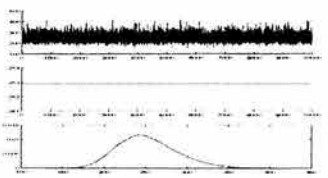
(k) β_{11}



(l) β_{12}



(m) β_{13}



(n) σ^2

Figura 5.3: Distribuciones finales de los parámetros del Modelo 3.

es un proceso determinista también cambiante en el tiempo. Los componentes λ_1 y λ_2 son constantes conocidas que denotan la longitud de periodo de los dos componentes armónicos de la serie $\{y_t - \mu_t\}$ considerándolas como parte de la especificación del modelo, y $\boldsymbol{\alpha}_t = (\alpha_{1,t}, \ldots, \alpha_{4,t})'$ son los coeficientes de regresión de los componentes armónicos que son cambiantes en el tiempo de acuerdo a una caminata aleatoria Gaussiana, i.e. $\{\boldsymbol{\omega}_{2,t}\}$ es una sucesión de vectores aleatorios independientes con distribución Normal multivariada, un vector de medias nulo y matrices de covarianzas cambiantes en el tiempo $\tau_{t-1}^{-1}\boldsymbol{W}_{2,t}$, con $\boldsymbol{W}_{2,t}$ una matriz cambiante el tiempo simétrica y positivo definida.

Suponemos que las varianzas $\{\tau_t^{-1}\}$ evolucionan de acuerdo a una caminata aleatoria multiplicativa en términos de la precisión como en el modelo 1 (vea la sección 5.1.1). Mientras que la evolución de las matrices de covarianzas $\boldsymbol{W}_t = \text{diagbloq}(W_{1,t}, \boldsymbol{W}_{2,t})$ se genera de manera determinista de acuerdo un factor de descuento $0 < \delta \leq 1$ fijo y conocido (West y Harrison, 1997, sección 6.3).

Este modelo corresponde a un caso particular de un modelo dinámico lineal (West y Harrison, 1997) con un vector de de estados $\boldsymbol{\theta}_t = (\mu_t, \alpha_{1,t}, \ldots, \alpha_{4,t})'$, un vector de observación $\boldsymbol{F}_t = (1, \cos(2\pi t/\lambda_1), \text{sen}(2\pi t/\lambda_1), \cos(2\pi t/\lambda_2), \text{sen}(2\pi t/\lambda_2))'$ y una matriz de sistema $\boldsymbol{G}_t = \boldsymbol{I}_5$ para todo t. Bajo los supuestos mencionados y asignando una distribución inicial sobre $(\boldsymbol{\theta}_0, \tau_0)$ en la familia Normal/Gama, es posible obtener un procedimiento conjugado de actualización sobre las distribuciones finales y predictivas del modelo en cada tiempo t, semejante al modelo 1. Los detalles sobre el proceso de actualización de la información se encuentran en West y Harrison (1997).

Después de monitorear diferentes variantes del modelo usando algunas medidas de diagnóstico predictivo descritas en el apéndice C, definimos para el IMECA tres componentes armónicos con periodos $\lambda_1 = 2$, $\lambda_2 = 13$ y $\lambda_3 = 21$, y especificamos el factor de descuento para la varianza $\beta = 0.95$ y para la matriz de covarianzas del estado como $\delta = 0.945$. Asignamos una distribución inicial para $(\boldsymbol{\theta}_0, \tau_0)$ en la familia conjugada Normal/Gamma con un vector de medias nulo, una matriz de covarianzas $10\boldsymbol{I}_7$, y los valores $a_0 = b_0 = 0.01$. Después de efectuar el proceso de aprendizaje secuencial, obtenemos que

la densidad predictiva del IMECA para el tiempo $T+1$ es

$$p_4(y_{T+1}|\boldsymbol{y}_T) = St(y_{T+1}|83.4450\ ,\ 10.8924\ ,\ 120). \tag{5.10}$$

5.1.5 Modelo 5

El quinto modelo propuesto es un caso particular de la clase de modelos autorregresivos lineales donde los errores tienen una distribución t-Student con ν grados de libertad. Este modelo es más robusto al Gaussiano usual. Semejante al Modelo 3 (sección 5.1.3), podemos expresarlo de manera equivalente como una mezcla continua de escalas de Normales. En el Modelo 3 la distribución de la mezcla correspondía a una distribución estable estrictamente positiva, con un exponente característico fijo. Para este caso, la distribución de la mezcla corresponde a una distribución Gamma-Inversa$(\nu/2, \nu/2)$, con $\nu > 0$, y el modelo queda representado como

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + \sigma\omega_t^{1/2}\varepsilon_t, \tag{5.11}$$

donde $\{\varepsilon_t\}$ es una sucesión de ruido blanco Gaussiano con varianza uno, σ es el parámetro de escala desconocido, y $\{\omega_t\}$ es una sucesión de variables aleatorias i.i.d. desconocidas, con distribución Gamma-Inversa$(\nu/2, \nu/2)$. Marginalizando respecto a ω_t se tiene, en cada tiempo t, un error aleatorio con distribución t-Student $St(0, 1, \nu)$. De nuevo, al igual que en el caso del Modelo 3, consideramos que el parámetro ν forma parte de la especificación del modelo. De esta forma las cantidades desconocidas de interés, para un número T de datos, son $(\boldsymbol{\beta}, \sigma, \boldsymbol{\omega})$, con $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_T)'$. La distribución final de éstos parámetros no es manejable analíticamente en conjunto sin embargo, es posible definir un esquema de muestreo semejante al empleado en el Modelo 3 para obtener muestras de la distribución final conjunta de estos parámetros mediante el muestreador de Gibbs por bloques. Si asignamos una distribución inicial no informativa de la forma $\pi(\boldsymbol{\beta}, \sigma, \boldsymbol{\omega}) = \pi(\boldsymbol{\beta}, \sigma)\pi(\boldsymbol{\omega})$, con $\pi(\boldsymbol{\beta}, \sigma) \propto \sigma^{-1}$, condicional en $\boldsymbol{\omega}$, el modelo (5.11) corresponde al caso de un modelo lineal Gaussiano heteroscedástico, de manera que podemos obtener las distribuciones finales condicionales completa para $\boldsymbol{\beta}$ y σ, como en el Modelo 3. Por otro lado, condicional en $\boldsymbol{\beta}$ y σ, se tiene que los elementos en $\boldsymbol{\omega}$ son mutuamente

independientes, de manera que cada ω_t tiene una distribución final condicional completa Gamma-Inversa$((\nu+1)/2,(\sigma^{-2}u_t^2+\nu)/2)$, donde $u_t = y_t - \boldsymbol{x}_t'\boldsymbol{\beta}$, para $t = 1,\ldots,T$.

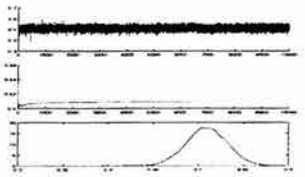
Para modelar el IMECA elegimos un modelo autorregresivo de orden $p = 13$ con errores t-Student con $\nu = 3$ grados de libertad. En la figura 5.3 presentamos los histogramas de las distribuciones finales marginales de los parámetros con 10,000 datos, obtenidas después de haber considerado un periodo inicial de calentamiento de 8,000 iteraciones. La densidad predictiva Rao-Blackwellizada, $p_5(y_{T+1}|\boldsymbol{y}_T)$, se calcula de manera semejante a la expresión (5.8) del modelo 3, simplemente sustituyendo en esa expresión los valores de $\lambda_{T+1}^{(i)}$ por $\omega_{T+1}^{(i)}/2$, con la sucesión $\{\omega_{T+1}^{(i)}\}$ formada por una muestra simulada i.i.d. de la distribución Gamma-Inversa$(\nu/2,\nu/2)$, i.e.

$$p_3(y_{T+1}|\boldsymbol{y}_T) = \frac{1}{N}\sum_{i=1}^{N} N(y_{T+1}|\boldsymbol{x}_{T+1}'\boldsymbol{\beta}^{(i)},\omega_{T+1}^{(i)}(\sigma^{(i)})^2). \tag{5.12}$$

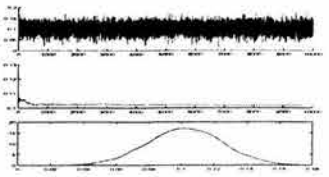
En todos los modelos donde fue requerido monitoreamos la convergencia de la cadena de Markov usando las gráficas de los promedios ergódicos de cada parámetro, que corresponden a los páneles centrales de las gráficas que mostramos. En los páneles superiores mostramos las trayectorias de las cadenas individuales de parámetro especificado. En todos los modelos graficamos los datos extraídos después del periodo inicial de calentamiento de las cadenas. En los páneles inferiores graficamos las densidades finales marginales aproximadas por *kernels*, para cada parámetro de interés. Estas se calcularon usando funciones *kernel* Optiïnales o Gaussianas, considerando los parámetros de suavizamiento usuales (vea el apéndice B.3). En los modelos donde es requerido, podemos suponer que la cadena se encuentra dentro de su periodo ergódico.

5.2 Comparación y Selección de Modelos

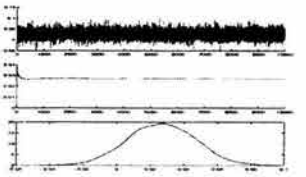
En este trabajo desarrollamos parte de la teoría en torno al problema de selección de modelos. Consideramos que la alternativa más honesta surge de plantear el problema con la perspectiva abierta (en este caso semiabierta) sobre la clase modelos contendientes. Como mencionamos, instrumentar un criterio de selección con esta perspectiva es aún limitado pues a la fecha no existe un modelo que sea completamente no paramétrico



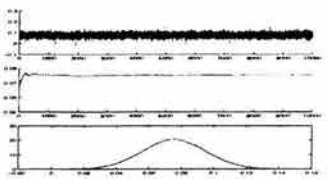
(a) β_1



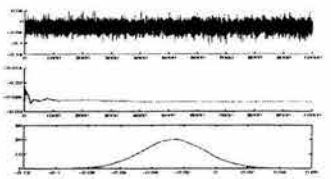
(b) β_2



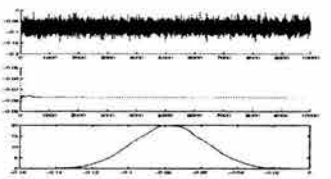
(c) β_3



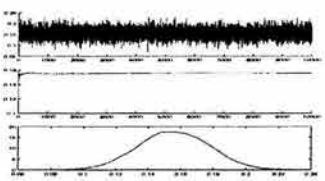
(d) β_4



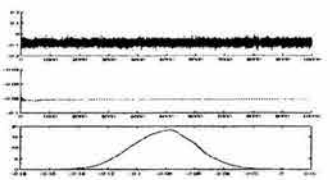
(e) β_5



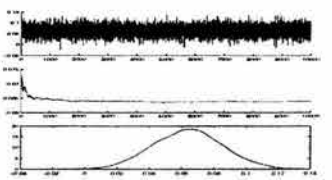
(f) β_6



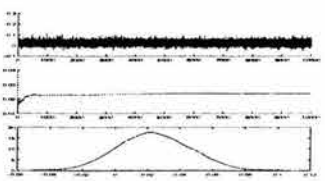
(g) β_7



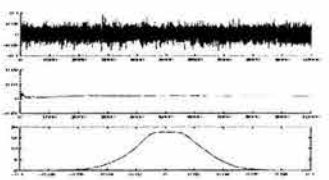
(h) β_8



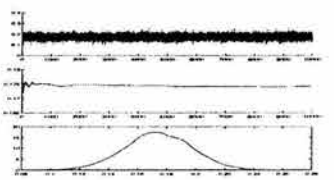
(i) β_9



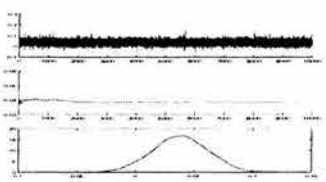
(j) β_{10}



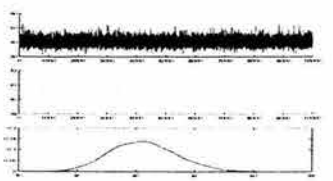
(k) β_{11}



(l) β_{12}



(m) β_{13}



(n) σ^2

Figura 5.4: Distribuciones finales de los parámetros del Modelo 5.

para modelar series de tiempo. La alternativa que proponemos en el capítulo 4 es la de aproximar el verdadero modelo con un modelo semiparamétrico flexible, en nuestro caso definido como la mezcla Bayesiana de dos modelos semiparamétricos. El objetivo de este capítulo es el de ilustrar la implementación del criterio predictivo de selección que surge de esta perspectiva, por lo que antes de continuar con la comparación de modelos presentamos una breve descripción de los lineamientos y resultados del análisis semiparamétrico de la serie del IMECA.

5.2.1 Análisis Semiparamétrico del IMECA

Los modelos paramétricos que consideramos para modelar la serie del IMECA son generalizaciones o casos particulares de la clase general de modelos autorregresivos, con excepción del modelo 4 que es una generalización del modelo armónico para series de tiempo. En este sentido, y dado que no incorporamos información adicional relevante para modelar la serie, el modelo semiparamétrico que hará el papel de juez de los modelos contendientes es un modelo autorregresivo semiparamétrico. A continuación describimos los lineamientos que empleamos para construir los modelos semiparamétricos que forman esta mezcla flexible, en cada uno de sus componentes.

Modelo M-I

El modelo que denotamos M-I es la representación semiparamétrica en forma de autorregresión de bases radiales de onduletas con un número de bases desconocido. En esta serie elegimos la onduleta radial de Marr y consideramos que el orden de autorregresión es $p = 13$. La representación del modelo es

$$y_t = \hat{f}(\boldsymbol{y}_{t-1}) + \varepsilon_t, \tag{5.13}$$

donde $\boldsymbol{y}_{t-1} = (y_{t-1}, \ldots, y_{t-p})'$, $\hat{f}(\cdot)$ es una aproximación por bases radiales de onduletas a la verdadera función de autorregresión, (4.19), y $\{\varepsilon_t\}$ es un proceso de ruido blanco Gaussiano con varianza σ^2 desconocida. Para evitar la sobre parametrización del modelo debida a una excesiva incorporación de bases radiales, el modelo considera la modelación

de esta incertidumbre. Nosotros especificamos que el número de bases radiales está entre 0 y 100, en el caso en que se tienen cero bases radiales el modelo (5.13) se reduce al modelo autorregresivo lineal Gaussiano. La determinación del número efectivo de bases radiales estará especificada por la naturaleza de los datos.

El modelo requiere de la especificación de centroides para cada base radial, pero éstos en realidad son desconocidos. En la implementación del modelo que consideramos que los centroides de las bases radiales son extraídos aleatoriamente sin reemplazo de la colección de vectores retrasados observados. Inicialmente todos los parámetros de dilatación son elegidos iguales a uno. En esta especificación encontramos situaciones donde la especificación inicial de estos parámetros retrasa el tránsito de la cadena a su fase estacionaria, pero no de manera significativa. La elección de los valores iniciales iguales a uno representa nuestra incertidumbre sobre los valores que mejor pueden ayudar a ajustar el modelo a la serie. En términos generales, la especificación de este conjunto de parámetros para aproximaciones usando bases radiales en general es un problema no resulto aún, y en algunos casos las aproximaciones que se obtienen con diferentes especificaciones generan resultados opuestos.

Para el conjunto de parámetros de regresión de la aproximación $\hat{f}$ elegimos una distribución Normal multivariada de dimensión adecuada, centrada en un vector de media nulo y una matriz de covarianzas igual a $10\boldsymbol{I}$, donde $\boldsymbol{I}$ denota la matriz identidad con la dimensión correspondiente. La elección del vector de medias nulo refleja nuestra incertidumbre sobre la localización de la distribución sobre los parámetros, a su vez esta matriz de covarianzas intenta reflejar nuestro nivel de incertidumbre respecto a la dispersión de estos parámetros. A la varianza le asignamos una distribución inicial Gamma difusa con parámetros $\delta_1 = \delta_2 = 0.01$.

Para la construcción de la cadena de Markov consideramos un número inicial de 50 bases radiales. En general el algoritmo que implementamos con esta serie no es sensible ante cambios en este valor inicial. Para mantener balanceado los movimientos en el número de bases radiales, especificamos como parámetros asignamos una probabilidad igual a los movimientos de NACIMIENTO y MUERTE de bases radiales, con una masa

de 0.3 en cada caso. En el movimiento dimensional estático de la cadena, consideramos una probabilidad simétrica de TRASLACIÓN y DILATACIÓN con 0.2 en cada caso, de esta manera hemos definido probabilidades simétricas de transición de estados de la cadena. Sobre los grados de libertad de los modelos asignamos una distribución inicial propia Gamma$(0.1, 0.1)$.

La construcción de la cadena de Markov usando el algoritmo 2 adolece de un periodo retardado para su tránsito a la fase estacionaria. Esto es debido a la especificación de un gran número de parámetros iniciales. Es por esto que consideramos que el periodo de calentamiento sea de 80,000 iteraciones para obtener una muestra final de 10,000 observaciones de la distribución final para $(I(k), \boldsymbol{\omega}_k, \tau_k)$.

Durante la implementación del algoritmo 2 para la generación de la muestra final de $(I(k), \boldsymbol{\omega}_k, \tau_k)$ no consideramos saltos en la recolección de los datos, pues en general las autocorrelaciones de cada uno de los componentes paramétricos fueron observadas cercanas a cero, incluso para retrasos de longitud corta. Además para efectos del cálculo de puntaje logarítmico esperado para los modelos contendientes invocaremos principalmente a la propiedad ergódica de las cadenas de Markov (vea la sección 2.3.3).

En general para esta serie, el algoritmo es poco sensible ante cambios en la especificación de las distribuciones iniciales de sus componentes. Este comportamiento se refleja en las distribuciones predictivas generadas en diferentes escenarios donde modificamos la dispersión de la distribución inicial sobre los parámetros de regresión del modelo y definimos diferentes hiperparámetros en la distribución inicial sobre los grados de libertad del modelo. En general las distribuciones predictivas se concentraron alrededor del nivel medio de 133 unidades (en la escala original), y fueron evidentemente unimodales con una dispersión semejante en los casos expuestos. En la gráfica 5.5 la línea sólida gruesa representa la distribución predictiva final para el tiempo $T + 1$ considerando las especificaciones que mencionamos anteriormente.

El comportamiento de la corrida con la que recolectamos la muestra definitiva del modelo es aceptable, en general el algoritmo explora eficientemente una gran cantidad de modelos durante su periodo de calentamiento, concentrándose en una vecindad del

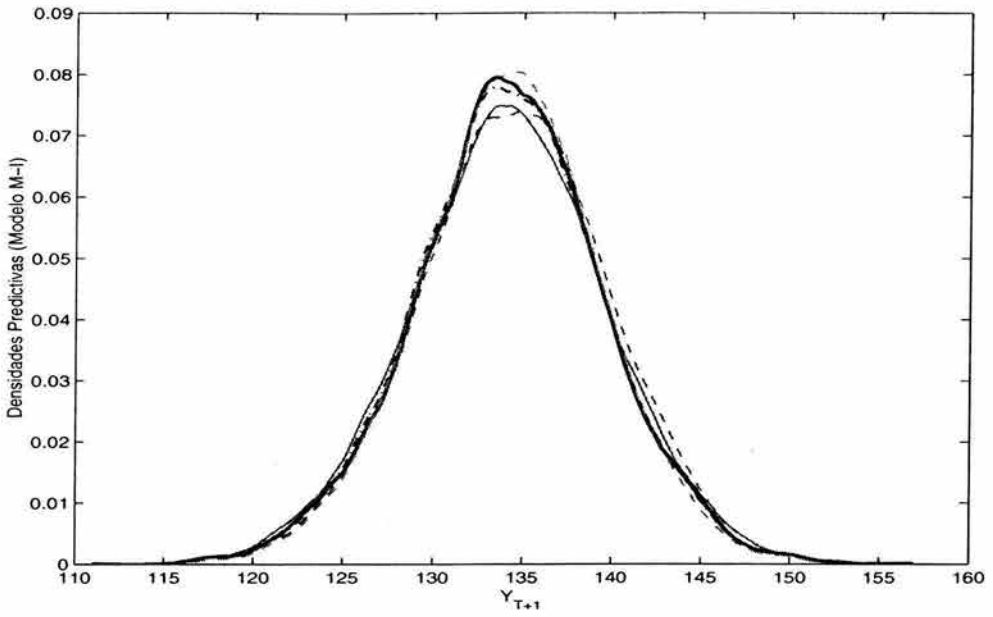


Figura 5.5: Densidades predictivas usando el modelo M-I con diferentes distribuciones iniciales.

modelo con 21 bases de onduletas radiales. Para el caso del movimiento NACIMIENTO obtuvimos un porcentaje de aceptación de 73.91%, en los movimientos de MUERTE observamos 89.64% de aceptación, para el cambio del parámetro de dilatación 70.61% y en el parámetro de traslación 16.22%. Finalmente observamos que el número de ondueltas con mayor probabilidad final fue de 21, con alta probabilidad en una vecindad a este número.

Modelo M-II

El segundo componente de la mezcla semiparamétrica es el modelo descrito en la sección 4.3, considerando como variables de regresión a vectores de retrasos de la serie de orden $p = 13$ y el componente lineal GARCH(1,1), i.e. es un modelo AR(p)-GARCH(1,1)

semiparamétrico especificado como:

$$\begin{aligned} y_t &= \boldsymbol{y}_{t-1}'\boldsymbol{\beta} + \sigma_t\varepsilon_t, \\ \sigma_t^2 &= \alpha_0 + \alpha_1\varepsilon_{t-1}^2 + \alpha_2\sigma_{t-1}^2, \\ \varepsilon_t|F &\overset{iid}{\sim} F, \end{aligned} \tag{5.14}$$

donde $\boldsymbol{y}_{t-1} = (y_{t-1}, \ldots, y_{t-p})'$ y $F \sim PT(\Delta_L, \Gamma_L)$, para un nivel L aceptable. Diferentes autores sugieren efectuar los cálculos hasta un nivel de $L = 8$ (vea Walker *et al.* (1999)), nosotros elegimos efectuar los cálculos varios niveles más abajo con $L = 11$ para obtener una mayor precisión en los cálculos. Al momento de especificar el nivel de las particiones del árbol de Pólya se debe tomar en cuenta el costo computacional que esto implica, pues el número de actualizaciones de los hiperparámetros de las distribuciones finales aumentan geométricamente. Con nuestra consideración estaremos trabajando con un árbol de particiones con $2^{11} = 2,048$ elementos en el último nivel.

Las distribuciones finales y la distribución predictiva de este modelo las obtenemos usando el algoritmo 3, que describimos en la sección 4.3. En nuestra implementación consideramos $R = 3$ cadenas simultáneas, de las que se obtendrá una muestra mediante saltos en entre estas cadenas para una mejor exploración del espacio parametral del modelo, particularmente para los parámetros del componente GARCH. Consideramos que este número de cadenas nos brinda la flexibilidad suficiente para explorar adecuadamente el espacio parametral del modelo, sobre todo del espacio asociado al componente GARCH(1,1). Recordemos que el aumento en el número de cadenas simultáneas tiene un impacto directo de letargo en el desempeño del algoritmo, en la medida que el número de operaciones necesarias es considerablemente mayor.

En la parte sistemática del modelo, asignamos una distribución inicial Normal multivarada para el vector de autorregresión $\boldsymbol{\beta}$, centrada en un vector de medias nulo de dimensión 13, y una matriz de covarianzas $10\boldsymbol{I}_{13}$. Elegimos esta distribución como una forma de manifestar nuestra incertidumbre inicial sobre el comportamiento de estos parámetros. En general el algoritmo no es sensible ante cambios en la matriz de covarianzas de la distribución inicial. Para los parámetros del componente GARCH elegimos una distribución inicial no informativa, uniforme en el simplex bidimensional. Esta distribución inicial

también refleja nuestra incertidumbre sobre estos parámetros.

Existen algunas consideraciones importantes respecto a la especificación del componente semiparamétrico. Particularmente el árbol de Pólya es extremadamente sensible ante la especificación de dos de sus componentes, uno es la distribución de probabilidad en la que se centra el proceso, la cual determina las particiones, y la segunda es la especificación de los parámetros en la clase Γ_L, las cuales determinan el tipo de distribuciones que se muestrearán. Respecto a la primera, nosotros elegimos centrar el árbol de Pólya en la distribución Normal con media cero y varianza 10. La elección de una varianza tan grande brinda una mayor flexibilidad a la distribución de los errores, que en componente GARCH tienden a explorar persistentemente regiones en las colas de la distribución, así evitamos que los errores muestreados caigan en un número pequeño de regiones en la partición del último nivel (vea Denison y Mallick (1999)).

El segundo elemento importante en la especificación del árbol de Pólya es la definición de los parámetros iniciales en las probabilidades de pertenencia a las regiones del árbol de particiones, i.e. Γ_L. Como mencionamos en la sección 4.3, éstos parámetros determinan el tiempo de distribuciones que muestrearemos. Para garantiza que muestrearemos distribuciones absolutamente continuas[2] con probabilidad uno, elegimos definir $\gamma_{\epsilon_1,...\epsilon_m} = Cm^2$, para todo $m = 1, 2, ..., L$, y $C > 0$.

A través del parámetro C se controla la velocidad de convergencia a la distribución continua que muestrea el proceso, valores grandes de C contribuyen a que esta convergencia sea más rápida. Aún cuando en la especificación el proceso muestrea de distribuciones absolutamente continuas, debemos recordar que en la práctica sólo es posible implementar aproximaciones a densidades de este tipo, por lo que básicamente realizamos todos los cálculos con errores aleatorios en un soporte de 2,048 valores distintos, en nuestro caso. En el proceso de actualización de los parámetros para la generación de las muestras de distribuciones finales, condicionales en los demás parámetros muestreados, el parámetro C juega un papel fundamental para el peso que tengan los datos en esta actualización. Para valores pequeños de C estaremos concentrando la actualización de los parámetros en

[2] Es decir distribuciones que admitan distribuciones de densidad.

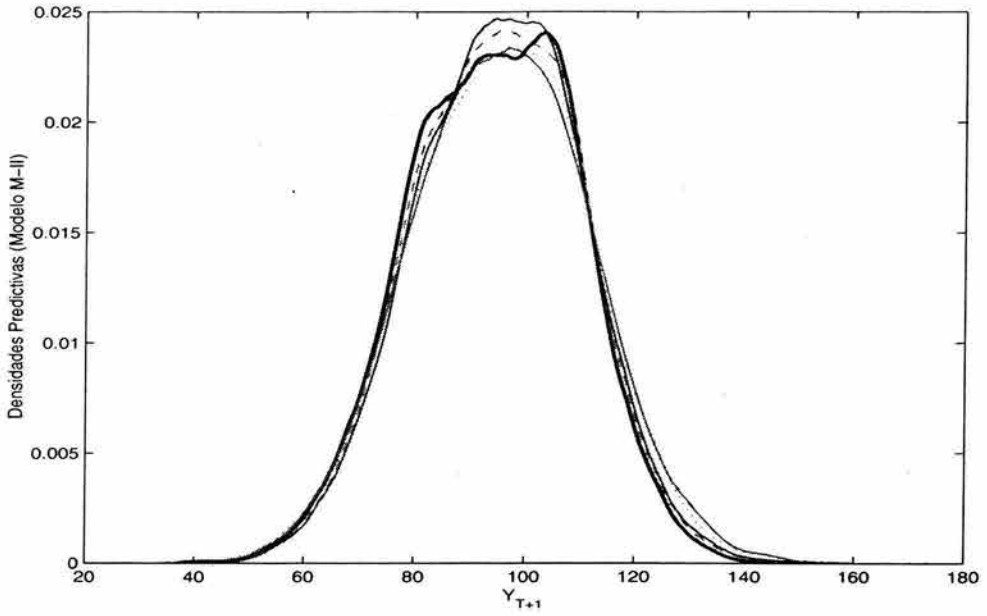


Figura 5.6: Densidades predictivas usando el modelo M-II para diferentes especificaciones de los parámetros del árbol de Pólya.

los datos observados, y las muestras generadas de las distribuciones estarán centradas en la distribución empírica de los errores, mientras que para valores grandes de C se estará muestreando consistentemente con mayor peso en la distribución de centralidad del proceso. En nuestro caso hemos elegido centrar el árbol de Pólya en una distribución Normal, que es unimodal y simétrica, pero esperamos que la distribución de los errores puede no tener este comportamiento. Exploramos los resultados que se obtienen en términos de predicción con diferentes valores del parámetro C, $C = 0.1, 0.5, 1, 10, 20$.

En la figura 5.6 graficamos las densidades predictivas del el IMECA al tiempo $T + 1$ para los cinco valores de C, considerando la misma especificación de los modelos en los demás componentes. La densidad suave que presenta unimodalidad y simetría evidentes corresponde al valor $C = 20$. Conforme el valor de este parámetro se reduce, el comportamiento del modelo es flexible, lo que se ve reflejado en la densidad predictiva para el valor $C = 0.1$, que es la densidad que presenta un ligero sesgo. En general, estas densidades están localizadas en una vecindad cercana, y su diferencia radica en el comportamiento de las colas de sus distribuciones y en el sesgo.

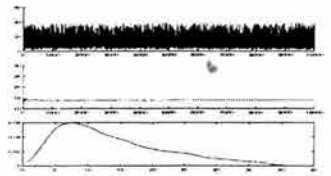
Finalmente el modelo que elegimos considera al parámetro $C = 0.1$, con el propósito de ser más flexibles en la definición y desempeño del modelo. La muestra final de las distribuciones finales de los parámetros asociados al componente lineal de este modelo se muestran en la figura 5.7, y en la figura 5.8 las correspondientes a los parámetros del componente GARCH(1,1). En todos ellos observamos un comportamiento estable en la evolución de la cadena. La flexibilidad que aporta el árbol de Pólya se refleja también en la forma de las densidades finales de los parámetros del componente lineal, que presentan sesgos en algunos casos evidentes.

Mezcla

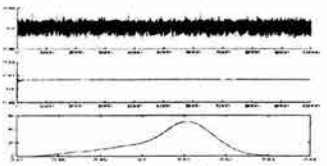
El modelo semiparamétrico flexible que toma el papel de modelo juez en nuestro criterio de selección está determinado como la mezcla de los modelos semiparamétricos M-I y M-II. Como se espera, cada uno de estos modelos captura elementos distintos de la serie. El modelo M-I está enfocado fundamentalmente a la modelación flexible del nivel de la serie, esto se ve reflejado en su comportamiento histórico del modelo ajustado. En general, posee un buen comportamiento de ajuste, y captura los cambios abruptos en nivel de manera eficiente (vea la figura 5.9(a)).

El modelo M-II contiene elementos de flexibilización en la distribución de los errores aleatorios de la serie. Este comportamiento se refleja en el ajuste a la serie, donde la dispersión de la serie es capturada por el modelo (vea la figura 5.9(b)).

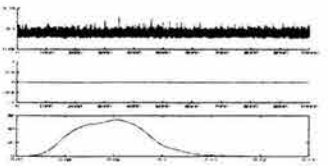
Existe una gran controversia respecto a la asignación de la distribución inicial sobre el conjunto de modelos de una mezcla. Mediante un enfoque simplista tenderíamos a pensar en asignar un mayor peso inicial a modelos complicados en espera de que tengan una mayor capacidad de ajuste a los datos observados respecto a sus modelos alternativos. Otra alternativa consiste en establecer o manifestar una especie de navaja de Ockham a priori, asignando un mayor peso a los modelos más simples (vea Jefferys y Berger (1992)). Estas controversias surgen y tienen sentido en la asignación de distribuciones iniciales sobre modelos paramértricos, pero recordemos que en este momento estamos trabajando con modelos semiparamétricos, en los que potencial y deseablemente el número de parámetros



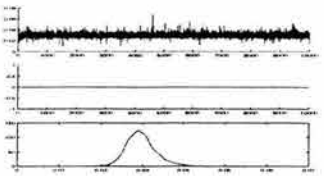
(a) β_1



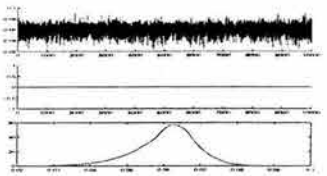
(b) β_2



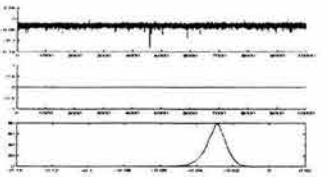
(c) β_3



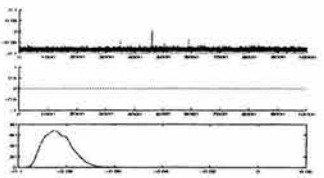
(d) β_4



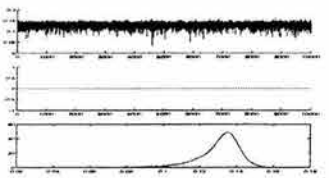
(e) β_5



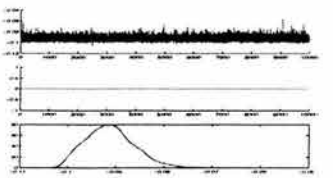
(f) β_6



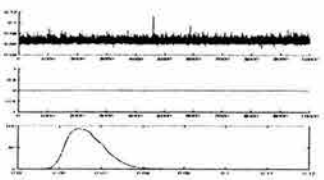
(g) β_7



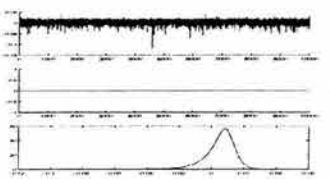
(h) β_8



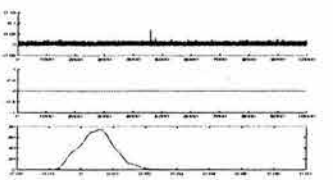
(i) β_9



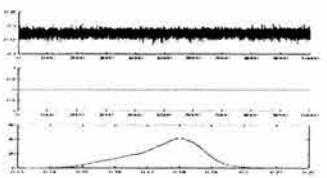
(j) β_{10}



(k) β_{11}

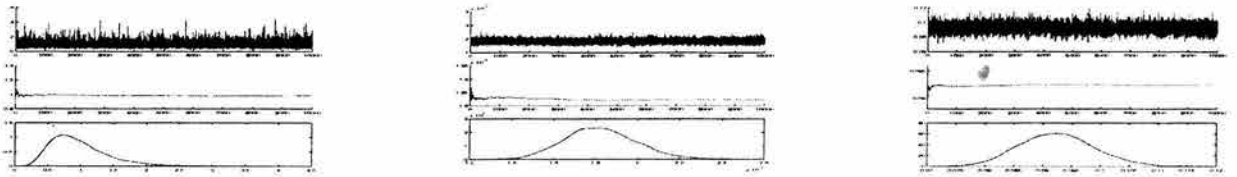


(l) β_{12}



(m) β_{13}

Figura 5.7: Muestra de la distribución final de los parámetros del componente de regresión del modelo M-II.



(a) γ_0 (b) γ_1 (c) γ_2

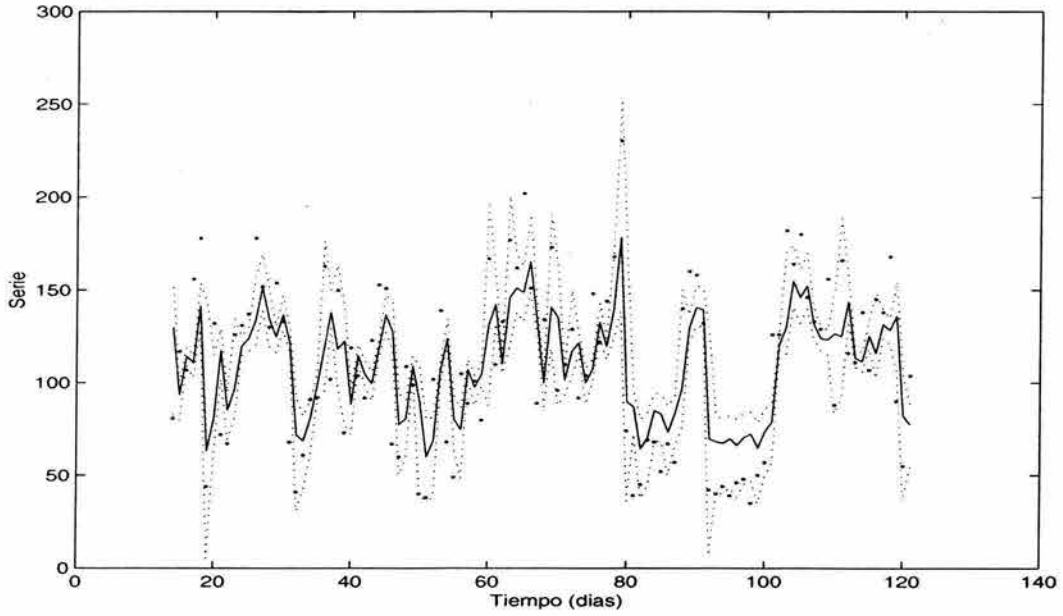
Figura 5.8: Muestra de la distribución final de los parámetros del componente GARCH(1,1) del modelo M-II.

debe ser bastante grande. En estos modelos la noción de complejidad no está aún definida, y de hecho en los modelos paramétricos tampoco. Tratar de discernir cual modelo entre M-I y M-II es más complicado no es simple.

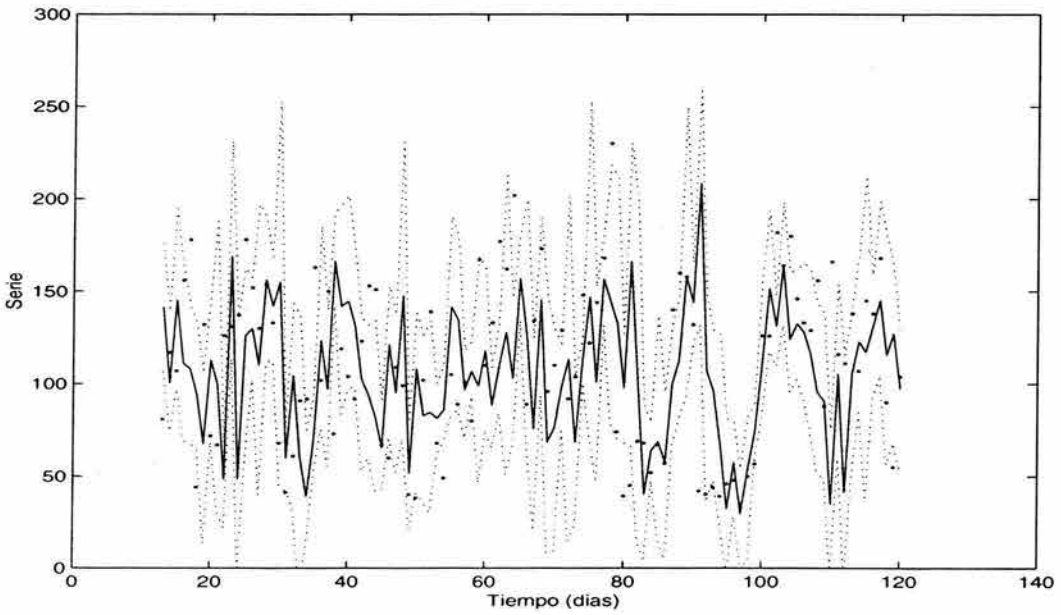
A priori, tampoco tenemos elementos sólidos para suponer que uno de los modelos pueda tener una mejor representatividad de la serie que el otro o una mejor capacidad predictiva, por lo que asignamos un peso inicial equitativo entre los dos modelos. De esta forma los pesos finales en la mezcla de las distribuciones predictivas está completamente determinada por las verosimilitudes integradas de los dos modelos, que se calculan mediante los estimadores de las verosimilitudes integradas para cada modelo (vea la sección 4.4). Estos estimadores en general resultan consistentes en tamaños de muestra grandes, mayores a 10,000 datos.

El ajuste de estos modelos tiene por objeto el de generar una aproximación de la verdadera distribución predictiva del IMECA para el tiempo $T + 1$. Las densidades predictivas generadas por los dos modelos son graficadas en la figura 5.10. Podemos observar que existe un evidente distanciamiento en la localización de estas distribuciones. Este resultado no debe causar sorpresa, pues evidentemente cada modelo contiene elementos distintivos en su concepción, algo que resultaba evidenciado en las gráficas de ajuste de los modelos (vea la figura 5.9).

Las diferencias en términos de predicción entre los dos modelos no se restringen solamente al nivel, sino que cada modelo tiene una dispersión distinta. El modelo M-I tiene una concentración alta alrededor de su nivel, mientras que el modelo M-II muestra una



(a) Modelo M-I



(b) Modelo M-II

Figura 5.9: Observaciones del IMECA (puntos), nivel medio (línea sólida) y el intervalo de credibilidad del 95% (líneas punteadas).

esperada mayor dispersión en su densidad predictiva. En cierta forma este resultado es predecible, pues es precisamente la forma distribucional lo que es modelado por M-II, mientras que la flexibilidad del modelo M-I para capturar la evolución de la serie en nivel hace que éste apueste por una menor dispersión respecto al nivel.

La mezcla de los dos modelos tiene sentido, pues no existe evidencia contundente de la dominancia de uno sobre el otro, mientras que la flexibilidad que otorga cada uno de ellos se ve fortalecida a través de la mezcla, i.e. estamos considerando un modelo que contiene elementos flexibles en nivel y en distribución. Finalmente el modelo M-II tiene la mayor verosimilitud integrada, lo que se refleja directamente en un mayor peso para la mezcla. La densidad predictiva resultante se grafica en la misma figura 5.10 con la línea sólida. Evidentemente el peso que tiene el modelo M-I es más reducido, aunque finalmente esta distribución, que toma el papel del modelo juez para los modelos paramétricos contendientes, es bimodal.

Debemos considerar que este modelo flexible lo emplearemos como modelo juez exclusivamente y es al que le asignaremos probabilidad uno, i.e. es el modelo que consideramos nuestra mejor aproximación a la distribución predictiva verdadera. Aunque dentro de las consideraciones generales del proceso de selección se tendrá que las mejores predicciones de la serie para el IMECA serán generadas por este modelo, debemos notar que el modelo puede presentar ciertas dificultades en esta dirección. Una consideración importante en términos de predicción, es el de generar pronóstios puntuales de la serie. Usando este modelo flexible, encontramos que para este efecto nos enfrentaríamos a un problema de identificabilidad para este pronóstico, pues como es sabido la media o mediana no resultan ser estimadores eficientes para distribuciones con más de una moda. Incluso en la construcción de predicción por intervalos, tendríamos el problema de tener finalmente intervalos de predicción disjuntos para ciertos niveles de probabilidad, lo que nos lleva de nuevo a un problema de identificabilidad en términos de predicción.

Este comportamiento se debe principalmente a que el modelo flexible que construimos está formado por modelos que sólo son parcialmente flexibles, en diferentes direcciones, pero que individualmente poseen todavía características restrictivas debidas a sus corres-

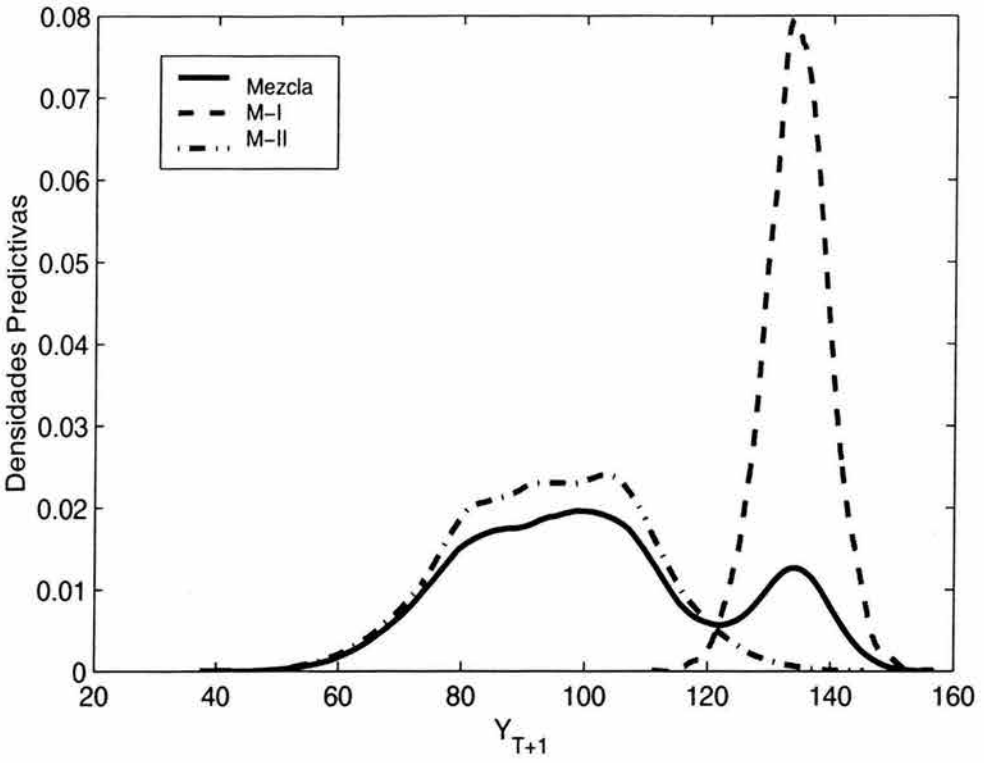


Figura 5.10: Densidades predictivas finales del IMECA para el tiempo $T+1$ del modelo semiparamétrico flexible, junto con las densidades predictivas para los modelos I y II.

pondientes componentes paramétricos. Parcialmente podríamos resolver esto si consideramos solamente el modelo que tenga un mayor peso final, sin embargo estaríamos dejando de lado elementos flexibles que el modelo alternativo puede aportar, y la asignación de probabilidad uno en ese modelo sería cuestionable. Sobre todo sesgaríamos el proceso de selección de modelos como veremos más adelante. Dado que no existe un modelo que sea completamente no paramétrico, lo que sugerimos es potenciar la capacidad predictiva de la mezcla mediante la incorporación de más modelos semiparamétricos, que contengan elementos distintivos.

Por el momento dejaremos de lado estas consideraciones y nos restringiremos a emplear el modelo flexible exclusivamente como modelo juez de los demás modelos paramétricos contendientes para efectos de selección.

5.2.2 Comparación y Selección

La clase de modelos contendientes, $\mathcal{M}$, está formada por los cinco modelos paramétricos descritos en la sección anterior. Suponemos que ninguno de estos modelos postulados representa al verdadero modelo de la serie del IMECA. La comparación de los modelos la realizamos usando el modelo flexible semiparamétrico a través de su densidad -distribución- predictiva (vea la figura 5.10). Los elementos del problema de decisión los describimos a continuación:

- **Espacio de 'estados de la naturaleza'**

Es el espacio de los posibles valores futuros de la serie del IMECA en el tiempo $T+1$, $Y_{T+1} \in \mathcal{Y}_{T+1}$, i.e. 31 de diciembre de 1998. Recordemos que el horizonte de predicción puede ampliarse para un periodo de mayor longitud.

- **Espacio de Acciones**

El espacio de acciones está formado por el conjunto de todas las densidades - distribuciones - predictivas de los modelos en la clase $\mathcal{M}$, i.e.

$$\mathcal{A} = \{p_k(\cdot|\boldsymbol{y}_T) : k = 1, ..., 5\}, \tag{5.15}$$

donde $p_k(\cdot|\boldsymbol{y}_T)$ denota la densidad (distribución) predictiva del modelo k, para $k = 1, \ldots, 5$.

En la figura 5.11 graficamos las densidades predictivas del espacio de acciones $\mathcal{A}$ en la misma escala. Para los modelos 1 y 4, las densidades son calculadas de manera analítica. Para el modelo 2 se estimó mediante la aproximación por *kernels*. Las densidades de los modelos 3 y 5 corresponden a las densidades predictivas Rao-Blackwellizadas (Gelfand y Smith, 1990). En la misma gráfica presentamos la densidad predictiva del modelo juez, i.e. el modelo semiparamétrico. Es notorio que estamos comparando densidades unimodales con una densidad bimodal, pero recordemos que esta bimodalidad tiene su origen en la mezcla de dos modelos semiparamétricos que aún tiene componentes estructurales rígidos. Por ejemplo, el componente M-II de la mezcla es una flexibilización de un

modelo autorregresivo lineal, y los modelos paramétricos 3 y 4 postulados que se encuentran localizados en una región cercana a la densidad semiparamétrica donde M-II tiene una mayor representación, son también modelos lineales.

El modelo 1 es una flexibilización del modelo autorregresivo lineal, este modelo interpreta patrones no lineales en la serie de manera eficiente. Debida a esta flexibilidad los modelos dinámicos tienden a generar mejores predicciones respecto a la que se obtienen con un modelo rígido. La distribución predictiva para este modelo se encuentra desplazada hacia la región donde el componente M-I está representado en la mezcla. Visualmente podemos observar que la densidad predictiva del modelo 1 está más próxima a la densidad semiparamétrica. por otro lado, de haber elegido solamente al modelo M-II que es el componente semiparamétrico con mayor en la mezcla, observamos que los dos modelos con la misma forma estructural en su parte sistemática estaría favorecidos o más próximos a la densidad juez.

- **Función de Utilidad**

Utilizamos la función de puntaje logarítmico como función de utilidad

$$u\left\{p_k(\cdot|\boldsymbol{y}_T), y_{T+1}\right\} = \log p_k(y_{T+1}|\boldsymbol{y}_T). \tag{5.16}$$

El modelo (óptimo) que elegiremos es el que maximice la utilidad esperada en la clase $\mathcal{A}$,

$$\overline{u}(p_k(\cdot|\boldsymbol{y}_T)) = \int \log\left(p_k(y_{T+1}|\boldsymbol{y}_T)\right) dF(y_{T+1}|\boldsymbol{y}_T), \tag{5.17}$$

donde $F(y_{T+1}|\boldsymbol{y}_T)$ denota la distribución predictiva un paso adelante del modelo semiparamétrico *flexible* formada por la mezcla de los modelos semiparamétricos M-I y M-II. Recordemos que el puntaje logarítmico esperado (5.17) tiene una estrecha relación con la medida de Kullback-Leibler de divergencia dirigida entre dos funciones de densidad, y maximizar esta cantidad es equivalente a minimizar la medida de discrepancia de la densidad predictiva paramétrica respecto a la predictiva semiparamétrica. Esta es una medida de cercanía de lo que uno puede apreciar como el la figura 5.11. A simple vista no es fácil distinguir cual de las densidades predictivas entre los modelos 1 y 3 es más

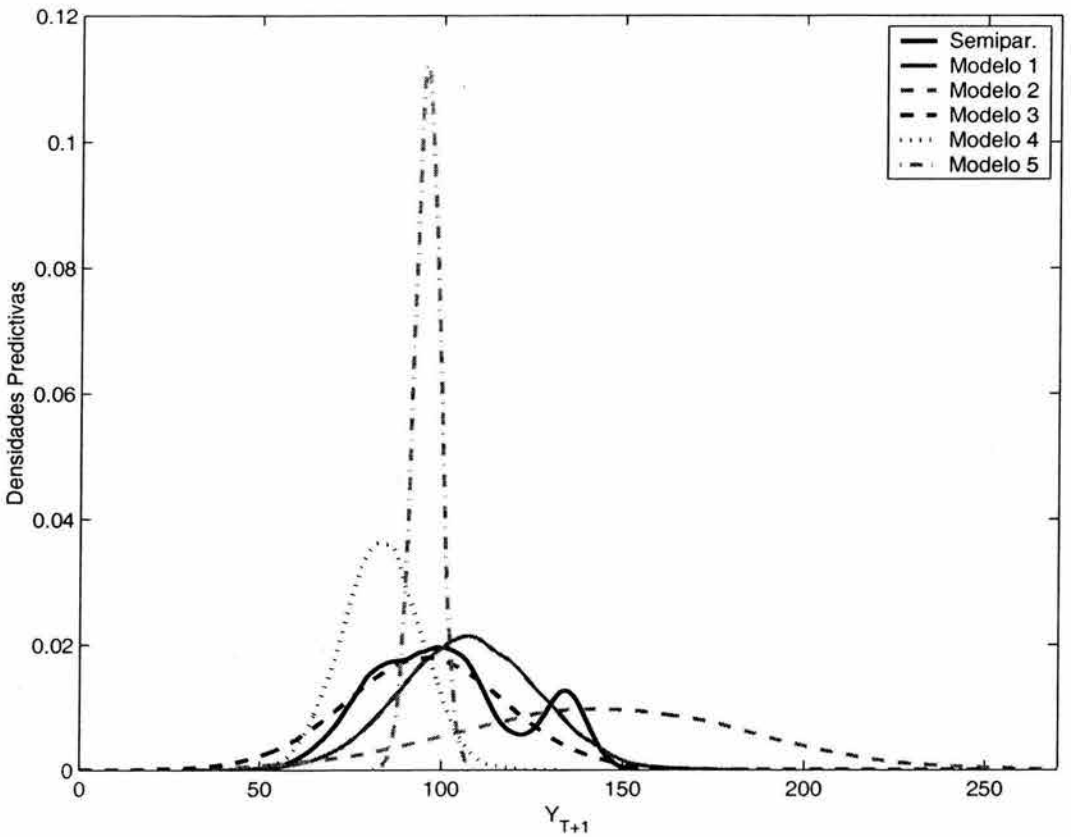


Figura 5.11: Densidades predictivas de los modelos paraméticos de la clase $\mathcal{A}$.

Modelo	Utilidad Esperada	
Modelo 1	**-3.3906**	(0.0057)
Modelo 2	-5.3347	(3.3927e-5)
Modelo 3	-4.4666	(1.6823e-4)
Modelo 4	-6.3154	(0.0015)
Modelo 5	-4.8466	(0.6962e-5)

Cuadro 5.1: Utilidades esperadas de los modelos en el espacio de acciones $\mathcal{A}$.

cercana a la densidad predictiva flexible. A través de esta medida encontramos una forma discernir este tipo de cuestionamientos que son identificados a través de una inspección visual de las gráficas de las densidades.

En el cuadro 5.1 presentamos las aproximaciones de Monte Carlo de los puntajes logarítmicos (utilidades) esperados para estos modelos. Las cantidades entre paréntesis corresponden a la varianza estimada del estimador. En este caso el modelo óptimo corresponde al modelo 1, i.e. al modelo TVAR(13). Los modelos alternativos no tan cercanos al óptimo son el modelo 3 -AR(13) α-estable- y el modelo 5 -AR(13) con errores t- ambos pertenecientes a la clase de modelos autorregresivos lineales. Este resultado no es sorprendente, pues son favorecidos por el componente M-II de la mezcla que es una extensión del modelo semiparamétrico usual y es el componente con mayor peso en la mezcla.

Como anteriormente comentamos, de haber elegido un solo modelo semiparamétrico, digamos el modelo con mayor peso final en la mezcla que corresponde al M-II, habríamos sesgado los resultados del procedimiento de selección favoreciendo a los modelos dentro de la clase de modelos semiparamétricos. La incorporación del segundo componente en el modelo juez, brinda la flexibilidad adicional de considerar modelos paramétrico flexibles en el nivel. Es por esto que el modelo TVAR tuvo una mayor aceptación, pues representa una flexibilización paramétrica de un modelo lineal convencional y modela además variaciones en la distribución de los errores.

Por otro lado, de acuerdo a la especificación de nuestro modelo semiparamétrico juez, es posible pensar que el modelo paramétrico óptimo deba poseer la característica de

bimodalidad, que generalmente se observa a través de las mezclas de los modelos semiparamétricos. Esto nos induce a pensar que la serie del IMECA pueda tener patrones conductuales que no son capturados por un modelo unimodal, i.e. puede estar formada por componentes estructurales distintos. En cierta forma esto tiene sentido, pues la serie como tal está formada como un indicador agregado de mediciones de diferentes partículas que existen en el ambiente (vea la sección 5.2), y posiblemente dos o más de ellas tengan. Esto nos sugiere que puede resultar conveniente el modelar de manera individual a cada uno de los componentes que conforman el IMECA, y posteriormente. Finalmente podemos almacenar la información y los resultados obtenidos en el problema de selección para considerar la posibilidad de replantear al análisis que estemos realizando del fenómeno de interés, enriqueciendo nuestro ciclo de aprendizaje.

En este capítulo hemos presentado una aplicación de un criterio Bayesiano sólido de selección de modelos, para el caso de series de tiempo escalares. Hemos evidenciado que es posible instrumentar un problema tan importante, que en esencia es teórica y computacionalmente demandante. El criterio predictivo $\mathcal{M}$-semiabierto que aplicamos representa una alternativa conceptualmente atractiva y que no ha sido tratada aún en la practica. Por otro lado, su planteamiento es bastante general por lo que la extensión para seleccionar modelos para series de tiempo ordinales, categóricas, de conteo, etc., puede ser extendida fácilmente salvo por la especificación del modelo flexible semiparamétrico que juegue el papel de modelo juez durante el proceso de selección.

Capítulo 6

Conclusiones

En este trabajo hemos abordado el problema de selección de modelos para series de tiempo como un problema de Bayesiano de decisión cuando por diferentes circunstancias es necesario seleccionar sólo un modelo paramétrico. Tales circunstancias surgen en contextos específicos en áreas como la economía, las ciencias sociales, las ciencias biológicas, etc., y en general donde el análisis de una serie de tiempo esté encaminado no sólo a generar predicciones, sino también en inferir sobre algunas características de la serie o sobre posibles relaciones de causalidad con otras variables observables. Realizamos una breve revisión de diferentes perspectivas que usualmente se asumen respecto a la clase de modelos por comparar y presentamos algunas soluciones que se han propuesto en la literatura usando el enfoque Bayesiano de inferencia. Por otro lado, abogamos por el uso del enfoque predictivo de selección porque en la mayoría de los casos generar predicciones generalmente es el objetivo central del análisis, inclusive cuando éste no es el caso consideramos que este enfoque debe ser empleado para seleccionar un modelo pues las distribuciones predictivas son características comparables entre los diferentes modelos paramétricos que sean postulados.

Realizamos una aplicación al análisis de series de tiempo del criterio de selección de modelos que Gutiérrez-Peña y Walker (2001) propusieron e implementaron para variables aleatorias intercambiables. Este criterio tiene dos elementos distintivos respecto a la

mayoría de las propuestas en la literatura, uno de ellos es que se considera que generar predicciones es el objetivo final del problema de decisión. Con esta consideración y los elementos del problema de decisión que describimos se obtiene implícitamente que las distribuciones predictivas de los modelos sean las características o elementos de comparación para efectos de selección. Debemos considerar que en algunas aplicaciones el problema de predicción no es el objetivo final del análisis, sin embargo considerar un criterio que permita la comparación entre los modelos postulados usando las distribuciones predictivas nos brinda la certeza de realizar las comparaciones con características que son comparables entre cualquiera de los modelos propuestos.

El segundo elemento distintivo de este criterio es que se asume una perspectiva completamente abierta respecto a la clase de modelos contendientes. En nuestra aplicación a los modelos de series de tiempo encontramos que asumir esta posición no es viable en primera instancia debido a la ausencia de un modelo completamente no paramétrico para series de tiempo, por lo que solamente es posible acercarnos a este perspectiva usando un modelo semiparamétrico flexible como aproximación a la "verdadera" distribución predictiva de la serie. Este acercamiento a la perspectiva pura abierta la denotamos como semiabierta, pues en esencia estamos aproximando la verdadera distribución del proceso con modelos que aún posees componentes estructurales restrictivos. Bajo esta perspectiva la elección del modelo semiparamétrico es subjetiva dentro del criterio, e implícitamente lo que el analista hace es asignar probabilidad uno al modelo semiparamétrico elegido cuidadosamente. No debe sorprendernos esta elección, pues el carácter subjetivo en la especificación de los elementos del problema de selección está presente también en la perspectiva cerrada, inclusive de manera más restrictiva. Bajo la perspectiva cerrada el analista asigna implícitamente probabilidad uno a la clase de modelos postulados, violando así el principio de Cromwell pues de manera indirecta se está asignando probabilidad cero a los modelos fuera de esta clase, y sin importar la naturaleza de los datos estos modelos continuarán teniendo probabilidad cero a posteriori. En los casos donde el analista o tomador de decisión tenga elementos suficientes para realizar su análisis dentro de una clase o colección de modelos específica, la perspectiva cerrada es completamente válida.

No necesariamente lo es cuando se carece de bases teóricas que sustenten esta perspectiva.

Bajo la perspectiva semiabierta que hemos asumido en este trabajo no existe un juicio apreciativo sobre la plausibilidad de los modelos postulados. La violación en que se incurre bajo la perspectiva cerrada está parcialmente presente en esta perspectiva, pero definitivamente de una manera menos restrictiva que en la cerrada pues en este caso se asigna probabilidad uno a una clase de modelos mucho más flexible que la de los modelos propuestos. El aspecto importante en este contexto es que se debe ser consistente con la elección de la probabilidad uno en el modelo flexible, y su plausibilidad no deberá ser cuestionada a posteriori, i.e. en términos prácticos el analista apuesta a que con el modelo semiparamétrico obtiene la mejor aproximación a la verdadera distribución de la serie, mientras que con los modelos postulados obtiene elementos interpretativos adicionales sobre el comportamiento de la serie o patrones conductuales de la misma.

Debemos considerar que de ser capaces de especificar un modelo completamente no paramétrico, éste debería ser el modelo considerado en el análisis como el modelo juez de los demás modelos postulados, sin embargo como mencionamos, a la fecha no conocemos de la existencia de un modelo con estas características para analizar series de tiempo. Consideramos que la perspectiva que hemos asumido es la aproximación más cercana a la perspectiva completamente abierta del problema.

Al igual que en su momento el carácter subjetivo que representa la designación de una distribución inicial generó controversias entorno al uso del paradigma Bayesiano, en este caso la elección de un modelo semiparamétrico puede generar controversias semejantes en la misma dirección. Preguntas como ¿qué modelo semiparamétrico elegir?, ¿elegir un sólo modelo semiparamétrico o más?, ¿en qué dirección se debe flexibilizar un modelo paramétrico para convertirlo en semiparamétrico?, etc., aún no tienen respuestas. Cada una de estas preguntas puede ser respondida de diversas maneras particulares, y en sí mismas abren la puerta para un trabajo futuro. Inclusive si se tiene un problema muy particular donde existan razones sólidas para elegir una clase específica de modelos paramétricos, el modelo flexible semiparamétrico puede especificarse como una extensión semiparamétrica flexible de esta familia. De manera general y como tema de trabajo

futuro queda estudiar la sensibilidad del criterio de selección ante cambios en la especificación del modelo semiparamétrico. La sensibilidad ante los cambios en esta especificación puede reducirse si alternativamente especificamos al modelo flexible mediante mezclas de modelos semiparamétricos pertenecientes a clases de modelos distintas.

En este trabajo nosotros proponemos usar el modelo semiparamétrico formado por la mezcla Bayesiana de dos modelos semiparamétricos. Con el propósito de obtener una flexibilidad general, cada uno de los modelos planteados en la mezcla modelan semiparamétricamente diferentes componentes de la serie. Uno modela semiparamétricamente la media condicional, $\mathbb{E}(y_t|\boldsymbol{y}_{t-1})$, mediante onduletas radiales preservando el supuesto de Normalidad en los errores aleatorios. Este modelo relaja significativamente el supuesto de linealidad que usualmente se supone en series de tiempo. El segundo modela semiparamétricamente la forma distribucional de los errores aleatorios usando un proceso de árboles de Pólya, relajando los supuestos usuales sobre la distribución de la serie, tales como unimodalidad, simetría, etc. Finalmente el modelo flexible queda completamente especificado como la mezcla Bayesiana de los dos modelos, en la que los pesos finales de los modelos quedan completamente determinados por la naturaleza de los datos observados. En conjunto estos dos modelos brindan una flexibilidad aceptable respecto a la modelación de la serie de tiempo, como se muestra en la aplicación del Capítulo 5, pero aún podemos ser más ambiciosos esperando desarrollar modelos que provean de toda la flexibilidad deseada para analizar un proceso en general. También como trabajo futuro se tiene la exploración de otros modelos semiparamétricos usando bases aditivas, *splines*, árboles de clasificación, etc., en cuanto a la modelación del nivel medio, o usando otros procesos para modelar la distribución de los errores usando por ejemplo mezclas de procesos Dirichlet, mezclas de árboles de Pólya, etc. Una posibilidad no explorada en este trabajo consiste en fusionar los componentes semiparamétricos en el componente sistemático y de la distribución de los errores dentro para definir una sola clase de modelos semiparamétricos.

Además de las diferencias entre las dos perspectivas que hemos mencionado, bajo el enfoque predictivo de selección, es que bajo la perspectiva abierta (o semiabierta)

el puntaje asignado a cada modelo no depende de la clase de modelos contendientes postulados. Bajo la perspectiva cerrada el puntaje asignado a cada modelo se calcula usando la mezcla Bayesiana de modelos, que depende de la composición de la clase $\mathcal{M}$, i.e. nuestro nivel de preferencia para cada modelo estará determinado también por la colección de sus modelos contendientes. Bajo la perspectiva abierta (o semiabierta), los puntajes asociados a cada modelo son independientes de la clase de modelos contendientes, de manera que el proceso de selección está basado en comparaciones de características propias para cada una de los modelos postulados.

Formalmente la elección del modelo semiparamétrico en sí misma comprende un problema de decisión. Su solución puede plantearse con los elementos que describimos en este trabajo elevando el problema una jerarquía arriba, pero depende de la existencia de un modelo completamente no paramétrico, que como mencionamos no está disponible, al menos por ahora.

Apéndice A

Medidas de Discrepancia de Funciones de Distribución

A.1 Discrepancia de Kullback-Leibler

Sea $Y \in \mathcal{Y}$ una variable aleatoria, y $p(\cdot)$ y $q(\cdot)$ dos densidades definidas en $\mathcal{Y}$. La medida de discrepancia logarítmica, o discrepancia de Kullback-Leibler, que denotamos por $\mathcal{I}_Y(p,q)$, mide qué tan distante está una densidad, por ejemplo q, respecto a la densidad p, en una escala logarítmica, y se define como

$$\mathcal{I}_Y(p,q) = \int \log \frac{p(y)}{q(y)} p(y) dy. \tag{A.1}$$

En general esta medida de discrepancia no es simétrica, i.e. $\mathcal{I}_Y(p,q) \neq \mathcal{I}_Y(q,p)$, y no cumple con la propiedad de transitividad, por lo que esta medida no puede ser considerada como distancia entre densidades propiamente, sino que se interpreta como una medida de discrepancia de una función respecto a la otra. Esta medida tiene propiedades importantes, relacionadas con el valor de la información proporcionada por nuevos datos, y en términos de suficiencia para estimación parametral. Algunas de las características más importantes de esta medida son (Bernardo y Smith, 1994, sección 3.4.):

- La medida de discrepancia $\mathcal{I}_Y(p,q)$ es más grande conforme la densidad q se aleja

de p en su soporte, y es no negativa, i.e. $\mathcal{I}_Y(p,q) \geq 0$. Además $\mathcal{I}_Y(p,q) = 0$ si y sólo si $q = p$.

- Si tenemos dos variables independientes Y_1 y Y_2, entonces $\mathcal{I}_{Y_1,Y_2}(q,p) = \mathcal{I}_{Y_1}(q,p) + \mathcal{I}_{Y_2}(q,p)$.

- Si q es un elemento de una clase de densidades Q, entonces encontrar la densidad q que minimiza $\mathcal{I}_Y(p,q)$, es equivalente encontrar la densidad q en Q que maximiza la integral $\int_{\mathcal{Y}} \{\log q(y)\}\, p(y)dy$.

A.2 Discrepancia Cuadrática

Sea $Y \in \mathcal{Y}$ una variable aleatoria y sea $q(\cdot)$ una función de densidad en una familia $\mathcal{Q}$ de densidades definidas en $\mathcal{Y}$. La función de puntaje cuadrática $u : \mathcal{Q} \times \mathcal{Y} \to \Re$ se define como (Bernardo y Smith, 1994, sección 3.4.):

$$u(q(\cdot), y) = A\left\{2q(y) - \int q^2(y)dy\right\} + B(y), \tag{A.2}$$

donde $A > 0$ es una constante conocida y $B(\cdot)$ es una función arbitraria. Esta función es propia (ver sección 2.2.2). Maximizar la utilidad esperada de $u(q(\cdot), y)$ respecto a una densidad $p(\cdot)$ definida en $\mathcal{Y}$ es equivalente a maximizar la cantidad

$$-\int \{p(y) - q(y)\}^2\, dy.$$

Apéndice B

Generales

B.1 Modelo Lineal Bayesiano

Una gran gama de modelos estadísticos pueden ser vistos como un caso particular del modelo de regresión, y algunos de éstos en particular pertenecen a la familia de los modelos de regresión lineal (por ejemplo vea los capítulos 2 y 5). En este apéndice presentamos resultados generales útiles relacionados con los modelos de regresión lineal desde la perspectiva Bayesiana, para el caso en que la matriz de regresión es de rango completo.

Sea $\boldsymbol{y} = (y_1, ..., y_n)'$ un conjunto de observaciones de la variable aleatoria Y, y sea $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)'$, con $\boldsymbol{x}_i = (x_{i,1}, ..., x_{i,p})$ para $i = 1, ..., p$, un conjunto de datos posiblemente relacionados con el nivel medio de Y a través del modelo de regresión lineal

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{B.1}$$

donde $\boldsymbol{\beta}$ es un vector columna de dimensión p que denota a los coeficientes de regresión del modelo, y $\boldsymbol{\varepsilon}$ es un vector columna de errores aleatorios con media $\mathbf{0}$ y una matriz de varianzas-covarianzas $\boldsymbol{\Sigma}$, con ambos parámetros desconocidos. Usualmente se supone que $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_n$. Para completar el modelo con el enfoque Bayesiano es necesario asignar una distribución inicial sobre los elementos de incertidumbre relevantes en (B.1), $(\boldsymbol{\beta}, \sigma^2)$. El análisis de este modelo se simplifica considerablemente si suponemos que los errores

aleatorios tienen una distribución Normal o Gaussiana. En este caso podemos asignar una distribución inicial conjugada para $(\boldsymbol{\beta}, \sigma^2)$, que resulta ser miembro de la familia Normal/Gamma-Inversa (Bernardo y Smith, 1994). Esta distribución se denota por

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma^2) &= NGaI(\boldsymbol{\beta}, \sigma_2 | \boldsymbol{b}_0, \boldsymbol{C}_0, \alpha_0, \delta_0) \\ &\propto (\sigma^2)^{-(\alpha_0+1)} \exp\left\{-\delta_0/\sigma^2\right\} \\ &\quad \times (\sigma^2)^{-p/2} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \boldsymbol{b}_0)' \boldsymbol{C}_0^{-1} (\boldsymbol{\beta} - \boldsymbol{b}_0)\right\}, \end{aligned} \tag{B.2}$$

donde $\boldsymbol{b}_0$ es un vector columna de dimensión p, $\boldsymbol{C}_0$ es una matriz de dimensión $p \times p$ simétrica positivo definida, y α_0 y δ_0 son dos valores reales positivos.

La distribución final para $(\boldsymbol{\beta}, \sigma^2)$ dado un conjunto de datos, denotados por $\mathcal{D}$, es Normal/Gamma-Inversa

$$p(\boldsymbol{\beta}, \sigma^2 | \mathcal{D}) = NGaI(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{b}_1, \boldsymbol{C}_1, \alpha_1, \delta_1), \tag{B.3}$$

con

$$\begin{aligned} \boldsymbol{C}_1 &= (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{C}_0^{-1})^{-1}, \\ \boldsymbol{b}_1 &= \boldsymbol{C}_1(\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{C}_0^{-1}\boldsymbol{b}_0), \\ \alpha_1 &= \alpha_0 + n/2, \\ \delta_1 &= \delta_0 + \frac{1}{2}\left(\boldsymbol{y}'\boldsymbol{y} + \boldsymbol{b}_0\boldsymbol{C}_0^{-1}\boldsymbol{b}_0 - \boldsymbol{b}_1\boldsymbol{C}_1^{-1}\boldsymbol{b}_1\right). \end{aligned}$$

En el capítulo 4 vimos la necesidad de calcular la verosimilitud integrada para evaluar la probabilidad de aceptación de un movimiento en el algoritmo de MCCMSR. Bajo los supuestos antes mencionados es posible calcularla mediante una expresión analítica cerrada como

$$\begin{aligned} \int\int p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}, \sigma^2) d\boldsymbol{\beta} d\sigma^2 &= \int\int N_n(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n) NGaI(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{b}_0, \boldsymbol{C}_0, \alpha_0, \delta_0) d\boldsymbol{\beta} d\sigma^2 \\ &= \frac{|\boldsymbol{C}_1|^{1/2} \delta_0^{\alpha_0} \Gamma(\alpha_1)}{|\boldsymbol{C}_0|^{1/2} (2\pi)^{n/2} \Gamma(\alpha_0)} \delta_1^{-\alpha_1}, \end{aligned}$$

con $\boldsymbol{C}_1$, $\boldsymbol{b}_1$, α_1, y δ_1 definidos como antes.

Consideremos ahora que tenemos dos modelos de regresión con diferentes matrices de regresión $\boldsymbol{X}$ y $\boldsymbol{X}^*$ de diferentes dimensiones. En este caso tenemos dos modelos

de regresión de dimensiones diferentes. Suponiendo que los errores en ambos modelos tienen una distribución Normal o Gaussiana con la misma varianza, y que asignamos una distribución inicial de la familia conjugada para $(\boldsymbol{\beta}, \sigma^2)$ y $(\boldsymbol{\beta}^*, \sigma^2)$ respectivamente, entonces encontramos que el factor de Bayes (FB) entre estos modelos es

$$FB = \frac{|\boldsymbol{C}_1^*|^{1/2} |\boldsymbol{C}_0|^{1/2}}{|\boldsymbol{C}_1|^{1/2} |\boldsymbol{C}_0^*|^{1/2}} \left(\frac{\delta_1}{\delta_1^*}\right)^{\alpha_1}, \tag{B.4}$$

donde los coeficientes $\boldsymbol{C}_1^*$, $\boldsymbol{b}_1^*$ y δ_1^* son los coeficientes de la distribución final definidos anteriormente para el modelo con la matriz de regresión $\boldsymbol{X}^*$, y $\boldsymbol{C}_1$, $\boldsymbol{b}_1$ y δ_1 correspondientes a la matriz de regresión $\boldsymbol{X}$, con $\alpha_1 = \alpha_0 + n/2$, definidos como antes.

B.2 Distribuciones α-estables

Las distribuciones α-estables se encuentran generalmente definidas mediante su función característica, como (Samorodnitsky y Taqqu, 1994):

$$\varphi(t) = \begin{cases} \exp\left\{i\mu t - \sigma|t|^\alpha \left[1 - i\delta \operatorname{sign}(t) \tan\left(\frac{\alpha\pi}{2}\right)\right]\right\}, & \text{si } \alpha \neq 1, \\ \exp\left\{i\mu t - \sigma|t| \left[1 + i\delta \operatorname{sign}(t) \frac{2}{\pi} \log(|t|)\right]\right\}, & \text{si } \alpha = 1, \end{cases} \tag{B.5}$$

donde t es un número real, $i = \sqrt{-1}$, $\operatorname{sign}(t) = 1$ si $t > 0$, $\operatorname{sign}(t) = 0$ si $t = 0$ y $\operatorname{sign}(t) = -1$ si $t < 0$. En este caso se dice que la variable aleatoria X tiene una distribución α-estable $S_{\alpha,\delta}(\mu, \sigma)$, donde $\alpha \in (0, 2]$ denota al *exponente característico* que determina las características de la distribución respecto al comportamiento de las colas, $\delta \in [-1, 1]$ denota al parámetro de sesgo de la distribución, $\sigma > 0$ denota el parámetro de escala o dispersión, y $\mu \in \Re$ denota el parámetro de localización. Entre menor sea el exponente característico se tiene que la distribución captura colas más pesadas. Por otro lado el parámetro de sesgo determina, como su nombre lo indica, el grado de sesgo de la distribución. En el caso en que $\delta = 0$ se tienen distribuciones simétricas respecto al parámetro de localización. Como caso extremo tenemos variables aleatorias estrictamente positivas cuando $\delta = 1$ y $\alpha \in (0, 1]$, si el parámetro de localización es mayor a cero. El parámetro de localización es una medida de tendencia central de la variable aleatoria y

en variables aleatorias simétricas se tiene que éste corresponde a la mediana de la distribución, cuando $0 < \alpha < 1$, mientras que cuando $1 < \alpha \leq 2$ éste corresponde a la media de la distribución.

Solamente en algunos casos particulares se conoce la forma funcional de la función de densidad asociada a una distribución α-estable de manera analítica cerrada. Para el caso de distribuciones simétricas sólo se conoce la densidad en el caso de la distribución Normal o Gaussiana ($\alpha = 2, \delta = 0$) y la distribución Cauchy ($\alpha = 1, \delta = 0$). Un comentario adicional respecto a la distribución Gaussiana como caso particular de una distribución α-estable es que el parámetro de dispersión en esta última equivale a dos veces la varianza en el caso de su parametrización usual.

Algunas propiedades de las distribuciones α-estables

Existen dos características importantes de las distribuciones α-estables que son relevantes en este trabajo. La primera se relaciona con la cerradura de estas distribuciones ante traslaciones y cambios de escala de la variable aleatoria. La segunda consiste en representar cualquier distribución α-estable simétrica como una mezcla particular de escalas de Normales, en cuyo caso la distribución de mezcla es una transformación de una distribución estable. La descripción detallada sobre las propiedades generales de las distribuciones estables se encuentra en Samorodnitsky y Taqqu (1994).

- Si $X \sim S_{\alpha,\delta}(\mu, \sigma)$ y a y b son dos números reales, entonces la variable aleatoria $Y = aX + b$ tiene una distribución α-estable $S_{\alpha,\delta'}(\mu', \sigma')$ con parámetro de sesgo $\delta' = \text{sign}(a)\delta$, parámetro de escala $\sigma' = |a|\sigma$ y parámetro de localización $\mu' = a\mu + b$ si $\alpha \neq 1$ ó $\mu' = a\mu - \frac{2}{\pi}a\,(\log|a|\sigma\delta) + b$ si $\alpha = 1$.

- Si $X \sim N(0, 2\sigma^2)$, i.e. $X \sim S_{2,0}(0, \sigma)$, y $Y \sim S_{\alpha/2,1}\,(0, 1)$, con $\alpha \in (0, 2]$, son dos variables aleatorias mutuamente independientes, entonces la variable aleatoria $Z = Y^{1/2}X \sim S_{\alpha,0}(0, \sigma)$. Utilizando esta propiedad se obtiene la representación de variables α-estables simétricas, $Y \sim S_{\alpha,0}(0, \sigma)$, como una mezcla continua de escalas

Nombre	Kernel
Rectangular	$K(u) = \frac{1}{2}\ \mathbf{1}(\lvert u\rvert \leq 1) + 0\ \mathbf{1}(\lvert u\rvert < 0)$
Triangular	$K(u) = (1/6 - \lvert u\rvert/6)\ \mathbf{1}(\lvert u\rvert \leq \sqrt{6}) + 0\ \mathbf{1}(\lvert u\rvert > \sqrt{6})$
Gaussiano	$K(u) = (2\pi)^{-1/2} \exp\{-u^2/2\}$
Optimal	$K(u) = \frac{3}{4\sqrt{5}}(1 - u^2/5)\ \mathbf{1}(\lvert u\rvert \leq \sqrt{5}) + 0\ \mathbf{1}(\lvert u\rvert > \sqrt{5})$
Doble-peso	$K(u) = \frac{15}{16\sqrt{7}}(1 - u^2/7)^2\ \mathbf{1}(\lvert u\rvert \leq \sqrt{7}) + 0\ \mathbf{1}(\lvert u\rvert > \sqrt{7})$

Cuadro B.1: Algunas funciones *kernel* comunes.

de Normales, i.e.,

$$f_{\alpha,0}(y|0,\sigma) = \int N(y|0, 2\lambda\sigma^2) f_{\alpha/2,1}(\lambda|0,1) d\lambda, \tag{B.6}$$

donde $N(y|\mu,\sigma)$ denota la función de densidad Gaussiana con media μ y varianza σ evaluada en y, y $f_{\alpha,\delta}(x|\mu,\sigma)$ denota la función de densidad de la distribución $S_{\alpha,\delta}(\mu,\sigma)$ evaluada en x.

B.3 Aproximación de Densidades por *Kernels*

Supongamos que tenemos una muestra $\{x^{(n)} : n = 1, ..., N\}$ de una variable aleatoria X con función de densidad f desconocida. Para un valor específico $x \in \Re$, podemos aproximar $f(x)$ mediante el estimador por kernel de f, denotado por $f_N(x)$, dado por la expresión

$$f_N(x) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{b_n} K\left(\frac{x - x^{(n)}}{b_n}\right), \tag{B.7}$$

donde $(b_n)_{n\geq 1}$ es una sucesión de números reales positivos que satisfacen que $\lim_{n\to\infty} b_n = 0$ y $\lim_{n\to\infty} nb_n = \infty$, y $K(\cdot)$ es una función kernel positiva y que integra a uno en $\Re$ (algunas funciones *kernel* comunes se muestran en el cuadro B.1). La aproximación (B.7) tiene entre otras propiedades que para un valor x en particular la esperanza $\mathbb{E}[f_N(x)]$ converge a $f(x)$ y que $f_N(x)$ converge en probabilidad a $f(x)$ cuando $N \to \infty$. Comúnmente se define $b_n = cn^{-1/5}$, para algún valor constante c positivo. Detalles sobre ésta y otras aproximaciones de funciones de densidad se pueden encontrar en Silverman (1986).

Apéndice C

Medidas de Diagnóstico

Una vez que ha sido seleccionado un modelo, debemos evaluar su comportamiento y capacidad predictiva con el conjunto de datos ya observados con el objetivo de determinar su validez. Supongamos, al igual que en la sección anterior, que $\{y_t\}_{t=1}^{T}$ es una realización del proceso de interés hasta el tiempo T. La idea de los métodos de diagnóstico se basan en comparar la densidad predictiva $p(y_r|\boldsymbol{y}_{r-1})$ de la variable aleatoria Y_r con el valor observado del proceso y_r al tiempo r, el cual puede ser considerado como una muestra aleatoria de su densidad predictiva, para $r = 1, ..., T$. La comparación se realiza a través de la esperanza, respecto a la densidad predictiva, de una función de discrepancia, $g(Y_r, y_r)$, del valor observado y_r respecto a la variable aleatoria predictiva Y_r (Box, 1980; Gelfand *et al.*, 1992), condicional en toda la información disponible, $\boldsymbol{y}_{r-1}$. Recordemos que en el análisis de series de tiempo esta información arriba secuencialmente, teniendo así una estructura secuencial de dependencia respecto a los valores pasados de la serie. Esta noción está representada por la descomposición de la densidad conjunta del proceso hasta el tiempo T, en el producto de las densidades predictivas en cada tiempo condicionales en el proceso observado previamente, i.e.

$$p(y_1, ..., y_T) = \prod_{r=1}^{T} p(y_r|\boldsymbol{y}_{r-1}), \tag{C.1}$$

donde $\boldsymbol{y}_r = (y_1, ..., y_r)'$, para $r = 1, ..., T$. Diferentes definiciones de la función de discrepancia g producen diferentes medidas de diagnóstico, cada una con características interpretativas distintas. En este trabajo consideramos algunas medidas de diagnóstico propuestas por Gelfand *et al.* (1992). En particular consideraremos las medidas que se obtiene de definir $g_1(y_r, Y_r) = y_r - Y_r$, que es la discrepancia del valor observado del proceso respecta a la variable predictiva. en este caso la medida de discrepancia resultante está dada por $d_{1,r} = y_r - \mu_r$, donde $\mu_r = \mathbb{E}(Y_r|\boldsymbol{y}_{r-1})$. Esta medida es ampliamente utilizada en diferentes aplicaciones estadísticas, aunque no considera otras características de la distribución predictiva más que en términos de una tendencia central. Una medida global del comportamiento del modelo puede ser $D_{1,T} = \sum_{r=1}^{T} |d_{1,r}|$.

Otra medida de interés que surge de considerar características más generales de la distribución predictiva, en términos de sus colas, surge de definir $g_{2,r}(y_r, Y_r) = \frac{1}{2\varepsilon}\mathbf{1}_{C_r(\varepsilon)}(Y_r)$, donde $C_r(\varepsilon) = \{y \in \mathcal{Y}_r : y_r - \varepsilon \leq y \leq y_r + \varepsilon\}$, con $\varepsilon > 0$, que es una vecindad de Y_r respecto al valor observado del proceso. En este caso la medida de discrepancia está dada por $d_{2,r} = \frac{1}{2\varepsilon}P(C_r(\varepsilon)|\boldsymbol{y}_{r-1})$; y una medida global puede ser $D_{2,T} = \prod_{r=1}^{T} d_{2,r}$. Un caso particular de esta medida, para variables absolutamente continuas, consiste en tomar el límite de ε cuando tiende a 0, dando como resultado la medida $d'_{2,r} = p(y_r|\boldsymbol{y}_{r-1})$, que es la densidad predictiva al tiempo r evaluada en el valor observado del proceso y_r. La medida global es en este caso $D'_{4,T} = \prod_{r=1}^{T} p(y_r|\boldsymbol{y}_{r-1})$. Esta medida ha sido utilizada por algunos autores como un criterio de comparación y selección de modelos (West y Harrison, 1997).

Apéndice D

Descripción de los Códigos

En esta sección brindamos una breve descripción de las códigos implementados en MATLAB (MathWorks, 2000) que fueron utilizadas en este trabajo. Para el lector interesado se anexa una copia de éstos en un disco floppy. Para trabajar con ellos sólo se debe de añadir a los directorios de trabajo de MATLAB la ruta del directorio donde se hayan almacenado.

Regresión vía Onduletas Radiales o Redes Neuronales

El modelo se encuentra descrito en la sección 4.2. La función principal es **bwn** (vea el algoritmo 2), en la cual se deben especificar la longitud y el periodo de calentamiento de la cadena de Markov por simular, el tipo de base, el número máximo y mínimo de funciones bases y los parámetros de las distribuciones iniciales de los parámetros. Las funciones auxiliares son: **bwnmat** que genera la matriz dada por (4.24), y **bwnmv** con la que calculamos la log-verosimilitud del modelo. También de manera auxiliar, las funciones **bwndil**, **bwntras**, y **bwnnac** son incorporadas para efectuar la dilatación, traslación y nacimiento en (4.24) según sea el caso. La función **bwn_pred** genera una muestra de Monte Carlo de la distribución predictiva de la serie un paso adelante, requiere la salida de la función **bwn**. La función **bwn_verint** aproxima la verosimilitud integrada del modelo mediante el estimador por importancia, usando al *kernel* de la distribución final como la

función de importancia.

Arboles de Pólya

Para implementar los árboles de Pólya se requieren de ciertas funciones genéricas. Las funciones que aquí describimos se aplican al caso de los árboles de Pólya parcialmente especificados como describimos en la sección 4.3. La función **polya_lv** calcula la log-verosimilitud de un conjunto de datos observados, en ésta las probabilidades en el último nivel son calculadas de acuerdo a la definición del árbol de Pólya (4.35). La función **polya_prob** calcula la probabilidad marginal de un conjunto de datos de acuerdo a la ecuación (4.36). La función **polya_fin** calcula los parámetros finales Γ' del árbol de Pólya con base en (4.37). Una muestra de la distribución F en el último nivel de particiones se obtiene con la función **polya_sim**, siguiendo la definición del árbol de Pólya. Como casos particulares tenemos las funciones principales: **polya_lineal** implementa el algoritmo 3 para el modelo de regresión lineal semiparamétrico, la función **polya_regarch** implementa el algoritmo 4 para el modelo de regresión lineal semiparamétrico con errores GARCH que describimos en la sección 4.3.4. La función **polya_reggarch_verint** calcula el estimador de Monte Carlo simple de la verosimilitud integrada del modelo M-II.

Modelo 2

La función **bar_shift** calcula muestras de la distribución final del modelo autorregresivo con cambios estructurales aleatorios en el nivel, descrito en la sección 5.1.2. Las muestras de la distribución final de los parámetros se obtiene mediante el muestreador de Gibbs por bloques. La función también genera muestras de la distribución predictiva de la serie un paso adelante. Se utilizan adicionalmente las funciones **bar_shift_phifin**, que genera muestra de la distribución condicional completa de los coeficientes de autorregresión del componente latente autorregresivo; y **bar_shift_deltafin**, que genera una muestra de las distribuciones condicionales completas del vector de indicadoras de cambios aleatorios y del vector de las magnitudes de los cambios. La función **bayeslineal_barshift_kl** calcula el estimador de Monte Carlo del logaritmo de la densidad respecto a una distribución F

arbitraria, y la varianza estimada del estimador de Monte Carlo, usando una muestra de ésta.

Modelo 3

La función **rstable** genera muestras de tamaño N de una distribución α-estable $f_{\alpha,\delta}(\mu,\sigma)$ siguiendo el algoritmo de Chambers-Mallows-Stuck. La función **bayeslineal_estable** implementa el muestreador de Gibbs para el modelo de regresión lineal con errores estables descrito en la sección 5.1.3; las muestras de la distribución final de los parámetros en la mezcla de Normales se obtiene con la función auxiliar **bayeslineal_rndestable**. La función **bayeslineal_estable_denpred** calcula la densidad Rao-Blackwellizada evaluada en un punto arbitrario, con base en la muestra generada por la función **bayeslineal_estable**.

Modelo 4

La función **bdin_nivelarmonico** implementa el proceso de filtrado del modelo dinámico armónico con cambios en nivel y parámetros de regresión que describimos en la sección 5.1.4. En esta caso el programa puede incorporar al modelo un número arbitrario de componentes armónicos. El algoritmo sólo efectúa una análisis prospectivo de la serie, aunque es fácil extenderlo para filtrar retrospectivamente. Las entradas de la función son la serie de interés y los hiperparámetros de las distribuciones iniciales de los parámetros. También se deben especificar los factores de descuento observacionales y de sistema. La función tiene incorporada por omisión 0.995 y 0.97 respectivamente, aunque pueden ser especificados por el usuario. El lector interesado en los detalles del filtrado puede consultar el libro de West y Harrison (1997).

Adicionales

El modelo **ar_jump** implementa el algoritmo 1 para el promedio Bayesiano de modelos autorregresivos lineales descrito en la sección 3.6. La función **ar_sim** genera una trayec-

toria simulada de un proceso autorregresivo lineal Gaussiano. La función **nar_ref** realiza inferencias sobre un modelo autorregresivo no lineal, empleando la función **nar_mat** para generar la matriz de regresión polinomial. Por otro lado, la función **pdfkernel** aproxima por *kernels* a una densidad $f(\cdot)$ evaluada en el punto $x \in \mathcal{X}$, especificando previamente la función *kernel* (vea el apéndice B.2). La función **barma** ajusta un modelo ARMA descrito en la sección 2.4.2. La función **ar_lin** realiza la comparación de modelos autorregresivos con base en el modelo NAR o con la mezcla de modelos autorregresivos lineales (BMAR).

Bibliografía

Aitkin, M. (1991). Posterior Bayes factors (con discusión). *Journal of the Royal Statistical Society, Serie B*, **53**, 111-142.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. En *Proceedings of the Second International Symposium on Information Theory*, (Editores: B. N. Petrov y F. Czaki). Budapest: Akademia Kiado, 267-281.

Anandalingan, G. y Chen, L. (1989). Linear combination of forecasts: a general Bayesian model. *Journal of Forecasting*, **8**, 199-214.

Bates, J. M. y Granger, C. W. (1969). The combination of forecasts. *Operational Research Quarterly*, **20**, 451-468.

Berger, J. O. y Pericchi, L. R. (1996). The intrinsec Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, **91**, 109-122.

Bernardo, J. M. (1979). Expected information as expected utility. *The Annals of Statistics*, **7**, 686-690.

Bernardo, J. M. y Smith, A. F. M. (1994). *Bayesian Theory*. Chichester: Wiley.

Bordley, R. F. (1982). The combination of forecasts: a Bayesian approach. *Journal of Operational Research Society*, **33**, 234-249.

Box, G. E. P. (1980). Sampling and Bayes´s inference in scientific modelling and robustness. *Journal of the Royal Statistical Society, Serie A*, **143**, 383-430.

Buckle, J. D. (1995). Bayesian inference for stable distributions. *Journal of the American Statistical Association*, **90**, 605-613.

Bunn, D. W. (1975). A Bayesian approach to the linear combination of forecasts. *Operational Research Quarterly*, **26**, 325-329.

Chambers, J. M., Mallows, C. L. y Stuck, B. W. (1976). A method for simulating stable random variables. *Journal of the American Statistical Association*, **71**, 340-344.

Cheng, B. y Titterington, D. M. (1994). Neural networks: a review from a statistical perspective (con discusión). *Statistical Science*, **9**, 2-54.

Clemen, R. T. (1989). Combining forecasts: a review and annotated bibliography. *International Journal of Forecasting*, **5**, 559-583.

Clyde, M. (1999). Bayesian model averaging and model search strategies (con discusión). En *Bayesian Statistics 6*, (Editores: J. M. Bernardo, J. O. Berger, A. P. Dawid y A. F. M. Smith). Oxford: Oxford University Press, 157-186.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. Filadelfia: SIAM.

Denison, D. G. T. y Mallick, B. K. (1999). Analysing financial data using Pólya trees. *Technical Report*. Imperial College of Science, Londres. (Disponible en `http://www.stat.tamu.edu/~bmallick/`)

Draper, D. (1995). Assesment and propagation of model uncertainty (con discusión). *Journal of the Royal Statistical Society, Serie B*, **57**, 45-97.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics*, **2**, 615-629.

Fitzgerald, W. J., Godsill, S. J., Kokaram, A. C. y Stark, J. A. (1999). Bayesian methods in signal and image processing. En *Bayesian Statistics 6*, (Editores: J. M. Bernardo, J. O. Berger, A. P. Dawid y A. F. M. Smith). Oxford: Oxford University Press, 239-254.

Gelfand, A. E. (1998). Approaches for semiparametric Bayesian regression. En *Asymptotics, Nonparametrics and Time Series*. (Editor: Subir Ghosh). Nueva York: Marcel Dekker, 615-638.

Gelfand, A. E. y Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society, Serie B*, **56**, 501-514.

Gelfand, A. E., Dey, D. K. y Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. En *Bayesian Statistics 4*, (Editores: J. M. Bernardo, J. O. Berger, A. P. Dawid y A. F. M. Smith). Oxford: Oxford University Press, 147-167.

Gelfand, A. E. y Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398-409.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, **24**, 1317-1339.

Geweke, J. (1995). Bayesian comparison of econometric models. *Working Paper 532.* Research Department, Federal Reserve Bank of Minneapolis. (Disponible en `http://www.biz.uiowa.edu/faculty/jgeweke/`)

Ghysels, E., Harvey, A. y Renault, E. (1996). Stochastic volatility. En *Handbook of Statistics, Vol. 14: Statistical Methods in Finance*, (Editores: G. S. Maddala y C. R. Rao). Amsterdam: North-Holland, 119-191.

Gilks, W. R. y Roberts, G. O. (1996). Strategies for improving MCMC. En *Markov Chain Monte Carlo in Practice*, (Editores: W. R. Gilks, S. Richardson y D. J. Spiegelhalter). Londres: Chapman & Hall, 89-114.

Godsill, S. J. (2001). On the relationship between MCMC model uncertainty methods. *Journal of Computational Graphics & Statistics*, **10**, 230-248.

Godsill, S. J. y Kuruoğlu, E. E. (1999). Bayesian inference for time series with heavy-tailed symmetric alpha-stable noise processes. En *Proceedings on Applications of Heavy Tailed Distributions in Economics, Engineering and Statistics.* Washington DC.

Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711-732.

Gutiérrez-Peña, E. (1997). A Bayesian predictive semiparametric approach to variable selection and model comparison in regression. En *Proceedings of the 51st Biennial Session of the International Statistical Institute, Invated Papers: Book 1*, Estambul, 17-29.

Gutiérrez-Peña, E. y Walker, S. G. (2001). A Bayesian predictive approach to model selection. *Journal of Statististical Planning and Inference*, **93**, 259-276.

Hald, A. (1998). *History of Mathematical Statistics from 1750 to 1930.* Nueva York: John Wiley.

Härdle, W., Lütkepohl, H. y Chen, R. (1997). A review of nonparametric time series analysis. *International Statistical Review*, **65**, 49-72.

Hastie, T. J. y Tibshirani, R. J. (1992). *Generalized Additive Models.* Boca Raton: Chapman & Hall.

Holmes, C. C. y Mallick, B. K. (1999). Bayesian radial basis functions of variable dimension. *Neural Computation*, **10**, 1217-1233.

Holmes, C. C. y Mallick, B. K. (2000). Bayesian wavelet networks for nonparametric regression. *IEEE Transactions on Neural Networks*, **11**, 27-35.

Jefferys, W. H. y Berger, J. O. (1992). Ockham´s razor and Bayesian analysis. *American Scientist*, **80**, 64-72.

Jeffreys, H. (1961). *Theory of Probability.* Londres: Oxford University Press.

Key, J. T., Pericchi, L. R. y Smith, A. F. M. (1999). Bayesian model choice: What and Why? (con discusión). En *Bayesian Statistics 6*, (Editores: J. M. Bernardo, J. O. Berger, A. P. Dawid y A. F. M. Smith). Oxford: Claredon Press, 343-372.

Kraft, C. H. (1964). A class of distribution function processes which have derivatives. *Journal of Applied Probability*, **1**, 385-388.

Laud, P. W. e Ibrahim, J. G. (1995). Predictive model selection. *Journal of the Royal Statistical Society, Serie B*, **57**, 247-262.

Lavine, M. (1992). Some aspects of Pólya tree distributions for statistical modelling. *The Annals of Statistics*, **20**, 1222-1235.

Lavine, M. (1994). More aspects of Pólya tree distributions for statistical modelling. *The Annals of Statistics*, **22**, 1161-1176.

Madigan, D. y Raftery, A. (1994). Model selection and accounting for model uncertainty in graphical models using the Occam´s window. *Journal of the American Statistical Association*, **89**, 1535-1546.

Madigan, D. y York, J. (1995). Bayesian graphical model for discrete data. *International Statistical Review*, **63**, 215-232.

MathWorks Inc. (2000). MATLAB*: The Language for Technical Computing, Version 6.0.* Massachusets.

Mauldin, R. D., Sudderth, W. D. y Williams, S. C. (1992). Pólya trees and random distributions. *The Annals of Statistics*, **20**, 1203-1221.

McCulloch, R. y Tsay, R. S. (1993). Bayesian inference and prediction for mean and variance shifts in autorregresive time series. *Journal of the American Statistical Association*, **88**, 968-978.

Molina Escobar, A. (2001). *Procesos de Lévy en* $\Re$. Tesis de Maestría, UNAM.

Min, C. y Zellner, A. (1993). Bayesian and non-Bayesian methods for combining models and forecast with application to forecast international growth rates. *Journal of Econometrics*, **56**, 89-118.

Müller, P. (1993). Alternatives to the Gibbs sampling scheme. *Technical Report.* ISDS, Duke University. (Disponible en `http://www.stat.duke.edu`)

Müller, P. y Pole, A. (1998). Monte Carlo posterior integration in GARCH models. *Sankhyã: The Indian Journal of Statistics*, **60**, 127-144.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison (con discusión). *Journal of the Royal Statistical Society, Serie B*, **57**, 99-138.

Powell, M. J. D. (1987). Radial basis functions for multivariable interpolation: a review. En *Algorithms for Approximation*, (Editores: J. C. Masen y M. G. Cox). Oxford: Oxford University Press, 143-167.

Rasmussen, C. E. y Ghahramani, Z. (2001). Occam´s razor. En *Advances in Neural Information Processing Systems* (Editores: T. Leen, T. Dietterich y V. Tresp). Massachusets: MIT Press.

Ripley, B. D. (1987). *Stochastic Simulation.* Nueva York: John Wiley & Sons.

Robert, C. P. y Casella, G. (1999). *Monte Carlo Statistical Methods.* Nueva York: Springer-Verlag.

Samorodnitsky, G. y Taqqu, M. S. (1994). *Stable non-Gaussian Random Processes: Stochastic Models with Infinite Variance.* Nueva York: Chapman & Hall.

San Martini, A. y Spezzaferri, F. (1984). A predictive model selection criterion. *Journal of the Royal Statistical Society, Serie B*, **42**, 296-303.

Schervish, M. (1995). *Theory of Statistics.* Nueva York: Springer-Verlag.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461-464.

Shephard, N. (1996). Statistical aspects of ARCH and stochastic volatility. En *Time Series Models in Econometrics, Finance and other fields.* (Editores: D. R. Cox, D. V. Hinkley y O. E. Barndorff-Nielsen). Londres: Chapman & Hall, 1-67.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Londres: Chapman & Hall.

Smith, A. F. M. y Gelfand, A. E. (1992). Bayesian statistics without tears: a sampling-resampling perspectives. *American Statistician*, **46**, 84-88.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics*, **22**, 1701-1762.

Tierney, L. y Kadane, J. B. (1986). Accurate approximations for the posterior moments and marginal densities. *Journal of the American Statistical Association*, **81**, 82-86.

Vidakovic, B. (1999). *Statistical Modeling by Wavelets*. Nueva York: John Wiley & Sons.

Waapetersen, R. y Sorensen, D. (2001). A tutorial on reversible jump MCMC with a view toward application in QTL-mapping. *International Statistical Review*, **69**, 49-61.

Walker, S. G., Damien, P., Laud, P. W. y Smith, A. F. M. (1999). Bayesian nonparametric inference for random distributions and related functions (con discusión). *Journal of the Royal Statistical Society, Serie B*, **61**, 485-527.

Walker, S. G., Gutiérrez-Peña, E. y Muliere, P. (2001). A decision theoretic approach to model averaging. *The Statistician*, **50**, 31-39.

Weron, R. (1996). On the Chambers-Mallows-Stuck method for simulating skewed stable random variables (con corrección). *Statistics and Probability Letters*, **28**, 165-171.

West, M. y Harrison, J. (1997). *Bayesian Forecasting and Dynamic Linear Models*. Segunda Edición, Nueva York: Springer-Verlag.

Zhang, Q. (1997). Using wavelet network in nonparametric estimation. *IEEE Transactions on Neural Networks*, **8**, 227-236.

Zhang, Q. y Benveniste, A. (1992). Wavelet networks. *IEEE Transactions on Neural Networks*, **3**, 889-898.