



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

---

---

FACULTAD DE CIENCIAS

Modelos Dinámicos Lineales: Un enfoque  
Bayesiano

T E S I S

QUE PARA OBTENER EL TÍTULO DE:  
ACTUARIO

PRESENTA:  
MARÍA FERNANDA GIL LEYVA VILLA

DIRECTOR DE TESIS:  
DRA. RUTH SELENE FUENTES GARCÍA



2014



Universidad Nacional  
Autónoma de México



**UNAM – Dirección General de Bibliotecas**  
**Tesis Digitales**  
**Restricciones de uso**

**DERECHOS RESERVADOS ©**  
**PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL**

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

1. Datos del alumno

Gil Leyva  
Villa  
María Fernanda  
(55)56305075  
Universidad Nacional Autónoma de México  
Facultad de Ciencias  
Actuaría  
30628168-6

2. Datos del tutor

Dra.  
Ruth Selene  
Fuentes  
García

3. Datos del sinodal 1

Dr.  
Fernando  
Baltazar  
Larios

4. Datos del sinodal 2

Mat.  
Margarita Elvira  
Chávez  
Cano

5. Datos del sinodal 3

Act.  
Jaime  
Vázquez  
Alamilla

6. Datos del sinodal 4

Dr.  
Ricardo  
Ramírez  
Aldana

7. Datos del trabajo escrito.

Modelos Dinámicos Lineales: Un enfoque Bayesiano  
162 p 2014

# Índice

<b>Introducción</b>	<b>6</b>
<b>1. Conceptos básicos de Estadística Bayesiana</b>	<b>8</b>
1.1. Conceptos básicos de Teoría de la Decisión . . . . .	8
1.1.1. Pérdida Esperada, Reglas de decisión y Riesgo . . . . .	9
1.1.2. Estimación y tres funciones de pérdida importantes . . . . .	11
1.1.3. Un caso que nos será muy útil . . . . .	14
1.2. Teorema de Bayes . . . . .	16
1.3. Conceptos básicos . . . . .	16
1.4. Estructuras de dependencia . . . . .	18
1.4.1. Independencia Condicional . . . . .	18
1.4.2. Un primer ejemplo . . . . .	19
1.4.3. Intercambiabilidad . . . . .	23
1.4.4. Observaciones heterogéneas . . . . .	24
<b>2. Modelos Dinámicos Lineales: Aspectos Teóricos.</b>	<b>24</b>
2.1. Un primer ejemplo de Modelos Dinámicos Lineales . . . . .	24
2.2. Un ejemplo más . . . . .	28
2.3. Modelos de Espacios de Estados . . . . .	34
2.3.1. Una representación gráfica de los MEE . . . . .	37
2.3.2. Estimación de estados y predicción . . . . .	37
2.4. Modelos Dinámicos Lineales . . . . .	43
2.4.1. Filtro de Kalman para un MDL univariado y un ejemplo . . . . .	44
2.4.2. Filtro de Kalman . . . . .	55
2.4.3. Suavizador de Kalman . . . . .	60
2.4.4. Algoritmo recursivo de predicción para un MDL . . . . .	62
2.4.5. Un último ejemplo: El Modelo de Crecimiento Lineal . . . . .	63
2.4.6. El proceso de inovaciones y la verificación del modelo . . . . .	70
<b>3. Especificación del modelo</b>	<b>75</b>
3.1. Modelos dinámicos lineales con observaciones univariadas . . . . .	75
3.1.1. Modelos de Tendencia . . . . .	76
3.1.2. Modelos de comportamiento cíclico . . . . .	87
3.1.3. La serie <i>nottem</i> . . . . .	101
3.1.4. Modelos Auto-regresivos y de Promedios Móviles (ARMA) como Modelos Dinámicos Lineales . . . . .	109
3.1.5. Modelos de Regresión . . . . .	114
<b>4. Aplicación de los Modelos Dinámicos Lineales a un caso práctico.</b>	<b>117</b>
4.1. Especificación del modelo . . . . .	118
4.2. Evaluación del modelo . . . . .	120
4.3. Estimación de estados . . . . .	122

4.4. Predicción de observaciones futuras . . . . .	125
4.5. Propuesta de un Modelo Dinámico . . . . .	127
4.6. Comparación Modelo Estático vs. Modelo Dinámico . . . . .	128
<b>Conclusiones</b>	<b>134</b>
<b>A. Anexos</b>	<b>136</b>
A.1. Gráficas . . . . .	136
A.2. Matrices . . . . .	139
A.2.1. Suma de matrices y multiplicación por un escalar . . . . .	140
A.2.2. Multiplicación de matrices . . . . .	140
A.2.3. Matrices transpuestas y matrices simétricas . . . . .	142
A.2.4. Matrices inversas . . . . .	142
A.2.5. Determinantes . . . . .	143
A.2.6. Matrices ortogonales y matrices definidas positivas . . . . .	143
A.3. Propiedades importantes de la distribución Normal . . . . .	146
A.3.1. Normal multivariada . . . . .	149

## Agradecimientos

Es la primera vez que escribo los agradecimientos de una tesis, no podía ser diferente pues es la primera vez que escribo una tesis. Así que aquí voy.

Las personas a las que les quiero dedicar este trabajo son personas muy especiales para mí, por lo que resulta todo un reto escribir estos párrafos de manera que estos párrafos merezcan ser agradecimientos para dichas personas.

Estas personas de quienes estoy hablando son aquellas, que (en menor o en mayor medida) han influido en mi forma de ser y de pensar, me han hecho una mejor persona y me han ayudado a alcanzar las metas que he alcanzado, aquellas personas sin las cuales yo no sería lo que soy hoy.

Claramente, y como era de esperarse, primero quiero agradecer a mi familia. En primer lugar quiero agradecer a mis papás por haberme creado y por haberme dado todas las oportunidades que he tenido, por tenerme paciencia y guíarme con tanto cariño, por haberme enseñado con el ejemplo y la palabra el significado de palabras tan importantes y abstractas como lo son el amor, la felicidad y la vida, por apoyarme en malos momentos y disfrutar conmigo los buenos. Quiero agradecerles simplemente por ser quienes son y por haberme hecho lo que soy (Porque, por muy arrogante que suene, la verdad es que estoy encantada con la persona que soy).

En segundo lugar quiero agradecerle al resto de mi familia. A mis dos abuelitas (Tita Berthita y Tita Rosita), que son una parte muy grande de lo que soy, por sus enseñanzas, por su dulzura, su amor, sus consejos y su paciencia. A mis tíos (Moni, Berthita, Coco, Mónica, Belinda, Rosita, Toña, Angeles y Lety), que son las personas más divertidas que conozco y mis segundas mamás y/o hermanas, por su amor y las risas que hemos compartido, por sus consejos y su enorme paciencia, por estar ahí cuando las necesito y que me hacen sentir muy querida y feliz, por ser las primeras amigas que tuve en mi vida (junto con mis papás). A mis primos, por hacerme sentir especial, y también por hacerme repelar tanto cuando se portan mal, que es tan divertido discutir con ellos.

Quiero agradecerle también a todos mis abuelitos por crear a mis papás (que aunque no los concí a todos, crearon a mis papás y por lo tanto a mí), a mis 8 bisabuelos por haber creado a mis abuelos (que crearon a mis papás y por lo tanto a mí), a mis 16 tatarabuelos por haber creado a mis bisabuelos (que crearon a mis abuelos, por lo tanto a mis papás y por lo tanto a mí)... y así hasta llegar a agradecer a la primer célula desde la que desciende mi linaje (bajo hipótesis de que es cierta la teoría evolutiva de Newton).

A mis amigas de clavados (Lorena, Pamela, Ximena, Xatziri, Karina, Abril y Celia) por ser mis hermanas, por enseñarme tantas cosas, porque sin ellas no sería quien soy, por los momentos vividos, por haber estado a mi lado cuando conocí el significado de la palabra derrota y victoria, por compartir nuestra hermosa locura. Quiero agradecerles a

todas por ser tan terriblemente geniales y divertidas, porque fueron las primeras personas (externas a mi familia) a las que admiré y amé.

Quiero agradecerle a mis dos entrenadores de clavados (Regina y Jorge) por haberme hecho fuerte y valiente, por haberme retado, por haberme hecho crecer, por enseñarme lo que es la determinación, por todo su cariño. Quiero también agradecerles porque gracias a ellos me pegué menos de lo que lo hubiera hecho si hubiera intentado aprender clavados yo sola.

Quiero agradecerles a mis mejores amigos de la facultad (Karen, David, Carlos y Fermín) por hacerme la vida más leve y haber soportado mi intesidad, porque me han enseñado a vivir la vida con ligereza, porque me apoyan en mis momentos malos (a pesar de los dolorosos golpes, diría David) y comparten conmigo en los buenos. Quiero agradecerles simplemente por compartir conmigo sus vidas y por las pláticas ridículmente interesantes que tenemos.

Quiero agradecer también a mis maestros en la facultad, por sus maravillosas clases y por enseñarme el hermoso lenguaje de las matemáticas. En especial a mi directora de tesis Ruth por su paciencia, y su dedicación, porque sin ella yo no estaría hoy aquí escribiendo esto en este documento.

Quiero agradecer al resto de mi familia, amigos y maestros por las intersecciones de nuestras vidas, que pequeñas o largas han influenciado en lo que hoy soy.

Quiero agradecerle a Andrea, por enseñarme lo que se siente amar, por reír y llorar conmigo, por haberme hecho más fuerte y mejor persona y por haber sido mi compañera a lo largo de mi carrera.

Quiero agradecer a mis enemigos, porque me hacen más fuerte y valiente, porque me retan y porque aprendo tanto de ellos.

Por último, quiero agradecer a quien sea que este leyendo esto, por haberle dado a mis palabras unos pocos minutos de tus valiosos minutos. Espero que tengas una vida increíblemente feliz seas quien seas.

Espero no haber defraudado a nadie con estas palabras. En realidad hay muchas más personas a quien desearía agradecer personalmente. Sin embargo, si lo hiciera, probablemente el título del trabajo debería cambiar a "Personas que Marifer agradece y un poco de Modelos Dinámicos Lineales".

Así que, sin más por el momento, gracias.

# Introducción

A lo largo de este trabajo discutiremos los llamados Modelos Dinámicos Lineales con un enfoque Bayesiano. Estos modelos resultan ser una herramienta alternativa muy útil para el análisis de series de tiempo debido a su enorme flexibilidad, ya que, entre otras características, estos modelos no son necesariamente constantes, es decir, sus parámetros e incluso su estructura pueden depender del tiempo.

En la primera parte de este documento se presentan conceptos básicos de estadística Bayesiana, esto con el objetivo de introducir al lector los conceptos más relevantes de dicha área del conocimiento así como presentar la notación que estaremos utilizando a lo largo de este documento.

Se presentan en seguida algunos aspectos teóricos de los Modelos Dinámicos Lineales. En esta parte del documento también analizaremos, dado un grupo de observaciones (una serie de tiempo) y dado un Modelo Dinámico Lineal especificado, los problemas de estimación de cantidades de interés y predicción de observaciones futuras. En este mismo capítulo se discutirá el conocido proceso de inovaciones que nos va a ayudar a realizar la evaluación del modelo, es decir, determinar si el Modelo Dinámico Lineal especificado se ajusta de manera adecuado a los datos.

El tercer capítulo está dedicado a la especificación del modelo: dada una serie de observaciones univariadas proponer un modelo viable. En este capítulo se explica la manera de describir tendencias, ciclos, componentes autoregresivos, de promedios móviles o de regresión que una serie de tiempo univariada pueda tener, a través de un Modelo Dinámico Lineal.

En la tesis se presentan ejemplos que acompañan a la teoría para resaltar algunas ideas principales así como para la mejor comprensión de los conceptos. En la mayor parte de estos ejemplos las observaciones fueron simuladas de algún proceso en específico. Sin embargo, también se discuten un par de casos de la vida real, donde se toman series de tiempo (del software estadístico R) y se les ajusta un Modelo Dinámico Lineal adecuado. El último capítulo se dedica al análisis de la serie de tiempo Nile, se hace una propuesta del modelo, el análisis de si el modelo es adecuado, se hace la estimación y construcción de intervalos de probabilidad para cantidades de interés, y predicción de observaciones futuras.

Este trabajo se ha construido siguiendo el libro *Dynamic Linear Models with R* de Giovanni Petris, Sonia Petrone y Patrizia Campagnoli ([1]). De igual manera se utiliza y explica, a lo largo del documento el paquete `d1m` de R desarrollado por los mismos autores que nos ayuda a realizar el análisis de series de tiempo a través de un Modelo Dinámico Lineal.

Al finalizar el desarrollo principal del documento se encuentra una pequeña sección de

anexos. Incluida por dos razones principales. La primera para lograr un mejor y más firme sustento de la teoría, y la segunda para que el lector pueda alcanzar un mejor entendimiento del tema.

En general y de manera muy resumida esta es la temática del documento y sus alcances. Espero que el lector encuentre un tanto útil e interesante lo que discutiremos a continuación.

# 1. Conceptos básicos de Estadística Bayesiana

En este primer capítulo se discuten algunos aspectos básicos y relevantes de la estadística Bayesiana para garantizar un mejor entendimiento.

## 1.1. Conceptos básicos de Teoría de la Decisión

En estadística Bayesiana la Teoría de la Decisión juega un papel muy importante, pues en muchas ocasiones nos veremos en la situación de tener que decidir, de entre un conjunto de opciones posibles, cual es el valor más factible de cierta cantidad de interés. A continuación se explica a que nos referimos con esto.

Consideremos una cantidad desconocida  $\theta$  que afecta el proceso de decisión, conocida como **estado de la naturaleza**, es importante considerar todos los posibles valores que puede tomar  $\theta$ , al conjunto de estos valores se le conoce como **espacio de estados de la naturaleza** y lo denotaremos por  $\Theta$ . Típicamente, cuando se realizan experimentos, para obtener información acerca de  $\theta$ , los experimentos son diseñados de tal manera que las observaciones estén distribuidas de acuerdo a alguna distribución de probabilidad con parámetro  $\theta$ . En dichas situaciones llamaremos a  $\theta$  el parámetro y a  $\Theta$  el espacio paramétrico.

Las decisiones también son conocidas como acciones, en este caso las denotaremos por  $a$  y  $\mathcal{A}$  se referirá al conjunto de todas las posibles decisiones. En un escenario como el que se planteó en el párrafo anterior, las acciones resultan ser una función de las observaciones. Es decir, la acción que tomaremos dependerá del resultado obtenido del experimento.

Un elemento clave en la teoría de la decisión es la función de pérdida. Si una acción particular  $a_1$  es tomada y  $\theta_1$  resulta ser el verdadero estado de la naturaleza, entonces tendremos una pérdida  $L(\theta_1, a_1)$ . Entonces, supondremos que la función de pérdida  $L : \Theta \times \mathcal{A} \rightarrow \mathbb{R}$  está definida para todo  $(\theta, a) \in \Theta \times \mathcal{A}$ .

Cuando es realizada una investigación estadística para obtener información acerca de  $\theta$ . El resultado (una variable aleatoria o una muestra aleatoria) lo denotaremos por  $Y$ , comúnmente  $Y = \{Y_1, Y_2, \dots, Y_n\}$  (donde  $n$  es el tamaño de la muestra) será un vector donde cada  $Y_i$  es una variable aleatoria independiente del resto y todas provenientes de alguna distribución común. Una realización particular del experimento se denotará por  $y = \{y_1, y_2, \dots, y_n\}$ . También es importante considerar el espacio de todas las posibles realizaciones de  $Y$ , en este caso lo denotaremos por  $\mathcal{Y}$  y evidentemente será un subconjunto de  $\mathbb{R}^n$  donde  $n$  es la dimensión de  $Y$  o bien el tamaño de la muestra.

### 1.1.1. Pérdida Esperada, Reglas de decisión y Riesgo

En ocasiones se cuenta con alguna información inicial acerca del estado de la naturaleza  $\theta$  que muchas veces puede ser expresada en términos de alguna distribución de probabilidad inicial (la cual por supuesto refleja la incertidumbre que se tiene acerca de  $\theta$ ). Debido a que el valor real de  $\theta$  es desconocido, nunca vamos a poder conocer con certeza a  $L(\theta^*, a)$  donde  $\theta^*$  es el verdadero valor del estado de la naturaleza. Por esto resulta natural y razonable considerar la **pérdida esperada** considerando la información inicial acerca de  $\theta$ .

**Definición 1.1.1.1.** Si  $\pi^*(\theta)$  es la función de densidad de probabilidad (f.d.p.) de  $\theta$  que se considera al momento de tomar la decisión (o acción) la **Pérdida Esperada Bayesiana** de una acción  $a$  es:

$$\rho(\pi^*, a) = \mathbb{E}^{\pi^*}[L(\theta, a)] = \int_{\Theta} L(\theta, a) dF^{\pi^*}(\theta)$$

donde  $F^{\pi^*}(\theta)$  es la función de distribución correspondiente a la f.d.p.  $\pi^*(\theta)$ .

**Nota 1:** Para quienes no están familiarizados con la notación, en el caso continuo

$$\int_{\Theta} L(\theta, a) dF^{\pi^*}(\theta) = \int_{\Theta} L(\theta, a) \pi^*(\theta) d\theta$$

mientras que en el caso discreto

$$\int_{\Theta} L(\theta, a) dF^{\pi^*}(\theta) = \sum_{\theta \in \Theta} L(\theta, a) \pi^*(\theta).$$

**Nota 2:** A lo largo de esta sección estaremos usando esta notación por simplicidad. Para entenderla con más claridad se requieren conceptos básicos de Teoría de la Medida

**Nota 3:** En la definición anterior, usamos  $\pi^*$  en lugar de  $\pi$ , ya que  $\pi$  generalmente se refiere a la f.d.p. inicial de  $\theta$ , mientras que  $\pi^*$  se refiere a la f.d.p. de  $\theta$  al momento de tomar la decisión (puede ser la f.d.p. inicial  $\pi(\theta)$  o bien la f.d.p. posterior  $\pi(\theta|y)$ , que no es más que el conocimiento acerca de  $\theta$  actualizado, incorporando la información que nos brindan las observaciones acerca de este).

**Principio Condicional de Bayes:** Una vez calculada la pérdida esperada Bayesiana, claramente preferiremos una acción  $a_1$  sobre una acción  $a_2$  si  $\rho(\pi^*, a_1) < \rho(\pi^*, a_2)$ . La acción que minimice la pérdida esperada Bayesiana será la que elegiremos, recibe el nombre de **acción de Bayes**.

La escuela clásica o frecuentista de la teoría de la decisión, adopta una pérdida esperada un poco distinta, que se describe a continuación.

**Definición 1.1.1.2.** Una **regla de decisión** (no aleatorizada) es una función  $\delta : \mathcal{Y} \rightarrow \mathcal{A}$  (la cual se supone que es una función medible<sup>1</sup>). Si  $Y = y$  es el valor observado entonces  $\delta(y)$  será la acción que tomaremos. Dos reglas de decisión  $\delta_1$  y  $\delta_2$  son consideradas equivalentes si  $\mathbb{P}_\theta[\delta_1(Y) = \delta_2(Y)] = 1$  para todo  $\theta \in \Theta$ .

**Nota:**  $\mathbb{P}_\theta[A]$  se refiere a la probabilidad de que suceda el evento  $A$  dado el estado de la naturaleza  $\theta$ .

**Definición 1.1.1.3.** La función de riesgo de una regla de decisión  $\delta(Y)$  está definida por

$$R(\theta, \delta) = \mathbb{E}_\theta^Y[L(\theta, \delta(Y))] = \int_{\mathcal{Y}} L(\theta, \delta(Y)) dF_Y(y|\theta)$$

donde  $F_Y(y|\theta)$  es la función de distribución correspondiente a la f.d.p. de  $Y$  dado  $\theta$ , ( $f(y|\theta)$ ).

Notemos que la Pérdida Esperada Bayesiana es un número, mientras que el riesgo es una función sobre  $\Theta$ , y como  $\theta$  es desconocido tenemos un conflicto en decidir que regla de decisión es mejor. El siguiente ordenamiento parcial resulta un primer paso para resolver el conflicto.

**Definición 1.1.1.4.** Decimos que una regla de decisión  $\delta_1$  es  $R$ -mejor que otra regla  $\delta_2$  si  $R(\theta, \delta_1) \leq R(\theta, \delta_2)$  para todo  $\theta \in \Theta$ , con la desigualdad estricta para algún  $\theta \in \Theta$ . Una regla de decisión  $\delta_1$  se dice  $R$ -equivalente a otra regla  $\delta_2$  si  $R(\theta, \delta_1) = R(\theta, \delta_2)$  para todo  $\theta \in \Theta$ .

**Definición 1.1.1.5.** Una regla de decisión  $\delta$  es admisible si ninguna regla de decisión es  $R$ -mejor. Una regla de decisión es inadmisibles si existe alguna regla de decisión que sea  $R$ -mejor.

Nótese que la pérdida esperada Bayesiana promedia sobre  $\Theta$ , mientras que el riesgo promedia sobre  $\mathcal{Y}$ , el riesgo de Bayes que definiremos a continuación promedia sobre ambos.

**Definición 1.1.1.6.** El **riesgo de Bayes** de una regla de decisión  $\delta$  con respecto a una distribución  $\pi^*$  se define como

$$r(\pi^*, \delta) = \mathbb{E}^{\pi^*}[R(\theta, \delta)] = \int_{\Theta} \left\{ \int_{\mathcal{Y}} L(\theta, \delta(Y)) dF_Y(y|\theta) \right\} dF^{\pi^*}(\theta).$$

**El Principio de Riesgo de Bayes:** Una regla de decisión  $\delta_1$  se prefiere a una regla de decisión  $\delta_2$  si  $r(\pi^*, \delta_1) < r(\pi^*, \delta_2)$ . Una regla de decisión que minimice  $r(\pi^*, \delta)$  es óptima, esta recibe el nombre de **regla de Bayes** y será denotada por  $\delta^{\pi^*}$ . La cantidad  $r(\pi^*, \delta^{\pi^*})$  recibe el nombre de **riesgo de Bayes** para  $\pi^*$ .

<sup>1</sup>Una variable aleatoria es un ejemplo de una función medible. En este caso, se pide que la regla de decisión  $\delta$  sea una función medible para que la composición  $\delta(Y)$  sea medible y por lo tanto una variable aleatoria. Para lograr un mejor entendimiento de esta discusión se requieren conceptos básicos de Teoría de la Medida.

### 1.1.2. Estimación y tres funciones de pérdida importantes

Consideremos que nos encontramos en la situación que hemos venido discutiendo, en que  $\theta$  resulta ser un parámetro desconocido de la distribución de  $Y$ . Por ahora suponemos que hemos podido calcular la función de densidad posterior  $\pi(\theta|y)$  (problema que será abordado en las siguientes secciones).

A continuación se desea encontrar un estimador para  $\theta$  dada  $\pi(\theta|y)$ . Para esto vamos a ver esta situación como un problema de teoría de decisión, donde la acción  $a$  que elijamos será el estimador de  $\theta$ .

Recordemos que dada una función de pérdida, se puede calcular la **pérdida esperada posterior**  $\mathbb{E}^{\pi(\theta|y)}[L(\theta, a)]$ , que no es más que la pérdida esperada Bayesiana considerando  $\pi^*(\theta) = \pi(\theta|y)$ . En este caso la acción de Bayes (la acción óptima) será aquella que minimice la pérdida Bayesiana posterior. Es claro que la pérdida esperada posterior depende del resultado del experimento  $Y = y$ , lo que nos hace pensar que la acción de Bayes  $a$  va a depender del resultado  $y$  (es decir, será una función de los posibles valores de  $Y$ ).

Por otro lado, podemos buscar la regla de decisión  $\delta^*$  que minimice el riesgo de Bayes  $r(\pi(\theta|y), \delta)$ , es decir la regla de Bayes. Una vez que la hemos encontrado, y dada una realización  $Y = y$ , basta evaluar  $\delta^*(y)$  y este será el estimador de  $\theta$  (notemos que  $\delta^*(y) \in \mathcal{A}$ , es decir  $\delta^*(y)$  es una acción).

Minimizar la pérdida esperada posterior puede parecer ser un problema mucho más sencillo que minimizar el riesgo de Bayes (ya que encontrar una función que minimice es usualmente mucho más complicado que encontrar un número que minimice). A pesar de esto, ambos problemas son esencialmente equivalentes, y esto es precisamente porque la acción de Bayes  $a$  va a depender del resultado  $y$ , como fue discutido anteriormente.

Formalmente, existe un resultado que dice que una regla de Bayes  $\delta^*$  puede ser encontrada, eligiendo para cada  $y \in \mathcal{Y}$  una acción de Bayes que minimice la pérdida esperada posterior, en dicho caso  $a^* = \delta^*(y)$  (donde  $Y = y$  es el resultado del experimento) se llama **estimador de Bayes**.

A continuación analizaremos este problema para tres funciones de pérdida muy importantes y recurridas en la literatura.

- La pérdida de error cuadrático:

La función de pérdida  $L(\theta, a) = (\theta - a)^2$  se llama **pérdida de error cuadrático**. En este caso la pérdida esperada posterior es

$$\int_{\Theta} (\theta - a)^2 dF^{\pi(\theta|y)}(\theta)$$

El valor de  $a$  que minimiza la expresión anterior lo podemos encontrar expandiendo la expresión cuadrática, derivando e igualando a cero (suponiendo que todas las integrales son finitas).

$$\begin{aligned}
0 &= \frac{d}{da} \left\{ \int_{\Theta} \theta^2 dF^{\pi(\theta|y)}(\theta) - 2a \int_{\Theta} \theta dF^{\pi(\theta|y)}(\theta) + a^2 \int_{\Theta} dF^{\pi(\theta|y)}(\theta) \right\} \\
&= -2 \int_{\Theta} \theta dF^{\pi(\theta|y)}(\theta) + 2a \int_{\Theta} dF^{\pi(\theta|y)}(\theta) \\
&= -2 \int_{\Theta} \theta dF^{\pi(\theta|y)}(\theta) + 2a \\
&= -2\mathbb{E}^{\pi(\theta|y)}[\theta] + 2a.
\end{aligned}$$

Lo que lleva a:

$$a^* = \mathbb{E}^{\pi(\theta|y)}[\theta].$$

Esto da pie al siguiente resultado:

Si  $L(\theta, a) = (\theta - a)^2$ , entonces la regla de Bayes será la media posterior, es decir:

$$\delta^*(y) = \mathbb{E}^{\pi(\theta|y)}[\theta].$$

■ La pérdida de error absoluto:

La función de pérdida  $L(\theta, a) = |\theta - a|$  se llama **pérdida de error absoluto**. En este caso cualquier mediana de  $\pi(\theta|y)$  es un estimador de Bayes de  $\theta$ .

Para ver esto consideremos  $m$  una mediana de  $\pi(\theta|y)$ . Sea  $a \in \mathcal{A}$  otra acción tal que  $a > m$ . Notemos que

$$L(\theta, m) - L(\theta, a) = \begin{cases} m - a & \text{si } \theta \leq m \\ 2\theta - (m + a) & \text{si } m < \theta < a \\ a - m & \text{si } \theta \geq a \end{cases}$$

de donde sigue que

$$L(\theta, m) - L(\theta, a) \leq (m - a)I_{(-\infty, m]}(\theta) + (a - m)I_{(m, \infty)}(\theta).$$

**Nota:** Esto último se debe a que si  $\theta \leq m$  entonces  $L(\theta, m) - L(\theta, a) = m - a$ . De igual forma si  $\theta \geq a > m$  entonces  $L(\theta, m) - L(\theta, a) = a - m$ . Por último, si  $a > \theta > m$  entonces  $2\theta - (m + a) > 2a - (m + a) = a - m$ .

Calculando esperanzas de ambos lados, y considerando que

- $\mathbb{P}[\theta \leq m|y] \geq \frac{1}{2}$  (definición de la mediana),

- $\mathbb{P}[\theta > m|y] \leq \frac{1}{2}$  (definición de la mediana),
- $\mathbb{P}[\theta \leq m|y] = 1 - \mathbb{P}[\theta > m|y]$ , y
- $a - m > 0 > m - a$ ,

se tiene que

$$\begin{aligned} \mathbb{E}^{\pi(\theta|y)}[L(\theta, m)] - \mathbb{E}^{\pi(\theta|y)}[L(\theta, a)] &= \mathbb{E}^{\pi(\theta|y)}[L(\theta, m) - L(\theta, a)] \\ &\leq (m - a)\mathbb{P}[\theta \leq m|y] + (a - m)\mathbb{P}[\theta > m|y] \\ &\leq (m - a)\frac{1}{2} + (a - m)\frac{1}{2} = 0. \end{aligned}$$

De donde podemos concluir que  $\mathbb{E}^{\pi(\theta|y)}[L(\theta, m)] \leq \mathbb{E}^{\pi(\theta|y)}[L(\theta, a)]$  para toda  $a \in \mathcal{A}$  tal que  $a > m$ . Bajo argumentos similares se puede mostrar que  $\mathbb{E}^{\pi(\theta|y)}[L(\theta, m)] \leq \mathbb{E}^{\pi(\theta|y)}[L(\theta, a)]$  también se cumple para toda  $a \in \mathcal{A}$  tal que  $a < m$ .

Con esto queda demostrado que la mediana de la distribución posterior,  $\pi(\theta|y)$  minimiza la pérdida esperada posterior dada la función de pérdida de error absoluto, y por lo tanto es el estimador de Bayes.

■ La pérdida 0-1:

Para definir esta función consideremos  $\{\Theta_i\}_{i \in \mathbb{N}}$  una partición de  $\Theta$  y sea  $\mathcal{A} = \{a_i\}_{i \in \mathbb{N}}$ , donde la acción  $a_i$  indica que se decidió considerar que el estado natural pertenece a  $\Theta_i$  y esto para toda  $i \in \mathbb{N}$ . En la literatura comúnmente se considera que tanto la partición de  $\Theta$  como el conjunto de posibles acciones  $\mathcal{A}$  son finitos, sin embargo, aquí trabajaremos con la generalización al caso numerable.

Entonces si la función de pérdida es de la forma

$$L(\theta, a_i) = \begin{cases} 0 & \text{si } \theta \in \Theta_i \\ 1 & \text{si } \theta \in \Theta_j \quad (j \neq i) \end{cases}$$

se llama **pérdida 0-1**. En este caso es fácil ver que para toda  $i \in \mathbb{N}$  se tiene que

$$\begin{aligned}
\mathbb{E}^{\pi(\theta|y)}[L(\theta, a_i)] &= \int_{\Theta} L(\theta, a_i) dF^{\pi(\theta|y)}(\theta) \\
&= \int_{\Theta_i^C} dF^{\pi(\theta|y)}(\theta) \\
&= \mathbb{P}[\Theta_i^C | y] \\
&= \sum_{\substack{j=1 \\ j \neq i}}^{\infty} \mathbb{P}[\Theta_j | y].
\end{aligned}$$

**Nota:**  $\Theta_i^C$  se refiere al complemento del conjunto  $\Theta_i$ .

Con esto terminamos la parte dedicada a encontrar un estimador de  $\theta$  a través de funciones de pérdida y minimizar pérdidas esperadas. Sin embargo antes de concluir la sección de estimación, cabe mencionar que una forma alternativa de hacer estimación sobre  $\theta$ , dado que se conoce su distribución posterior, es simplemente considerar el valor  $\hat{\theta}$  que maximiza  $\pi(\theta|y)$  (es decir, la moda más grande) conocido como **estimador de máxima verosimilitud generalizada**.

### 1.1.3. Un caso que nos será muy útil

De aquí en adelante, se trabajará con el caso en que  $\pi(\theta|y)$  resulta ser una distribución Normal con media  $\mu$  y varianza  $\sigma^2$ . Por esta razón resulta conveniente analizar cual es el estimador más factible de  $\theta$ . Para empezar notemos que en este caso  $\Theta = \mathbb{R}$ .

- En el caso del estimador máximo verosímil generalizado, es fácil ver que  $\mu = \mathbb{E}^{\pi(\theta|y)}[\theta]$  es el valor que maximiza  $\pi(\theta|y)$ .
- Para el caso en el que se desea encontrar el estimador de Bayes, y consideramos la función de pérdida  $L(\theta, a) = (\theta - a)^2$ , es decir la pérdida de error cuadrático, ya vimos que en este caso el estimador de Bayes es  $\mu = \mathbb{E}^{\pi(\theta|y)}[\theta]$ .
- Nuevamente para encontrar el estimador de Bayes, pero esta vez considerando la función de pérdida de error absoluto, se tiene que este resulta ser la mediana de  $\pi(\theta|y)$ . Sin embargo, debido a que estamos tratando con una distribución Normal, la media posterior coincidirá con la mediana posterior, por lo que en este caso de igual forma tendremos que el estimador de Bayes es  $\mu = \mathbb{E}^{\pi(\theta|y)}[\theta]$ .
- Por último en el caso de la función de pérdida 0-1, se tiene que

$$\mathbb{E}^{\pi(\theta|y)}[L(\theta, a_i)] = \sum_{\substack{j=1 \\ j \neq i}}^{\infty} \mathbb{P}[\Theta_j | y]. \tag{1.1.3.1}$$

Por lo que si consideramos  $\{\Theta_i\}_{i \in \mathbb{N}}$  cualquier partición de  $\mathbb{R}$  tal que todos los intervalos  $\Theta_i$  tengan exactamente la misma longitud, va a resultar que el intervalo digamos  $\Theta_k$  que minimice  $\mathbb{E}^{\pi(\theta|y)}[L(\theta, a_i)]$  será por supuesto aquel tal que  $\mu = \mathbb{E}^{\pi(\theta|y)}[\theta] \in \Theta_k$ . Haciendo cada vez más fina la partición. Podemos ver que en este caso también resulta óptima la media posterior  $\mu$ .

Por lo tanto si la distribución posterior es Normal, entonces el estimador óptimo para  $\theta$  es la media posterior.

Por ahora dejemos a un lado la Teoría de la decisión para presentar otros conceptos básicos y necesarios para de estadística Bayesiana. En lo que sigue del documento vamos a modificar un poco la notación para que sea más amigable. En términos prácticos, esta sección de Teoría de la decisión, simplemente nos servirá como sustento para la estimación del parámetro de interés.

## 1.2. Teorema de Bayes

La estadística Bayesiana toma su nombre precisamente del Teorema de Bayes, la razón por la que esto sucede es porque todo el desarrollo de la estadística Bayesiana surge como una consecuencia del siguiente e importante teorema. Considere el caso más sencillo y general.

Sean  $A$  y  $B$  dos eventos, entonces:

$$\mathbb{P}[A \cap B] = \mathbb{P}[A|B]\mathbb{P}[B] = \mathbb{P}[B|A]\mathbb{P}[A]$$

donde  $\mathbb{P}[A \cap B]$  es la probabilidad de que sucedan  $A$  y  $B$ ,  $\mathbb{P}[A|B]$  es la probabilidad de que suceda  $A$  dado que se sabe que ha sucedido  $B$  y  $\mathbb{P}[B]$  es la probabilidad marginal de  $B$ .

o dicho de otra forma:

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \frac{\mathbb{P}[B|A]\mathbb{P}[A]}{\mathbb{P}[B]}. \quad (1.2.0.2)$$

## 1.3. Conceptos básicos

En Estadística Bayesiana se considera que  $A$  es un evento de interés y que  $B$  es un resultado experimental que puede dar información acerca de  $A$ , es claro que tanto el evento de interés como el resultado experimental y la relación entre ellos depende del problema que se desea resolver.

Es común en la práctica de la inferencia estadística considerar una familia de variables o vectores aleatorios  $Y = \{Y_i : i = 1, 2, \dots\}$  que tienen cierta función de distribución con un vector de parámetros desconocido  $\theta$ . Supongamos que es posible tomar una muestra de la familia de vectores aleatorios  $y_{1:n} = \{y_1, y_2, \dots, y_n\}$ ,  $n \in \mathbb{N}$ , en este caso el resultado experimental serán las observaciones que se han tomado de  $Y$  y el evento de interés resulta ser el vector de parámetros  $\theta$ .

Se definen las siguientes funciones:

- La *función de verosimilitud*,  $\pi(y_{1:n}|\theta)$ , es la distribución de  $\{Y_i : i = 1, 2, \dots, n\}$  dado que se conoce el valor de sus parámetros  $\theta$ .
- La *distribución inicial*,  $\pi(\theta)$ , es la información inicial (o incertidumbre) que se tiene acerca de  $\theta$ .
- La *distribución posterior*,  $\pi(\theta|y_{1:n})$ , es la información acerca de  $\theta$  dado que se tiene el resultado experimental  $y_{1:n}$ .

**Nota:** En la sección anterior usamos la notación  $\pi(\theta|y)$  para describir a la función de densidad posterior, a partir de este momento estaremos usando la notación  $\pi(\theta|y_{1:n})$  para describirla, esta

nueva notación nos permite saber con respecto a que observaciones se está condicionando. De manera análoga sucede con la función de verosimilitud.

La estadística Bayesiana busca obtener la distribución posterior, para ver que valores, de los que posiblemente puede tomar el vector de parámetros, son más factibles tomando en cuenta la información que proporcionan las observaciones. Supongamos que basados en lo que se conoce acerca del problema, es posible asignar la verosimilitud  $\pi(y_{1:n}|\theta)$  para  $Y$  dado  $\theta$  y que además podemos expresar la incertidumbre o la información inicial de  $\theta$  mediante su distribución inicial  $\pi(\theta)$ . Entonces podemos escribir a la distribución posterior en términos de la función de verosimilitud y la distribución inicial usando una generalización del teorema de Bayes ( 1.2.0.2):

$$\pi(\theta|y_{1:n}) = \frac{\pi(y_{1:n}|\theta)\pi(\theta)}{\pi(y_{1:n})}. \quad (1.3.0.3)$$

Como  $\theta$  es un parámetro de la distribución de  $y_{1:n}$ , si no conocemos  $\theta$  es claro que  $\pi(y_{1:n})$  (la función de densidad conjunta) no la podemos escribir explícitamente, para resolver este problema escribimos a  $\pi(y_{1:n})$  de la siguiente forma:

$$\begin{aligned} \pi(y_{1:n}) &= \int \pi(y_{1:n} \cap \theta) d\theta = \int \pi(y_{1:n}|\theta)\pi(\theta) d\theta \\ \therefore \pi(\theta|y_{1:n}) &= \frac{\pi(y_{1:n}|\theta)\pi(\theta)}{\int \pi(y_{1:n}|\theta)\pi(\theta) d\theta}. \end{aligned}$$

En muchas ocasiones es muy complicado calcular  $\int \pi(y_{1:n}|\theta)\pi(\theta) d\theta$ . Sin embargo, si consideramos a  $\pi(\theta|y_{1:n})$  como función de  $\theta$  es claro que  $\int \pi(y_{1:n}|\theta)\pi(\theta) d\theta$  resulta ser una constante, por lo que podemos decir que:

$$\pi(\theta|y_{1:n}) \propto \pi(y_{1:n}|\theta)\pi(\theta). \quad (1.3.0.4)$$

Esto es,  $\pi(\theta|y_{1:n})$  es proporcional a  $\pi(y_{1:n}|\theta)\pi(\theta)$ .

Como  $\pi(\theta|y_{1:n})$  es una distribución, después de multiplicar la función de verosimilitud por la distribución inicial, puede conocerse el valor  $\int \pi(y_{1:n}|\theta)\pi(\theta) d\theta$  tratándola como una constante de normalización.

Otro problema que es interesante es el de *predecir* cual será el valor de la siguiente observación, es decir si hemos tomado las observaciones  $y_1, y_2, \dots, y_n$  es de interés conocer que valor es más probable que tome  $y_{n+1}$  tomando en cuenta toda la información que hemos recabado.

Definamos:

- La *distribución predecitiva*,  $\pi(y_{n+1}|y_{1:n})$ , es la información que se tiene acerca de  $y_{n+1}$  dada la muestra  $y_{1:n} = \{y_1, y_2, \dots, y_n\}$ .

En ocasiones resulta conveniente expresar a la distribución predictiva de la siguiente forma:

$$\pi(y_{n+1}|y_{1:n}) = \int \pi(y_{n+1}, \theta|y_{1:n})d\theta = \int \pi(y_{n+1}|\theta, y_{1:n})\pi(\theta|y_{1:n})d\theta.$$

## 1.4. Estructuras de dependencia

### 1.4.1. Independencia Condicional

En una gran parte de aplicaciones es razonable suponer que  $Y_1, Y_2, \dots, Y_n$  (para cualquier  $n \in \mathbb{N}$ ) son independientes e idénticamente distribuidas (i.i.d) dado  $\theta$ , esto es,  $\pi(y_{1:n}|\theta) = \prod_{i=1}^n \pi(y_i|\theta)$ .

Si es el caso, podemos reescribir (1.3.0.3) como:

$$\pi(\theta|y_{1:n}) = \frac{\pi(y_{1:n}|\theta)\pi(\theta)}{\pi(y_{1:n})} = \frac{\prod_{i=1}^n \pi(y_i|\theta)\pi(\theta)}{\pi(y_{1:n})}.$$

De lo que se sigue que:

$$\pi(\theta|y_{1:n}) \propto \prod_{i=1}^n \pi(y_i|\theta)\pi(\theta). \quad (1.4.1.1)$$

Esto significa, podemos ir actualizando la distribución posterior conforme vayamos recabando nueva información de manera muy sencilla. Suponga que para alguna  $n \in \mathbb{N}$  hemos recabado las observaciones  $y_1, y_2, \dots, y_n$  y que hemos calculado la distribución posterior  $\pi(\theta|y_{1:n})$ , supongamos que eventualmente se nos brinda un dato más  $y_{n+1}$  y que nos interesa incorporar la información que este nuevo dato nos brinda a la distribución posterior. Es decir, queremos conocer  $\pi(\theta|y_{1:n+1})$ .

La ecuación 1.4.1.1 indica que:

$$\pi(\theta|y_{1:n+1}) \propto \prod_{i=1}^{n+1} \pi(y_i|\theta)\pi(\theta) = \pi(y_{n+1}|\theta)\pi(\theta|y_{1:n}). \quad (1.4.1.2)$$

Es decir la nueva distribución posterior  $\pi(\theta|y_{1:n+1})$  es proporcional a la función de verosimilitud de  $y_{n+1}$  dado  $\theta$  por la distribución posterior que ya habíamos calculado  $\pi(\theta|y_{1:n})$ , de hecho, se puede pensar que para este nuevo problema  $\pi(\theta|y_{1:n})$  juega el papel de la distribución inicial.

Por lo que para resolver el nuevo problema basta multiplicar  $\pi(\theta|y_{1:n})$  por  $\pi(y_{n+1}|\theta)$  y calcular la nueva constante de normalización.

### 1.4.2. Un primer ejemplo

Imagine que una hormiga (conocedora de conceptos básicos de probabilidad y estadística) quiere transportar un pedazo de hoja a su hogar, debido a que esta hoja es muy grande obstruye su visión, y solamente cuando viento le ayuda a levantarla, la hormiga logra ver su hogar a lo lejos así como visualizar su camino. Esta hormiga quiere saber a que distancia se encuentra de su destino para administrar su energía. Cada vez que logra ver su hogar ella toma una observación empírica  $y_i$  de la distancia  $\theta$  entre ella y su destino. La hormiga ha logrado conseguir  $n$  observaciones  $y_1, y_2, \dots, y_n$  tales que:

$$Y_i = \theta + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$$

Las observaciones está afectadas por un error aleatorio debido a la percepción de la hormiga. Suponga que  $\theta$  y las  $\varepsilon_i$ 's son independientes y además que la hormiga tiene el conocimiento de qué tanta desviación tienen sus observaciones, por lo que para ella  $\sigma^2$  es una constante conocida. En otras palabras:

$$Y_i|\theta \sim \mathcal{N}(\theta, \sigma^2) \quad \forall i = 1, 2, \dots$$

También suponga que la hormiga tiene un conocimiento inicial sobre la distancia que la separa de su hogar y conviene en expresarla de la siguiente manera:

$$\theta \sim \mathcal{N}(\mu_0, C_0).$$

Notemos que si la hormiga no tiene tan claro a qué distancia se encuentra de su hogar, es decir, el conocimiento inicial de  $\theta$  no le es muy claro, le basta asignar un valor grande a  $C_0$  para que sea menos informativa acerca de  $\theta$ .

Otra manera de pensar este problema es que se desea conocer la posición  $\theta$  de una partícula que vive en un mundo univariado (unidimensional) es decir  $\theta \in S \subseteq \mathbb{R}$ . Se han tomado observaciones  $y_1, y_2, \dots, y_n$  de la posición de la partícula, estas están afectadas por un error aleatorio, de forma que:

$$Y_i = \theta + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2),$$

donde los errores  $\varepsilon_i$ 's y  $\theta$  son independientes y  $\sigma^2$  es una constante conocida.

En otras palabras:

$$Y_i|\theta \sim \mathcal{N}(\theta, \sigma^2) \quad \forall i = 1, 2, \dots \tag{1.4.2.1}$$

Supongamos que se conviene en expresar el conocimiento inicial que se tiene sobre  $\theta$  de la siguiente manera:

$$\theta \sim \mathcal{N}(\mu_0, C_0).$$

Una vez más, si a  $C_0$  le asignamos un valor grande, damos menos información de  $\theta$  que si le damos un valor pequeño. Es decir si la incertidumbre acerca de  $\theta$  es mucha, podemos

elegir un valor de  $C_0$  muy grande para ser menos informativos.

Lo primero que se hace en cualquiera de los dos planteamientos (que obviamente son equivalentes) es expresar la función de verosimilitud y la distribución inicial.

- La función de verosimilitud:  $\pi(y_{1:n}|\theta) = \prod_{i=1}^n \pi(y_i|\theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y_i - \theta)^2}{2\sigma^2}\right\}$ .
- La distribución inicial:  $\pi(\theta) = \frac{1}{\sqrt{2\pi C_0}} \exp\left\{-\frac{(\theta - \mu_0)^2}{2C_0}\right\}$ .

**Proposición 1.4.2.1.** Si  $\theta \sim \mathcal{N}(\mu_0, C_0)$ ,  $Y_i|\theta \sim \mathcal{N}(\theta, \sigma^2)$  para toda  $i \in \mathbb{N}$  y las  $Y_i$ 's son i.i.d dado  $\theta$ . Entonces para cualquier  $n \in \mathbb{N}$ :

$$\theta|y_{1:n} \sim \mathcal{N}(\mu_n, C_n).$$

donde

$$\mu_n = \mathbb{E}[\theta|y_{1:n}] = \frac{C_0}{C_0 + \sigma^2/n} \bar{y} + \frac{\sigma^2/n}{C_0 + \sigma^2/n} \mu_0, \quad (1.4.2.2)$$

$$C_n = \text{Var}(\theta|y_{1:n}) = \frac{\sigma^2 C_0}{\sigma^2 + n C_0} = \left( \frac{n}{\sigma^2} + \frac{1}{C_0} \right)^{-1}, \quad (1.4.2.3)$$

y

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i.$$

**Demostración:**

Ya habíamos visto  $\pi(\theta|y_{1:n}) \propto \pi(y_{1:n}|\theta)\pi(\theta)$  para toda  $n \in \mathbb{N}$  (ecuación 1.3.0.4). Como las  $Y_i$ 's son i.i.d dado  $\theta$  se cumple la ecuación 1.4.1.1, por lo que podemos escribir:

$$\begin{aligned}
\pi(\theta|y_{1:n}) &\propto \prod_{i=1}^n \pi(y_i|\theta)\pi(\theta) \\
&\propto \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{-\frac{(y_i - \theta)^2}{2\sigma^2}\right\} \frac{1}{\sqrt{2\pi C_0}} \exp\left\{-\frac{(\theta - \mu_0)^2}{2C_0}\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma^2} \left(\sum_{i=1}^n y_i^2 - 2\theta \sum_{i=1}^n y_i + n\theta^2\right) - \frac{1}{2C_0} (\theta^2 - 2\theta\mu_0 + \mu_0^2)\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma^2 C_0} \left(C_0 \sum_{i=1}^n y_i^2 - 2\theta n C_0 \bar{y} + n\theta^2 C_0 + \sigma^2 \theta^2 - 2\theta \sigma^2 \mu_0 + \mu_0^2 \sigma^2\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma^2 C_0} (\theta^2 (nC_0 + \sigma^2) - 2\theta (nC_0 \bar{y} + \sigma^2 \mu_0))\right\} \\
&\propto \exp\left\{-\frac{nC_0 + \sigma^2}{2\sigma^2 C_0} \left(\theta^2 - 2\theta \left(\frac{nC_0 \bar{y} + \sigma^2 \mu_0}{nC_0 + \sigma^2}\right)\right)\right\} \\
&\propto \exp\left\{-\frac{nC_0 + \sigma^2}{2\sigma^2 C_0} \left(\theta^2 - 2\theta \left(\frac{nC_0 \bar{y} + \sigma^2 \mu_0}{nC_0 + \sigma^2}\right) + \left(\frac{nC_0 \bar{y} + \sigma^2 \mu_0}{nC_0 + \sigma^2}\right)^2\right)\right\} \\
&\propto \exp\left\{-\frac{nC_0 + \sigma^2}{2\sigma^2 C_0} \left(\theta - \frac{nC_0 \bar{y} + \sigma^2 \mu_0}{nC_0 + \sigma^2}\right)^2\right\}.
\end{aligned}$$

Podemos reconocer que la última línea representa al kernel<sup>2</sup> de una distribución  $\mathcal{N}(\mu_n, C_n)$ , es decir  $\theta|y_{1:n} \sim \mathcal{N}(\mu_n, C_n)$ , donde:

$$\mu_n = \mathbb{E}[\theta|y_{1:n}] = \frac{C_0}{C_0 + \sigma^2/n} \bar{y} + \frac{\sigma^2/n}{C_0 + \sigma^2/n} \mu_0 \quad (1.4.2.4)$$

y

$$C_n = \text{Var}(\theta|y_{1:n}) = \frac{\sigma^2 C_0}{\sigma^2 + nC_0} = \left(\frac{n}{\sigma^2} + \frac{1}{C_0}\right)^{-1}. \quad (1.4.2.5)$$

**Nota:** En la sección anterior denotamos a la media de la distribución posterior como  $\mathbb{E}^{\pi(\theta|y)}[\theta]$  en esta sección estaremos usando la notación  $\mathbb{E}[\theta|y_{1:n}]$ .

Con esto queda demostrado que si las  $Y_i$ 's son i.i.d con distribución Normal dada su media  $\theta$  y que si la información inicial de  $\theta$  la escribimos a través de una distribución normal, entonces la distribución posterior de  $\theta$  dada la muestra también será Normal con los parámetros descritos anteriormente.

---

<sup>2</sup>kernel se refiere a la parte de la función de densidad que no es constante

Notemos que la media posterior  $\mu_n$  es la suma la media inicial  $\mu_0$  multiplicada por un peso  $k_{\mu_0}$  más la media muestral  $\bar{y}$  multiplicada por  $k_{\bar{y}} = 1 - k_{\mu_0}$ . Los pesos que se le asignan a cada una de las medias ( $k_{\mu_0}$  y  $k_{\bar{y}}$ ) dependen únicamente que de la varianza inicial  $C_0$ , de la varianza de los errores  $\sigma^2$  y del número de observaciones  $n$  que se hayan tomado. Es claro que:

- Si  $C_0 > \frac{\sigma^2}{n}$  entonces  $\bar{y}$  tendrá un mayor peso que  $\mu_0$ .

Es decir, entre más pequeña sea la varianza de los errores, entre más observaciones se hayan tomado, y entre más grande sea la varianza inicial, la media posterior se verá más afectada por la media muestral.

- Si  $C_0 < \frac{\sigma^2}{n}$  entonces  $\bar{y}$  tendrá un menor peso que  $\mu_0$ .

Esto es, si la varianza inicial es muy pequeña (se tiene poca incertidumbre acerca de la información inicial de  $\theta$ ), las observaciones tienen desviaciones grandes, y no se han tomado tantas observaciones, entonces la media posterior dependerá más de la media inicial.

- Si  $C_0 = \frac{\sigma^2}{n}$  entonces  $\bar{y}$  y  $\mu_0$  tendrán el mismo peso.

Si la varianza inicial es exactamente igual a la varianza de los errores entre el número de observaciones tomadas entonces la media posterior será el promedio de la media inicial  $\mu_0$  y la media muestral  $\bar{y}$ . Notemos que al tener disponible una o más observación extras, la media posterior se va viendo más afectada por la media muestral.

Esto era de esperarse ya que  $C_0$  representa la incertidumbre que se tiene acerca del conocimiento inicial de  $\theta$ ,  $\sigma^2$  representa que tan grandes serán las desviaciones de las observaciones con respecto a  $\theta$  y al tomar más observaciones se recaba más información de  $\theta$  lo que hace que la información empírica o inicial de  $\theta$  pierda importancia. Esto último resulta ser muy conveniente, ya que a veces la elección que se hace acerca de la distribución inicial puede no ser tan exacta, sin embargo, este problema puede ser minimizado tomando muchas observaciones.

En cuánto a la precisión posterior  $\frac{1}{C_n} = \frac{n}{\sigma^2} + \frac{1}{C_0}$  podemos ver que es la suma de la precisión muestral  $\left(\frac{\sigma^2}{n}\right)^{-1}$  y la precisión inicial  $(C_0)^{-1}$ , evidentemente la precisión posterior será mayor que la precisión muestral y que la precisión inicial ya que estás dos últimas son números positivos, también es fácil ver que entre más grande sea el tamaño de la muestra  $n$ , más grande será la precisión posterior.

Como ya se había visto anteriormente es fácil actualizar la distribución posterior conforme tengamos acceso a más observaciones, ya que  $\pi(\theta|y_{1:n+1}) \propto \pi(y_{n+1})\pi(\theta|y_{1:n})$ , donde

$\pi(y_{n+1})$  juega el papel de la verosimilitud y  $\pi(\theta|y_{1:n})$  juega el papel de la distribución inicial. En este caso  $Y_{n+1}|\theta \sim \mathcal{N}(\theta, \sigma^2)$  y  $\theta|y_{1:n} \sim \mathcal{N}(\mu_n, C_n)$ , como pudimos darnos cuenta, si la distribución inicial y la verosimilitud tienen una distribución Normal, entonces la distribución posterior (de la media)  $\pi(\theta|y_{1:n+1})$  también será una distribución Normal, con parámetros:

$$\begin{aligned}\mu_{n+1} &= \left(\frac{C_n}{C_n + \sigma^2}\right) y_{n+1} + \left(1 - \frac{C_n}{C_n + \sigma^2}\right) \mu_n \\ &= \mu_n + \frac{C_n}{C_n + \sigma^2} (y_{n+1} - \mu_n)\end{aligned}$$

y varianza

$$C_{n+1} = \left(\frac{1}{\sigma^2} + \frac{1}{C_n}\right)^{-1} = \frac{\sigma^2 C_n}{\sigma^2 + C_n}.$$

Ahora, imaginemos que ya tenemos los datos  $y_1, y_2, \dots, y_m$  para alguna  $m \in \mathbb{N}$ , también resulta de interés predecir cual será el valor de  $y_{m+1}$ . Como  $Y_{m+1} = \theta + \varepsilon_{m+1}$ , entonces  $Y_{m+1}|y_{1:m} = \theta|y_{1:m} + \varepsilon_{m+1}$  (ya que  $\varepsilon_{m+1}$  es independiente de  $y_{1:m}$ ). Como  $\theta|y_{1:m} \sim \mathcal{N}(\mu_m, C_m)$ , y es independiente de  $\varepsilon_{m+1} \sim \mathcal{N}(0, \sigma^2)$ , entonces  $Y_{m+1} \sim \mathcal{N}(\mu_m, C_m + \sigma^2)$ , por lo que  $\mu_m$  es el valor esperado posterior de  $\theta$  y también será nuestra la predeción de un paso adelante  $\mathbb{E}[Y_{m+1}|y_{1:m}] = \mu_m$ .

### 1.4.3. Intercambiabilidad

La estructura de dependencia básica de la estadística Bayesiana es la intercambiabilidad. Consideremos de nuevo una secuencia infinita de variables o vectores aleatorios  $\{Y_i : i = 1, 2, \dots\}$ . La intercambiabilidad se refiere a que el orden de la secuencia no es relevante, en el sentido de que para toda  $n \geq 1$   $(Y_1, Y_2, \dots, Y_n) \stackrel{d}{=} (Y_{p_1}, Y_{p_2}, \dots, Y_{p_n})$  (es decir  $(Y_1, Y_2, \dots, Y_n)$  tiene la misma distribución que  $(Y_{p_1}, Y_{p_2}, \dots, Y_{p_n})$ ) donde  $\{p_1, p_2, \dots, p_n\}$  es cualquier permutación de  $\{1, 2, \dots, n\}$ , es decir . Existe un resultado importante conocido como el teorema de representación de *de Finetti* que afirma que la suposición de intercambiabilidad es equivalente a la suposición de independencia (e idéntica distribución) condicional.

**Teorema 1.4.3.1** (Teorema de representación de de Finetti). *Sea  $\{Y_i : i = 1, 2, \dots\}$  una secuencia infinita de vectores aleatorios intercambiables. Entonces*

(i) *Con probabilidad 1, la secuencia de funciones de distribución empíricas*

$$F_n(y) = F_n(y; Y_1, Y_2, \dots, Y_n) = \frac{1}{n} \sum_{i=1}^n I_{(-\infty, y]}(Y_i)$$

*converge débilmente a una función de distribución aleatoria  $F$  cuando  $n \rightarrow \infty$ .*

(ii) Para cualquier  $n \geq 1$ , la función de distribución de  $(Y_1, Y_2, \dots, Y_n)$  puede ser representada como

$$\mathbb{P}[Y_1 \leq y_1, Y_2 \leq y_2, \dots, Y_n \leq y_n] = \int \prod_{i=1}^n \pi(y_i) d\pi(F)$$

donde  $\pi$  es la ley de probabilidad del límite débil  $F$  de la secuencia de funciones de distribución empíricas.

Notemos que este teorema nos dice que si suponemos que la secuencia  $\{Y_i\}$  es intercambiable, entonces podemos pensarla como una secuencia de vectores (o variables) aleatorios condicionalmente independientes e idénticamente distribuidos dada la función de distribución  $F$  (donde  $F$  es el límite débil de las funciones de distribución empíricas). La distribución inicial  $\pi$  se refiere a ley de probabilidad sobre el espacio  $\mathcal{F}$  de todas las funciones de distribución sobre el espacio muestral  $\mathcal{Y}$  y expresa nuestras creencias acerca del límite de las funciones de distribución empíricas. En el caso de una distribución inicial  $\pi$  paramétrica, el teorema de representación implica que  $Y_1, Y_2, \dots$  son i.i.d. con distribución común  $\pi(\bullet|\theta)$  y  $\theta$  tiene una distribución inicial  $\pi(\theta)$ , esta es precisamente la independencia condicional i.i.d. que discutimos anteriormente.

#### 1.4.4. Observaciones heterogéneas

La intercambiabilidad es la forma de dependencia más simple y es apropiada cuando se cree que las observaciones son homogéneas. Sin embargo, en muchos casos la dependencia de las observaciones es más compleja y con frecuencia resulta adecuado asumir heterogeneidad entre las observaciones. Esto es suponer que para toda  $n \geq 1$

$$(Y_1, Y_2, \dots, Y_n) | (\theta_1, \theta_2, \dots, \theta_n) \sim \prod_{i=1}^n f_i(y_i | \theta_i).$$

En otras palabras,  $Y_1, \dots, Y_n$  son condicionalmente independientes dado un vector  $(\theta_1, \dots, \theta_n)$  donde  $Y_i$  depende únicamente de  $\theta_i$ . Es claro que si  $\theta_i = \theta_j$  y  $f_i = f_j$  para toda  $i, j = 1, 2, \dots$ , entonces regresamos al caso en que las observaciones son homogéneas.

El tipo de dependencia heterogénea es con el que estaremos trabajando en la parte restante del documento, y más adelante regresaremos a discutirla.

## 2. Modelos Dinámicos Lineales: Aspectos Teóricos.

### 2.1. Un primer ejemplo de Modelos Dinámicos Lineales

Empecemos con un pequeño ejemplo. Recordemos que en capítulo pasado se vió el planteamiento de un problema donde una hormiga (extremadamente inteligente) quería calcular la distancia a la que se encontraba de su casa. A diferencia de ese ejemplo, esta vez la hormiga se mueve por lo que la distancia entre ella y su hogar  $\theta_t$  depende del tiempo  $t$ .

En este caso, las observaciones que la hormiga logra hacer se comportaran de la siguiente forma:

$$Y_t = \theta_t + \varepsilon_t, \quad \text{donde} \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2).$$

Notemos que  $Y_t|\theta_t \sim \mathcal{N}(\theta_t, \sigma^2)$  y es independiente de  $Y_{1:t-1}$  y de  $\theta_{1:t-1}$ . La hormiga tiene la posibilidad de expresar el conocimiento inicial de su posición mediante una distribución Normal de forma que conoce que  $\theta_0 \sim \mathcal{N}(\mu_0, C_0)$  donde  $\mu_0 = 10$  y  $C_0 = 2$ . La hormiga no se moverá aún, así que sabe que la distancia no cambiará para el siguiente tiempo, por lo que  $\theta_1 = \theta_0$ , lo cual significa que  $\theta_1 \sim \mathcal{N}(a_1, R_1)$  donde  $a_1 = \mu_0$  y  $R_1 = C_0$ . Imaginemos también que la hormiga puede cuantificar que tanta desviación tienen sus observaciones de la realidad por lo que asigna  $\sigma^2 = 0.4$ .

En este momento la hormiga puede hacer una predicción de la primera observación:  $f_1 = \mathbb{E}[Y_1] = \mathbb{E}[\theta_1 + \varepsilon_1] = \mathbb{E}[\theta_1] = a_1 = 10$ .

Al tiempo  $t = 1$  la hormiga consigue su primera observación  $Y_1 = y_1 = 10.426$  y calcula el error de su predicción  $e_1 = y_1 - f_1$ .

Para poder calcular la distancia que la separa de su hogar al tiempo  $t = 1$  dada su primera observación la hormiga hace lo siguiente: Considera  $\pi(\theta_1)$  como la distribución inicial y a  $\pi(y_1|\theta_1)$  como la verosimilitud y teniendo en cuenta que  $\theta_1 \sim \mathcal{N}(a_1, R_1)$  y  $Y_1|\theta_1 \sim \mathcal{N}(\theta_1, \sigma^2)$  se tiene que

$\theta_1|y_1 \sim \mathcal{N}(\mu_1, C_1)$  donde:

$$\begin{aligned} \mu_1 &= a_1 + \frac{R_1}{R_1 + \sigma^2}(y_1 - a_1) \\ &= a_1 + \frac{R_1}{R_1 + \sigma^2}(y_1 - f_1) = 10.332 \end{aligned}$$

y

$$C_1 = \left( \frac{1}{R_1} + \frac{1}{\sigma^2} \right)^{-1} = 0.312.$$

Démonos cuenta que  $\mu_1$  se obtiene sumándole a  $\mathbb{E}[\theta_1]$  el error de predicción  $e_1 = y_1 - f_1$  multiplicado por un factor  $K_1 = \frac{R_1}{R_1 + \sigma^2}$  que depende de la varianza inicial y la varianza de los errores. Esta es una forma de corregir la información que se tenía de  $\theta_1$  con la información que nos proporciona  $y_1$ . No está de más hacer notar que esto nos indica que entre más pequeña sea  $\sigma^2$  y más grande  $R_1$  se confía más en las observaciones y se tiene más incertidumbre acerca del valor inicial de  $\theta$ .

La hormiga continua sin moverse, entonces  $\theta_2 = \theta_1$ , por lo que  $\theta_2|y_1 \sim \mathcal{N}(a_2, R_2)$  donde  $a_2 = \mu_1$  y  $R_2 = C_1$ , en este tiempo la hormiga puede predecir cual será su siguiente observación calculando  $f_2$  de igual manera que como calculó  $f_1$ . Cuando la hormiga tiene acceso a su segunda observación  $Y_2 = y_2 = 10.965$ , le es posible calcular el error de predicción  $e_2$  y la f.d.p. de  $\theta_2|y_1, y_2$  de manera similar a como lo hizo anteriormente, con diferencia de que esta vez tomará  $\pi(\theta_2|y_1)$  como distribución inicial y  $\pi(y_2|\theta_2)$  como verosimilitud, de forma que obtiene:

$$\theta_2|y_1, y_2 \sim \mathcal{N}(\mu_2, C_2) \text{ donde } \mu_2 = 10.609 \text{ y } C_2 = 0.175.$$

Al tiempo  $t = 2$  la hormiga decide comenzar a moverse en dirección a su casa, pretende acercarse a velocidad constante  $v$  (unidades de distancia/ unidades de tiempo), sin embargo, debido a la fuerza del viento está velocidad se ve afectada por un factor aleatorio  $w_3$  de manera que:

$$\theta_3 = \theta_2 - v + w_3, \quad w_3 \sim \mathcal{N}(0, \sigma_w^2),$$

donde  $\theta_2$  y  $w_3$  son independientes,  $v = 4$  y  $\sigma_w^2 = 0.9$ . El proceso observable  $\{Y_t\}$  continua dependiendo de  $\{\theta_t\}$  de la misma manera que lo hacía, es decir al tiempo  $t = 3$ ,  $Y_3 = \theta_3 + \varepsilon_3$ .

La hormiga continua repitiendo los pasos que ha venido haciendo desde el principio del ejemplo para poder estimar su nueva posición  $\theta_3$ .

- **Paso inicial:** Al tiempo  $t = 2$  se tiene que

$$\theta_2|y_1, y_2 \sim \mathcal{N}(\mu_2, C_2).$$

- **Paso predictivo:** Al tiempo  $t = 2$  la hormiga puede intentar adivinar a que distancia se encontrará al siguiente tiempo, como  $\theta_3 = \theta_2 - v + w_3$  entonces  $\theta_3|y_1, y_2 = (\theta_2 - v + w_3)|y_1, y_2 = \theta_2|y_1, y_2 - v + w_3$  (esto ya que  $v$  es una constante y  $w_3$  es independiente de  $y_1$  y  $y_2$ ) lo que lleva a que

$$\theta_3|y_1, y_2 \sim \mathcal{N}(a_3, R_3).$$

con

$$\begin{aligned} a_3 &= \mathbb{E}[\theta_3|y_1, y_2] = \mathbb{E}[\theta_2|y_1, y_2 - v + w_3] = \mu_2 + v = 6.609, \\ R_3 &= \text{Var}(\theta_3|y_1, y_2) = \text{Var}(\theta_2|y_1, y_2 - v + w_3) = C_2 + \sigma_w^2 = 1.075 \end{aligned}$$

También puede predecir cual será la siguiente observación. Como  $Y_3 = \theta_3 + \varepsilon_3$ , entonces

$$Y_3|y_1, y_2 \sim \mathcal{N}(f_3, Q_3)$$

con

$$\begin{aligned} f_3 &= \mathbb{E}[Y_3|y_1, y_2] = \mathbb{E}[\theta_3|y_1, y_2 + \varepsilon_3] = a_3 = 6.609, \\ Q_3 &= \text{Var}(Y_3|y_1, y_2) = \text{Var}(\theta_3|y_1, y_2 + \varepsilon_3) = R_3 + \sigma^2 = 1.075 \end{aligned}$$

- Paso de estimación (filtración):** Al tiempo  $t = 3$  la hormiga obtiene una nueva observación  $Y_3 = y_3 = 6.189$ . Para el tiempo  $t = 3$  puede calcular el error de predicción  $e_3 = y_3 - f_3 = -0.420$ . Esto dice que ha sobre estimado  $Y_3$  e intuitivamente también ha sobre estimado  $\theta_3$  por lo que  $\mu_3 = \mathbb{E}[\theta_3|y_{1:3}]$  será menor que  $a_3 = \mathbb{E}[\theta_3|y_{1:2}]$ . Para calcular la distribución posterior de  $\theta_3|y_{1:3}$  toma a  $\pi(\theta_3|y_{1:2})$  (la información que se tiene acerca de  $\theta_3$ ) como la distribución inicial y la verosimilitud será  $\pi(y_3|\theta_3)$ , como ya se ha visto anteriormente la distribución inicial y la verosimilitud son distribuciones Normales  $\theta_3|y_{1:2} \sim \mathcal{N}(a_3, R_3)$  y  $Y_3|\theta_3 \sim \mathcal{N}(\theta_3, \sigma^2)$  por lo que la distribución posterior de  $\theta_3|y_{1:3}$  será Normal con parámetros:

$$\mu_3 = \mathbb{E}[\theta_3|y_{1:3}] = a_3 + \frac{R_3}{R_3 + \sigma^2}(y_3 - a_3) = 6.303$$

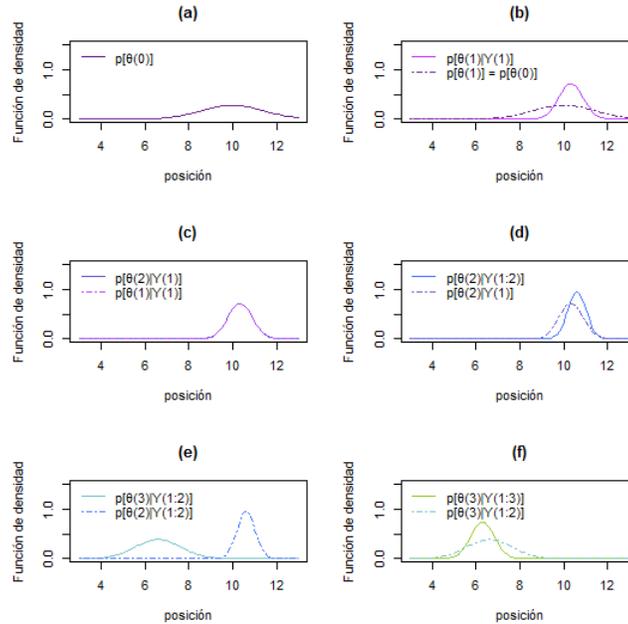
y

$$C_3 = Var(\theta_3|y_{1:3}) = \left( \frac{1}{R_3} + \frac{1}{\sigma^2} \right)^{-1} = 0.292.$$

La hormiga puede proceder repitiendo estos 3 pasos recursivamente para actualizar las estimaciones y predicciones conforme tenga disponibles nuevas observaciones.

A continuación se muestra una gráfica de como evoluciona f.d.p. de la posición de la hormiga con respecto a su hogar en estos tres primeros tiempos:

Figura 2.1.0.1: Funciones de densidad de la posición de la hormiga con respecto a su hogar



En las gráficas anteriores  $p(\bullet)$  denota la función de densidad de  $\bullet$ .

Además usamos  $\theta(\ast)$  para denotar a  $\theta_\ast$  y análogamente  $Y(\ast)$  se refiere a  $Y_\ast$ .

La gráfica (a) muestra la función de densidad de  $\theta_0 \sim \mathcal{N}(10, 2)$ .

La gráfica (b) muestra la función inicial de densidad de  $\theta_1 = \theta_0$  con una línea punteada y también muestra la función posterior de densidad de  $\theta_1|y_1 \sim \mathcal{N}(10.332, 0.312)$ . Como se puede ver, es claro que la varianza posterior es menor que la varianza inicial, esto es porque conforme se va recabando información la incertidumbre disminuye, así mismo podemos darnos cuenta de que la media posterior se encuentra entre la media inicial  $a_1 = 10$  y la primera observación  $y_1 = 10.426$ , y que se encuentra a una menor distancia  $y_1$  que de  $a_1$ , esto es porque la varianza inicial es mayor que la varianza de los errores que afectan las observaciones.

En la tercera gráfica, (c), se pueden apreciar la función de densidad posterior de  $\theta_1|y_1$  y la función de densidad inicial de  $\theta_2|y_1$ , como  $\theta_2 = \theta_1$ , es decir la hormiga aún no se movía, estas dos variables aleatorias tienen la misma función de densidad.

La gráfica (d) contiene la función de densidad inicial de  $\theta_2|y_1 \sim \mathcal{N}(10.332, 0.312)$  así como la función de densidad posterior de  $\theta_2|y_{1:2} \sim \mathcal{N}(10.609, 0.175)$ . De la misma forma que sucedió con las funciones de densidad inicial y posterior de  $\theta_1$ , la varianza posterior es menor que la varianza inicial. Recordemos que la segunda observación  $y_2 = 10.965$  por lo que la media posterior  $\mu_2 = 10.609$  se encuentra entre la media inicial  $a_2 = 10.332$  y esta última observación, podemos ver que la media posterior se encuentra muy cercana al punto medio del intervalo  $[a_2, y_2]$ , esto se debe a que la varianza inicial es muy similar a la varianza de los errores.

En la gráfica (e) podemos ver a la función posterior de densidad de  $\theta_2|y_{1:2}$  y la función inicial de densidad de  $\theta_3|y_{1:2}$ . Llama la atención el hecho de que la varianza de  $\theta_3|y_{1:2}$  es mucho mayor que la de  $\theta_2|y_{1:2}$ , la razón por lo que esto sucede es porque  $\theta_3 = \theta_2 - v + w_3$  por lo que  $Var(\theta_3|y_{1:2}) = Var(\theta_2|y_{1:2}) + Var(w_3) = 0.175 + 0.9 = 1.075$ . En cuanto a la media, nos podemos dar cuenta que  $\mathbb{E}[\theta_3|y_{1:2}] = \mathbb{E}[\theta_2|y_{1:2}] - 4$  ya que la posición de la hormiga con respecto a su hogar al tiempo  $t = 3$  será 4 unidades adelante más un error aleatorio (de media cero) de la posición de la hormiga al tiempo  $t = 2$ .

La última gráfica, (f), muestra las funciones de densidad inicial y posterior de  $\theta_3|y_{1:2}$  y  $\theta_3|y_{1:3}$  respectivamente. Una vez más vemos que la varianza posterior es menor que la varianza inicial y que la media posterior se encuentra entre la media inicial y  $y_3 = 6.189$  y que debido a que la varianza de los errores es menor que la varianza inicial, la media posterior está más cercana a la nueva observación que a la media inicial.

## 2.2. Un ejemplo más

Recordemos ahora el ejemplo que se vió en el capítulo anterior acerca de calcular la posición de una partícula que vive en un mundo univariado (es decir,  $\theta \in S \subseteq \mathbb{R}$ ). Esta partícula se mueve, a diferencia del ejemplo que se vió en el capítulo pasado, por lo que la

posición de la partícula depende del tiempo. Esto es al tiempo  $t$ ,  $\theta_t$  representa la posición de la partícula, por simplicidad  $t = 0, 1, 2, \dots$ . Para darnos idea de cual es la posición de la partícula, a cada tiempo  $t$  se han tomado mediciones ( $Y_t : t = 1, 2, \dots$ ) que están afectadas por errores aleatorios independientes e idénticamente distribuidos, de tal forma que:

$$Y_t = \theta_t + \varepsilon_t, \quad \text{donde} \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2).$$

El propósito de este segundo ejemplo, además de asentar los conceptos, es darnos una idea de que tan acertada puede ser la estimación que nosotros hacemos acerca de  $\theta_t$  (dadas las observaciones hasta ese tiempo), y hacer una pequeña comparación entre la mejor predicción de  $\theta_t$  al tiempo  $t-1$  ( $a_t$ ) contra la mejor predicción de  $\theta_t$  al tiempo  $t$  ( $\mu_t$ ).

Supongamos esta vez que la partícula se comporta de la siguiente forma:

- Tiene una posición inicial  $\theta_0 \sim \mathcal{N}(\mu_0, C_0)$  donde  $\mu_0 = 1$  y  $C_0 = 2$
- La partícula se mantiene estática durante los siguientes 10 tiempos, es decir

$$\theta_t = \theta_{t-1} \quad \text{para } t = 1, 2, \dots, 10.$$

- Al tiempo  $t = 10$  la partícula se empieza a mover a una velocidad constante  $v_1 = 0.2$  y lo sigue haciendo hasta el tiempo  $t = 20$ , de tal forma que

$$\theta_t = \theta_{t-1} + v_1 \quad \text{para } t = 11, 12, \dots, 20.$$

- A partir del tiempo  $t = 20$  y hasta el tiempo  $t = 30$  la partícula continua moviéndose, pero esta vez lo hace a una velocidad aleatoria con una ligera tendencia de ser positiva.

$$\theta_t = \theta_{t-1} + v_2 + w_t \quad \text{para } t = 21, 22, \dots, 30,$$

donde  $v_2 = 0.4$  y las  $w_t$ 's son variables aleatorias i.i.d que se distribuyen  $\mathcal{N}(0, \sigma_w^2)$ , en este caso  $\sigma_w^2 = 0.9$ .

- Por último la partícula se estabiliza de nuevo a partir del tiempo  $t = 30$ .

$$\theta_t = \theta_{t-1} \quad \text{para } t = 31, 32, \dots$$

Supongamos que nosotros no tenemos forma de conocer la posición de la partícula, pero podemos tomar observaciones  $Y_t$  afectadas por un error aleatorio para darnos una idea de donde se encuentra la partícula a cada tiempo  $t = 1, 2, \dots$

$$Y_t = \theta_t + \varepsilon_t \quad \text{donde } \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2)$$

Se tiene el conocimiento de que  $\sigma_\varepsilon^2 = 0.4$ .

- Al tiempo  $t = 0$  sabemos que  $\theta_1 \sim \mathcal{N}(a_1, R_1)$  donde:

$$a_1 = \mathbb{E}[\theta_1] = \mathbb{E}[\theta_0] = \mu_0$$

y

$$R_1 = \text{Var}(\theta_1) = \text{Var}(\theta_0) = C_0.$$

Al tiempo  $t = 1$  se vuelve disponible la primera observación  $Y_1 = y_1$ . Por lo que nos interesa actualizar el conocimiento que se tiene acerca de  $\theta_1$ . Para esto  $\pi(\theta_1)$  tomara el papel de distribución inicial y  $\pi(y_1|\theta_1)$  es la verosimilitud, considerando que  $\theta_1 \sim \mathcal{N}(a_1, R_1)$  y  $Y_1|\theta_1 \sim \mathcal{N}(\theta_1, \sigma_\varepsilon^2)$  se tiene:

$$\theta_1|y_1 \sim \mathcal{N}(\mu_1, C_1)$$

donde

$$\mu_1 = \mathbb{E}[\theta_1|y_1] = a_1 + \frac{R_1}{R_1 + \sigma_\varepsilon^2}(y_1 - a_1)$$

y

$$C_1 = \text{Var}(\theta_1|y_1) = \left( \frac{1}{R_1} + \frac{1}{\sigma_\varepsilon^2} \right)^{-1}.$$

- Para  $t = 2, 3, \dots, 10$ , al tiempo  $t - 1$  podemos conocer la distribución de  $\theta_{t-1}|y_{1:t-1} \sim \mathcal{N}(\mu_{t-1}, C_{t-1})$ , como  $\theta_t = \theta_{t-1}$  entonces también sabemos que  $\theta_t|y_{1:t-1} \sim \mathcal{N}(a_t, R_t)$  donde  $a_t = \mu_{t-1}$  y  $R_t = C_{t-1}$ . Al tiempo  $t$  podemos tener acceso a una nueva observación  $Y_t = y_t$ . Por lo que podemos actualizar la información de  $\theta_t$  considerando la distribución inicial =  $\pi(\theta_t|y_{1:t-1})$  y la verosimilitud =  $\pi(y_t|\theta_t)$ . De esta forma se puede concluir que

$$\theta_t|y_{1:t} \sim \mathcal{N}(\mu_t, C_t)$$

con parámetros

$$\mu_t = \mathbb{E}[\theta_t|y_{1:t}] = a_t + \frac{R_t}{R_t + \sigma_\varepsilon^2}(y_t - a_t)$$

y

$$C_t = \text{Var}(\theta_t|y_{1:t}) = \left( \frac{1}{R_t} + \frac{1}{\sigma_\varepsilon^2} \right)^{-1}.$$

- Al tiempo  $t = 10$  la partícula se empieza a mover, por lo que habrá que hacer una pequeña modificación en la manera de estimar  $\theta_t$ . Para  $t = 11, 12, \dots, 20$  se tiene que  $\theta_t = \theta_{t-1} + v_1$ . Al tiempo  $t - 1$  conocemos que  $\theta_{t-1}|y_{1:t-1} \sim \mathcal{N}(\mu_{t-1}, C_{t-1})$  por lo que podemos fácilmente calcular que  $\theta_t|y_{1:t-1} \sim \mathcal{N}(a_t, R_t)$  donde  $a_t = \mu_{t-1} + v_1$  y  $R_t = C_{t-1}$ . En cuanto tengamos disponible la siguiente observación  $Y_t = y_t$  podemos calcular la distribución posterior  $\pi(\theta_t|Y_t)$  de la misma manera que lo hicimos cuando la partícula no se movía.
- Al tiempo  $t = 20$  el movimiento de la partícula cambia de nuevo, ahora se incluye un elemento aleatorio en el movimiento de esta, de forma que para  $t = 21, 22, \dots, 30$   $\theta_t = \theta_{t-1} + v_2 + w_t$ . Una vez más para el tiempo  $t - 1$  ya hemos calculado que  $\theta_{t-1}|y_{1:t-1} \sim \mathcal{N}(\mu_{t-1}, C_{t-1})$ . Nos interesa poder expresar la información que tenemos hasta ahorita de  $\theta_t$  por lo que calculamos  $\pi(\theta_t|y_{1:t-1})$ , sabemos que  $\theta_t|y_{1:t-1}$  sigue una distribución Normal ya que  $\theta_{t-1}|y_{1:t-1}$  y  $w_t$  (quien es independiente de  $Y_{1:t-1}$ ) se distribuyen Normal y son independientes entre ellas, por lo tanto lo único que nos hace falta para conocer como se distribuye  $\theta_t|y_{1:t-1}$  es calcular su media  $a_t$  y su Varianza  $R_t$ :

$$a_t = \mathbb{E}[\theta_t|y_{1:t-1}] = \mathbb{E}[\theta_{t-1}|y_{1:t-1} + v_2 + w_t] = \mu_1 + v_2$$

(ya que  $\mathbb{E}[w_t] = 0$  y  $v_2$  es una constante conocida)

y

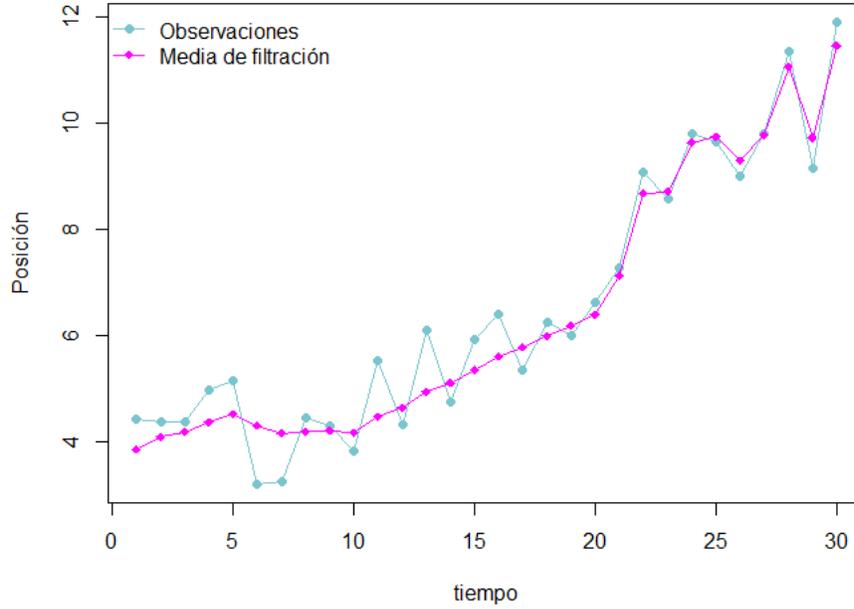
$$R_t = \text{Var}[\theta_t|y_{1:t-1}] = \text{Var}[\theta_{t-1}|y_{1:t-1} + v_2 + w_t] = C_1 + \sigma_w^2.$$

Una vez que ya hemos calculado  $\theta_t|y_{1:t-1}$  nos interesará, en cuanto obtengamos la observación  $Y_t = y_t$ , actualizar el conocimiento de  $\theta_t$  por lo que calculamos  $\pi(\theta_t|y_{1:t})$  de la misma forma que lo hemos venido haciendo, tomando como distribución inicial el conocimiento que teníamos de  $\theta_t$  ( $\pi(\theta_t|y_{t-1})$ ) y la función de verosimilitud que es la distribución de  $Y_t$  condicional dado  $\theta_t$ .

Para poder ver gráficamente lo anterior se ha hecho en R una pequeña simulación. Se ha simulado el movimiento de una partícula que se comporta como se describió al principio de este ejemplo. En base a este comportamiento se han tomado (simulado) observaciones afectadas por un error aleatorio y a cada tiempo (tomando en cuenta únicamente las observaciones hasta el tiempo  $t$ , como si las observaciones arribaran secuencialmente en el tiempo) se ha hecho una estimación de donde se encuentra la partícula, claramente esta será  $\mu_t = \mathbb{E}[\theta_t|y_{1:t}]$ , también conocida como *media de filtración* (esto porque estamos trabajando con distribuciones Normales, y como ya se vió en la primera parte de la tesis, en este caso el estimador más adecuado resulta ser la media posterior).

En la siguiente gráfica es posible ver las observaciones que hemos tomado así como la estimación de la posición de la partícula que vamos haciendo conforme avanza el tiempo y obtenemos nueva información.

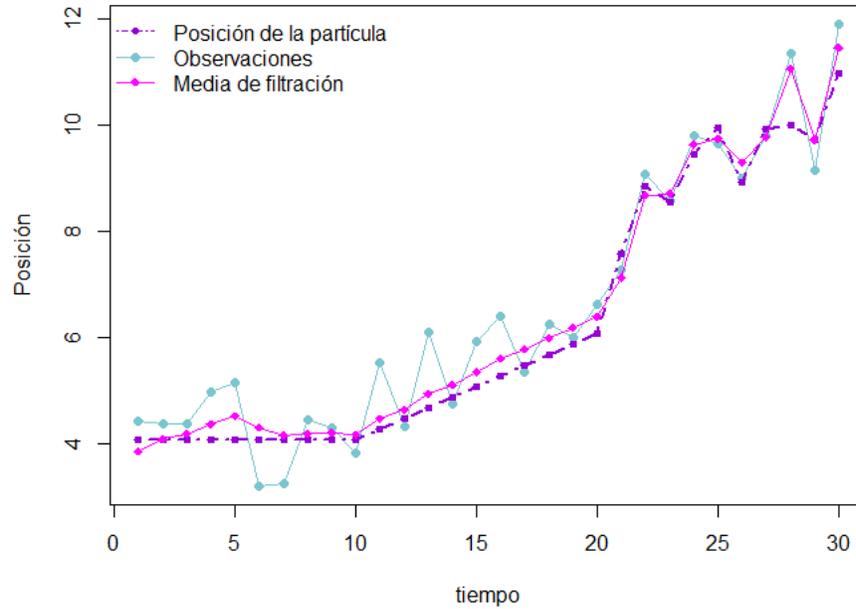
Figura 2.2.0.2: Estimación de la posición de la partícula



La estimación de la posición de la partícula a cada tiempo  $t$  se ha hecho tomando únicamente la información que se tiene al momento, es decir, la información que proporcionan las observaciones  $y_{1:t}$ .

Por lo general uno no tiene acceso a la posición de la partícula y es por esto que se realiza todo el procedimiento anterior, para poder hacer una predicción de en donde se encuentra la partícula a cada tiempo. Como uno de los propósitos de este ejemplo es darnos una idea de que tan acertada puede ser la estimación. En este caso, debido a que fue simulado el movimiento de la partícula, esta información. A continuación se presenta una gráfica que muestra la posición de la partícula (simulada), las observaciones tomadas y la estimación de la posición que vamos haciendo a cada tiempo.

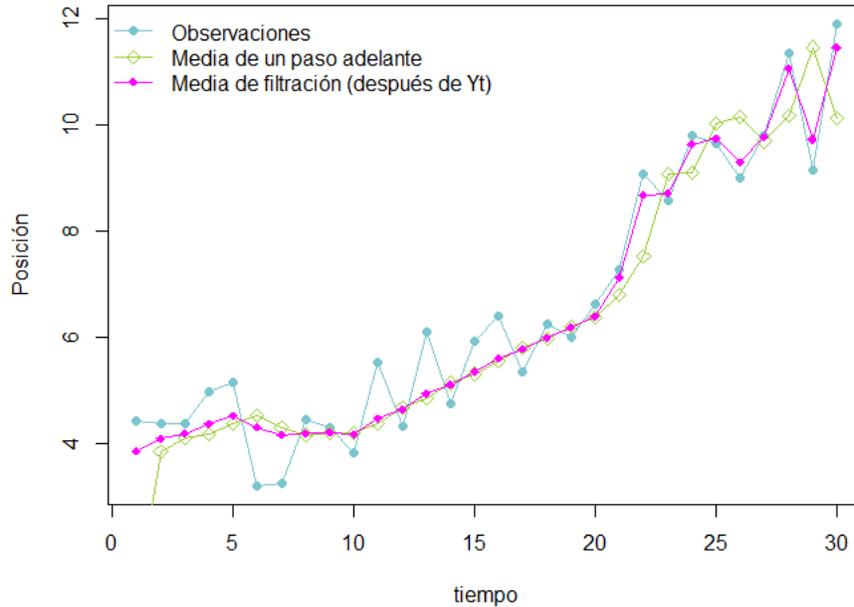
Figura 2.2.0.3: Comparación entre la posición real de la partícula y la estimación de ella.



La estimación que se ilustra en esta gráfica es la ilustrada en la figura 2.2.0.2.

La última gráfica que se presentará para este ejemplo es una gráfica donde se muestran las observaciones, la estimación de la posición a cada tiempo  $t$  tomando en cuenta la observación recabada en ese tiempo ( $\mu_t = \mathbb{E}[\theta_t | y_{1:t}]$ ), y la estimación de la posición al tiempo  $t$  antes de tomar la  $t$ -ésima observación ( $a_t = \mathbb{E}[\theta_t | y_{1:t-1}]$ ).

Figura 2.2.0.4: Comparación entre la estimación de la posición de la partícula (a cada tiempo) y la predicción de un paso adelante.



A cada tiempo  $t$  la estimación representada por la línea rosa se ha hecho tomando en cuenta las observaciones  $y_{1:t}$  mientras que para realizar la estimación ilustrada por la línea verde se han considerado las observaciones  $y_{1:t-1}$ , es decir la línea verde representa la predicción de un paso adelante de la posición de la partícula.

Como era de esperarse al cada tiempo  $t$ ,  $a_t$  está más alejada de  $y_t$  de lo que lo está  $\mu_t$ , es muy probable que también este más alejada  $a_t$  de la posición real de la partícula que  $\mu_t$  ya que está segunda estimación se hace teniendo más información.

Los ejemplos anteriores ilustran los aspectos básicos de un Modelo de Espacios de Estados (MEE), en particular un Modelo Dinámico Lineal (MDL), que se presentan en la siguiente sección.

### 2.3. Modelos de Espacios de Estados

Sea  $\{Y_t : t = 1, 2, \dots\}$  un proceso estocástico a tiempo discreto del cual se ha obtenido una realización  $\{Y_t = y_t : t = 1, 2, \dots, n\}$ . En la práctica es muy común que dada una realización finita, se desee especificar la distribución conjunta finito-dimensional de  $\{Y_1, Y_2, \dots, Y_n\}$  para alguna  $n \in \mathbb{N}$ , como generalmente la única información que se tiene acerca  $\{Y_t : t = 1, 2, \dots\}$  es precisamente alguna realización (o muestra) finita, especificar

la distribución conjunta finito-dimensional del proceso  $\{Y_t\}$  resulta un tanto complicado.

Una suposición común es asumir que las variables aleatorias del proceso estocástico son independientes o intercambiables. En muy pocas ocasiones estas suposiciones son realistas, ya que esencialmente estas hacen que el tiempo sea irrelevante. Una dependencia Markoviana entre las  $Y_t$ 's es la forma de dependencia más simple en la que el tiempo tiene un papel definitivo.

Recordemos, que se dice que  $\{Y_t\}$  es un proceso de Markov si:

$$\pi(y_t|y_{1:t-1}) = \pi(y_t|y_{t-1}) \quad \forall t > 1.$$

Esto significa que la información que proporcionan  $y_{1:t-1}$  acerca de  $Y_t$  es exactamente la misma información que proporciona  $y_{t-1}$ .

Es muy fácil demostrar que para un proceso de Markov, la distribución conjunta finito-dimensional se puede expresar de la siguiente forma:

$$\pi(y_{1:t}) = \pi(y_1) \prod_{i=2}^t \pi(y_i|y_{i-1}) \quad \forall t > 1.$$

**Demostración:**

$$\begin{aligned} \pi(y_{1:t}) &= \pi(y_t|y_{1:t-1})\pi(y_{1:t-1}) \\ &= \pi(y_t|y_{t-1})\pi(y_{1:t-1}). \end{aligned}$$

Si repetimos recursivamente  $t - 2$  veces estos dos pequeños pasos llegaremos a lo que queríamos demostrar.

Este tipo de dependencia entre las observaciones (las  $Y_t$ 's) no es apropiada en muchas aplicaciones. Los Modelos de Espacios de Estados se basan en este tipo de dependencia simple para construir modelos más complejos. Es decir, un Modelo Espacial de Estados supone un proceso  $\theta_t$  inobservable, y supone que las observaciones  $Y_t$ 's son una medida imprecisa de este proceso.

**Definición 2.3.0.1.** *Formalmente un Modelo de Espacios de Estados (MEE) consiste en dos procesos,  $\{\theta_t : t = 0, 1, \dots\}$  que toma valores en  $\mathbb{R}^p$  y  $\{Y_t : t = 1, 2, \dots\}$  que toma valores en  $\mathbb{R}^m$ , que satisfacen:*

- (i)  $\{\theta_t\}$  es un proceso de Markov.
- (ii) Dado  $\{\theta_t\}$ , las  $Y_t$ 's son independientes y  $Y_t$  depende únicamente de  $\theta_t$ .

El número (i) nos indica que a cada tiempo  $\theta_t$  no depende de toda la trayectoria anterior  $\theta_{0:t-1}$ , si no que solamente depende de su estado anterior  $\theta_{t-1}$ . Mientras que el número (ii) explica que dado  $\theta_t$ ,  $Y_t$  no depende de la trayectoria de  $\theta_{0:t-1}$  ni de su trayectoria  $Y_{1:t-1}$ , solamente depende del estado al tiempo  $t$  del proceso inobservable.

A  $\{\theta_t\}$  se le llama proceso de estados y a  $\{Y_t\}$  le llamamos proceso observable. Se puede pensar a  $\{\theta_t\}$  como un proceso auxiliar que nos ayuda a especificar la distribución del proceso observable  $\{Y_t\}$ .

Un MEE queda completamente especificado si conocemos la distribución inicial  $\pi(\theta_0)$  y las densidades condicionales  $\pi(\theta_t|\theta_{t-1})$  y  $\pi(y_t|\theta_t)$ . Ya que, para cualquier  $t > 0$ :

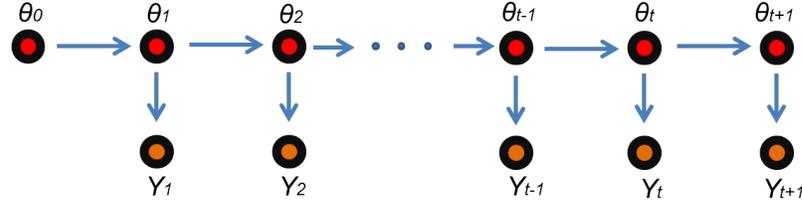
$$\begin{aligned}
\pi(\theta_{0:t}, y_{1:t}) &= \pi(y_t|\theta_{0:t}, y_{1:t-1})\pi(\theta_{0:t}, y_{1:t-1}) \\
&= \pi(y_t|\theta_t)\pi(\theta_{0:t}, y_{1:t-1}) \quad \dots \text{por (ii)} \\
&= \pi(y_t|\theta_t)\pi(y_{t-1}|\theta_{0:t}, y_{1:t-2})\pi(\theta_{0:t}, y_{1:t-2}) \\
&= \pi(y_t|\theta_t)\pi(y_{t-1}|\theta_{t-1})\pi(\theta_{0:t}, y_{1:t-2}) \\
&= \dots \\
&= \prod_{i=1}^t \pi(y_i|\theta_i)\pi(\theta_{0:t}) \\
&= \prod_{i=1}^t \pi(y_i|\theta_i) \prod_{i=1}^t \pi(\theta_i|\theta_{i-1})\pi(\theta_0) \quad \dots \text{por (i)} \\
&= \pi(\theta_0) \prod_{i=1}^t \pi(\theta_i|\theta_{i-1})\pi(y_i|\theta_i)
\end{aligned}$$

$$\therefore \pi(\theta_{0:t}, y_{1:t}) = \pi(\theta_0) \prod_{i=1}^t \pi(\theta_i|\theta_{i-1})\pi(y_i|\theta_i).$$

A los MEE's en los que los estados toman valores de un conjunto discreto con frecuencia se les llama *Modelos de Markov escondido*.

### 2.3.1. Una representación gráfica de los MEE

Figura 2.3.1.1: Representación gráfica de un Modelo Espacial de Estados



El flujo de información de un MEE se representa en la figura anterior. La gráfica en la figura 2.3.1.1 es una digráfica acíclica (ver sección de anexos para detalles sobre gráficas). En estos párrafos posteriores se tiene que pensar a los arcos de la gráfica como aristas no dirigidas, es decir, pensemos que los segmentos de recta que unen a los vértices (que en este caso representan variables aleatorias) no tienen dirección. La representación gráfica del modelo nos ayuda a deducir o representar la independencia condicional entre las variables aleatorias involucradas en este de la siguiente manera. Si  $A$  y  $B$  son conjuntos de variables aleatorias, entonces  $A$  es condicionalmente independiente de  $B$  dado un tercer conjunto  $C$ , si y sólo si  $C$  desconecta a  $A$  de  $B$ . Esto es, para cualquier  $a \in A$  y  $b \in B$  todas las  $(a, b)$ -trayectoria pasan por  $C$ , es decir todas las  $(a, b)$ -trayectorias tienen un vértice interno  $c \in C$ . Una vez dicho esto, podemos ver representado en la gráfica que  $Y_t$  es condicionalmente independiente de  $(\theta_{0:t-1}, Y_{1:t-1})$  dado  $\theta_t$  ya que cualquier trayectoria que va de  $Y_s$  a  $Y_t$  ( $s < t$ ) pasa por  $\theta_t$ , de igual manera, todas las trayectorias que van de  $\theta_s$  a  $Y_t$  pasa por  $\theta_t$ , es decir  $\{\theta_t\}$  separa a  $\{Y_t\}$  de  $\{\theta_{0:t-1}, Y_{1:t-1}\}$ .

$$\therefore \pi(y_t | \theta_{0:t}, y_{1:t-1}) = \pi(y_t | \theta_t).$$

Bajo el mismo argumento se puede ver que  $\theta_t$  es condicionalmente independiente de  $(\theta_{0:t-2}, y_{1:t-1})$  dado  $\theta_{t-1}$ .

$$\therefore \pi(\theta_t | \theta_{0:t-1}, y_{1:t-1}) = \pi(\theta_t | \theta_{t-1}).$$

Démonos cuenta que las dos ecuaciones anteriores son precisamente las que definen a un MEE.

### 2.3.2. Estimación de estados y predicción

Debido a su gran flexibilidad, los MEE tienen muchas aplicaciones en un rango enorme de problemas. Como en muchas aplicaciones estadísticas, un paso crucial y complicado en la mayoría de las ocasiones es la especificación del modelo.

Para empezar supongamos que ya conocemos el modelo, es decir, asumamos que las funciones de densidad  $\pi(\theta_0)$ ,  $\pi(y_t | \theta_t)$  y  $\pi(\theta_t | \theta_{t-1})$  ya están especificadas para toda  $t \in \mathbb{N}$ .

Para un MEE dado, una de las principales tareas es, basándonos en alguna parte de la secuencia de observaciones, hacer inferencia de los estados (inobservables) o predecir observaciones futuras. Estos dos problemas, llamados respectivamente *estimación* de estados y *predicción* de observaciones futuras, pueden resolverse calculando las densidades condicionales de las variables de interés dada la información disponible.

Para estimar el vector de estados es claro que hay que calcular  $\pi(\theta_s|y_{1:t})$  (con  $t, s \in \mathbb{N}$ ). Si ( $s = t$ ) entonces estamos hablando de una *filtración*, por otro lado si ( $s < t$ ) lo que se hace es *suavizar*.

La filtración se usa en problemas donde la información o los datos arriban secuencialmente en el tiempo. Por ejemplo: imaginemos un objeto que se mueve y a cada tiempo  $t$  (con  $t$  que pertenece a un conjunto numerable) podemos tomar una observación de su posición, o en alguna aplicación financiera donde día a día tenemos nueva información y es necesario actualizar la estimación diariamente. Para estos casos nos interesa encontrar un procedimiento que nos permita poder estimar el vector de estados  $\theta_t$  dada la información recabada hasta este momento, y poder actualizar esta estimación conforme nueva información se vuelva disponible, esto es calcular  $\pi(\theta_t|y_{1:t})$  y actualizarla calculando  $\pi(\theta_{t+1}|y_{1:t+1})$ .

Por otro lado, suavizamos cuando nos interesa hacer análisis retrospectivo, es decir, se tienen las observaciones  $y_{1:t}$  para cierto periodo y queremos estimar el vector de estados para cierto tiempo  $s < t$ . Un caso como este, por ejemplo, podría darse si quisieramos investigar el comportamiento de una o varias variables socio-económicas para cierto periodo en el pasado.

Muchas veces en el análisis de una serie de tiempo, el objetivo principal es poder predecir futuras observaciones, en este caso la estimación de los estados se vuelve un paso auxiliar. Si se desea conocer la predicción de un paso adelante, esto es dado  $y_{1:t}$  queremos predecir el valor que tomará  $Y_{t+1}$ , lo primero que haríamos es estimar el vector de estados  $\theta_{t+1}$  y basándonos en este valor, podemos calcular la predicción de  $Y_{t+1}$ . Para lograr esto nos apoyamos en la densidad de filtración  $\pi(\theta_t|y_{1:t})$  y en que se sabe como se distribuye  $\theta_{t+1}|\theta_t$ , para calcular la densidad predictiva de un paso adelante  $\pi(\theta_{t+1}|y_{1:t})$  del vector de estados. Una vez hecho esto, es posible calcular la densidad predictiva de un paso adelante de las observaciones  $\pi(y_{t+1}|y_{1:t})$  usando el hecho de que conocemos  $\pi(y_{t+1}|\theta_{t+1})$  y que ya calculamos la densidad predictiva del vector de estados.

De una manera similar es posible hacer predicción o estimaciones más a futuro. Esto es calcular  $\pi(y_{t+k}|y_{1:t})$  o calcular  $\pi(\theta_{t+k}|y_{1:t})$  para alguna  $k \in \mathbb{N}$ ,  $k > 1$ . Claramente, entre más grande sea  $k$  la predicción será más incierta, afortunadamente la densidad predictiva, a través de su varianza, ya sea del vector de estados o de las observaciones, nos dice que tanta incertidumbre existe en la predicción.

## Filtración

Una de las ventajas de los MEE es que debido a la estructura Markoviana y a la independencia condicional de las observaciones, es posible calcular las densidades de filtración y predictivas mediante un algoritmo recursivo. Esto lo podemos ver ilustrado en los ejemplos 2.1 y 2.2, claro que estos ejemplos son tan sólo un caso particular de un MEE. Por medio de la siguiente proposición se explica formalmente lo que se menciona en este párrafo.

**Proposición 2.3.2.1** (Algoritmo recursivo de filtración). *Para un MEE se tiene que:*

- (i) *La densidad predictiva del vector de estados de un paso adelante se puede calcular a partir de la densidad de filtración  $\pi(\theta_{t-1}|y_{1:t-1})$ .*

$$\pi(\theta_t|y_{1:t-1}) = \int \pi(\theta_t|\theta_{t-1})\pi(\theta_{t-1}|y_{1:t-1})d\theta_{t-1}.$$

- (ii) *La densidad predictiva de un paso adelante de las observaciones puede ser calculada a través de la densidad predictiva del vector de estados (calculada en el paso anterior).*

$$\pi(y_t|y_{1:t-1}) = \int \pi(y_t|\theta_t)\pi(\theta_t|y_{1:t-1})d\theta_t.$$

- (iii) *La densidad de filtración se puede calcular a partir de las densidades que calculamos en los incisos anteriores.*

$$\pi(\theta_t|y_{1:t}) = \frac{\pi(y_t|\theta_t)\pi(\theta_t|y_{1:t-1})}{\pi(y_t|y_{1:t})}.$$

### Demostración:

Inciso (i):

Para demostrar este inciso lo único que haremos es aprovechar el hecho de que  $\theta_t$  es independiente de  $Y_{1:t-1}$  dado  $\theta_{t-1}$ .

$$\begin{aligned}\pi(\theta_t|y_{1:t-1}) &= \int \pi(\theta_t, \theta_{t-1}|y_{1:t-1})d\theta_{t-1} \\ &= \int \pi(\theta_t|\theta_{t-1}, y_{1:t-1})\pi(\theta_{t-1}|y_{1:t-1})d\theta_{t-1} \\ &= \int \pi(\theta_t|\theta_{t-1})\pi(\theta_{t-1}|y_{1:t-1})d\theta_{t-1}.\end{aligned}$$

Inciso (ii):

Esta demostración se consigue gracias a que  $Y_t|\theta_t$  es independiente de  $Y_{1:t-1}$ .

$$\begin{aligned}\pi(y_t|y_{1:t-1}) &= \int \pi(y_t, \theta_t|y_{1:t-1})d\theta_t \\ &= \int \pi(y_t|\theta_t, y_{1:t-1})\pi(\theta_t|y_{1:t-1})d\theta_t \\ &= \int \pi(y_t|\theta_t)\pi(\theta_t|y_{1:t-1})d\theta_t.\end{aligned}$$

Inciso (iii):

El inciso (iii) se demuestra haciendo uso del Teorema de Bayes y de la independencia condicional entre  $Y_t$  y  $Y_{1:t-1}$  dado  $\theta_t$ .

$$\begin{aligned}\pi(\theta_t|y_{1:t}) &= \frac{\pi(y_{1:t}, \theta_t)}{\pi(y_{1:t})} \\ &= \frac{\pi(y_t|\theta_t, y_{1:t-1})\pi(\theta_t, y_{1:t-1})}{\pi(y_{1:t})} \\ &= \frac{\pi(y_t|\theta_t)\pi(\theta_t|y_{1:t-1})}{\pi(y_t|y_{1:t-1})}.\end{aligned}$$

Para que este algoritmo recursivo tenga sentido se tiene que empezar en algún punto, es decir, para alguna  $t$  se tiene que conocer o poder calcular la densidad de filtración para así poder calcular la densidad predictiva del vector de estados y la densidad predictiva de las observaciones al siguiente tiempo  $t + 1$ , y en base a estas calcular la densidad de filtración al tiempo  $t + 1$ . Por simplicidad, y para dar continuidad a la notación que hemos manejado, supongamos que la primera observación se tiene al tiempo  $t = 1$ . Recordemos que nos encontramos en el caso que el modelo está especificado, es decir, conocemos  $\pi(\theta_0)$ ,  $\pi(\theta_t|\theta_{t-1})$  y  $\pi(y_t|\theta_t)$  para toda  $t \geq 1$ . Dicho lo anterior, se tiene que:

- Para el tiempo  $t = 1$  la densidad predictiva del vector de estados es  $\pi(\theta_1) = \int \pi(\theta_1|\theta_0)\pi(\theta_0)d\theta_0$ , la densidad predictiva de las observaciones la escribimos como  $\pi(y_1) = \int \pi(y_1|\theta_1)\pi(\theta_1)d\theta_1$  y la densidad de filtración se escribe  $\pi(\theta_1|y_1) = \frac{\pi(y_1|\theta_1)\pi(\theta_1)}{\pi(y_1)}$ .
- Para el siguiente paso, cuando  $t = 2$ , la densidad predictiva se calcula de la siguiente manera  $\pi(\theta_2|y_1) = \int \pi(\theta_2|\theta_1)\pi(\theta_1)d\theta_1$  mientras que para calcular la densidad predictiva de las observaciones y la densidad de filtración basta sustituir  $t = 2$  en el inciso (ii) y (iii) de la proposición 2.3.2.1.

- Para calcular estas tres densidades en los tiempos restantes, es decir, cuando  $t > 2$  basta sustituir el valor de  $t$  en las ecuaciones de los incisos (i), (ii) y (iii) de la proposición anterior.

De una manera muy similar a como se calculó la distribución predictiva de un paso adelante, se puede calcular la distribución predictiva de  $k$  pasos adelante (con  $k > 0$ ) recursivamente de acuerdo a la fórmula:

$$\pi(\theta_{t+k}|y_{1:t}) = \int \pi(\theta_{t+k}|\theta_{t+k-1})\pi(\theta_{t+k-1}|y_{1:t})d\theta_{t+k-1}$$

y

$$\pi(y_{t+k}|y_{1:t}) = \int \pi(y_{t+k}|\theta_{t+k})\pi(\theta_{t+k}|y_{1:t})d\theta_{t+k}.$$

Las recursiones de filtración que acabamos de ver muestran que  $\pi(\theta_t|y_{1:t})$  resume la información recabada en las observaciones  $y_{1:t}$  de tal forma que es suficiente para predecir  $Y_{t+k}$ , para cualquier  $k > 0$ .

### Filtrando con observaciones faltantes

Al analizar una serie de tiempo es muy común que existan observaciones que por alguna razón falten, ya sea que para cierto tiempo  $t$  falte el vector completo  $y_t$  (observación completamente faltante) o que falte alguna entrada de este mismo (observación parcialmente faltante). Evidentemente en el caso univariado la observación está o no está, no existen observaciones parcialmente faltantes.

Afortunadamente la estructura de los MEE es tal que el problema de tener observaciones faltantes puede ser fácilmente acomodado en la recursión de filtración.

Primero analicemos el caso en el que para cierto tiempo  $t$  la observación  $y_t$  falta por completo, es claro que debido a que no la tenemos  $y_t$  no aporta información, de tal forma que

$$\pi(\theta_t|y_{1:t}) = \pi(\theta_t|y_{1:t-1}).$$

Eso significa que la densidad de filtración al tiempo  $t$  será la densidad predictiva (de un paso adelante) del vector de estados al tiempo  $t - 1$ . A partir del tiempo  $t + 1$  el algoritmo continua como lo conocemos hasta que encontremos que falta otra observación y resolvemos el problema de la misma manera que lo hicimos al tiempo  $t$ .

Ahora consideremos el caso en el que sólo faltan algunos componentes del vector de observaciones para cierto tiempo  $t$ , esto significa que  $y_t$  sí contiene (en los componentes que no faltan) cierta información acerca de  $\theta_t$ . Sea  $\hat{y}_t$  el vector compuesto solamente por los componentes no faltantes de  $y_t$ . Entonces basta con reemplazar, en la proposición 2.3.2.1,

$\pi(y_t|\theta_t)$  con  $\pi(\hat{y}_t|\theta_t)$  y  $\pi(y_t|y_{1:t-1})$  con  $\pi(\hat{y}_t|y_{1:t-1})$ .

### Suavizando

Como ya se había mencionado anteriormente, en el análisis de las series de tiempo con frecuencia se tiene acceso a observaciones de  $\{Y_t\}$  para cierto periodo de tiempo  $t = 1, 2, \dots, T$  y es de interés hacer una reconstrucción retrospectiva del sistema (estados) para analizar algún fenómeno físico o alguna variable socio-económica de la que se trate. Para este caso existe un algoritmo recursivo (hacia atrás) para calcular  $\theta_t$  dado  $Y_{1:T}$  para cualquier  $t < T$ . Empezamos por calcular la densidad de filtración  $\pi(\theta_T|y_{1:T})$  y vamos calculando las densidades hacia atrás de los estados dada la información disponible, de tal forma que a continuación calcularíamos  $\pi(\theta_{T-1}|y_{1:T})$  para después calcular  $\pi(\theta_{T-2}|y_{1:T})$  y seguimos hasta calcular la densidad o densidades que buscamos.

**Proposición 2.3.2.2** (Algoritmo recursivo de suavización). *Para un MEE general definido en la página 35 se tiene que:*

(i) *Dado  $y_{1:T}$  la secuencia  $(\theta_0, \theta_1, \dots, \theta_T)$  tiene probabilidades condicionales de transición (hacia atrás) dadas por:*

$$\pi(\theta_t|\theta_{t+1}, y_{1:T}) = \frac{\pi(\theta_{t+1}|\theta_t)\pi(\theta_t|y_{1:t})}{\pi(\theta_{t+1}|y_{1:t})} \quad \forall t = 0, 1, \dots, T-1.$$

(ii) *Las distribuciones de suavización de  $\theta_t$  dado  $y_{1:T}$  pueden ser calculadas de acuerdo a las siguientes recursiones hacia atrás sobre  $t$  empezando por  $\pi(\theta_T|y_{1:T})$ :*

$$\pi(\theta_t|y_{1:T}) = \pi(\theta_t|y_{1:t}) \int \frac{\pi(\theta_{t+1}|\theta_t)}{\pi(\theta_{t+1}|y_{1:t})} \pi(\theta_{t+1}|y_{1:T}) d\theta_{t+1}.$$

### Demostración:

Inciso (i):

Para probar este inciso se usa el hecho de que  $\theta_T$  es independiente de  $Y_{t+1:T}$  dado  $\theta_{t+1}$  así como  $\theta_{t+1}$  es independiente de  $Y_{1:t}$  dado  $\theta_t$ .

$$\begin{aligned}
\pi(\theta_t|\theta_{t+1}, y_{1:T}) &= \pi(\theta_t|\theta_{t+1}, y_{1:t}) \\
&= \frac{\pi(\theta_t, \theta_{t+1}, y_{1:t})}{\pi(\theta_{t+1}, y_{1:t})} \\
&= \frac{\pi(\theta_{t+1}|\theta_t, y_{1:t})\pi(\theta_t, y_{1:t})}{\pi(\theta_{t+1}, y_{1:t})} \\
&= \frac{\pi(\theta_{t+1}|\theta_t, y_{1:t})\pi(\theta_t|y_{1:t})}{\pi(\theta_{t+1}|y_{1:t})} \\
&= \frac{\pi(\theta_{t+1}|\theta_t)\pi(\theta_t|y_{1:t})}{\pi(\theta_{t+1}|y_{1:t})}.
\end{aligned}$$

Inciso (ii):

$$\begin{aligned}
\pi(\theta_t|y_{1:T}) &= \int \pi(\theta_t, \theta_{t+1}|y_{1:T})d\theta_{t+1} \\
&= \int \pi(\theta_t|\theta_{t+1}, y_{1:T})\pi(\theta_{t+1}|y_{1:T})d\theta_{t+1} \\
&= \int \pi(\theta_{t+1}|y_{1:T})\frac{\pi(\theta_{t+1}|\theta_t)\pi(\theta_t|y_{1:t})}{\pi(\theta_{t+1}|y_{1:t})}d\theta_{t+1} \quad \dots \quad (\text{Por inciso anterior}) \\
&= \pi(\theta_t|y_{1:t}) \int \pi(\theta_{t+1}|\theta_t)\frac{\pi(\theta_{t+1}|y_{1:T})}{\pi(\theta_{t+1}|y_{1:t})}d\theta_{t+1}.
\end{aligned}$$

□

En el inciso (ii) de la proposición 2.3.2.2 se ha podido expresar (para  $t < T$ )  $\pi(\theta_t|y_{1:T})$  en términos de densidades que hemos podido calcular y  $\pi(\theta_{t+1}|y_{1:T})$ . Esto significa que si podemos calcular  $\pi(\theta_s|y_{1:T})$  para alguna  $s \leq T$  entonces es posible calcular  $\pi(\theta_{s-1}|y_{1:T})$ , y por lo tanto  $\pi(\theta_t|y_{1:T})$  para toda  $t < T$ . Pero  $\pi(\theta_T|y_{1:T})$  es tan sólo la densidad de filtración al tiempo  $T$ , la cual ya sabemos como calcular.

## 2.4. Modelos Dinámicos Lineales

Ahora nos concentremos en los Modelos Dinámicos Lineales, estos son un caso particular de los MEE. Como veremos a continuación, debido a su estructura Gaussiana, para estos modelos es muy sencillo calcular las densidades de filtración, predicción y suavización.

**Definición 2.4.0.1.** *Formalmente un Modelo Dinámico Lineal (MDL) es un MEE que cumple que:*

$$\theta_0 \sim \mathcal{N}_p(\mu_0, C_0), \tag{2.4.0.1}$$

y para toda  $t > 1$  se tiene que:

$$\theta_t = G_t \theta_{t-1} + w_t, \quad w_t \sim \mathcal{N}_p(0, W_t), \quad (2.4.0.2)$$

$$Y_t = F_t \theta_t + v_t, \quad v_t \sim \mathcal{N}_m(0, V_t). \quad (2.4.0.3)$$

En lo anterior  $\theta_t$  ( $\forall t \geq 0$ ) y  $Y_t$  ( $\forall t \geq 1$ ) son vectores de dimensión  $p$  y  $m$  respectivamente, así como  $G_t$  y  $F_t$  son matrices conocidas de orden  $p \times p$  y  $m \times p$  respectivamente,  $\{w_t\}_{t>1}$  y  $\{v_t\}_{t>1}$  son dos secuencias independientes de vectores Gaussianos con media cero y matrices de varianzas conocidas  $W_t$  y  $V_t$ . Además,  $\theta_0$  es independiente de  $\{w_t\}$  y de  $\{v_t\}$ . La ecuación (2.4.0.2) se llama *ecuación de estados* o *ecuación del sistema*, mientras que a la ecuación (2.4.0.3) se le llama *ecuación de observación*.

Un MDL está completamente especificado si se conocen  $\mu_0$ ,  $C_0$ ,  $G_t$ ,  $F_t$ ,  $W_t$  y  $V_t$  para toda  $t$ .

Los ejemplos de la hormiga y la partícula comentados anteriormente son un ejemplo en particular de estos modelos.

Es muy fácil ver que un MDL es en efecto un caso particular de un MEE ya que  $\theta_t$  depende únicamente de  $\theta_{t-1}$  de forma que  $\theta_t | \theta_{t-1} \sim \mathcal{N}_p(G_t \theta_{t-1}, W_t)$  (ver página (149) en la sección de anexos para detalles acerca de la distribución Normal Multivariada) de forma que  $\{\theta_t\}$  es un proceso de Markov, además para cada  $t$ ,  $Y_t$  depende únicamente de  $\theta_t$  de la siguiente manera  $Y_t | \theta_t \sim \mathcal{N}_m(F_t \theta_t, V_t)$ .

Una característica muy importante de estos modelos que vale la pena resaltar es que tanto las matrices  $G_t$  y  $F_t$  como las varianzas de los errores dependen del tiempo, lo que hace que el modelo pueda variar en el tiempo. Al hacer análisis de series de tiempo esto resulta muy conveniente ya que existen sucesos que no tienen un comportamiento homogéneo en el tiempo y esta cualidad la podemos reflejar al ajustar un Modelo Dinámico Lineal adecuado.

### 2.4.1. Filtro de Kalman para un MDL univariado y un ejemplo

Para una mejor comprensión veamos primero el caso de un MDL univariado. El ejemplo en particular que analizaremos es un MDL univariado que se llama *Caminata aleatoria más un ruido* y está descrito por las siguientes ecuaciones:

$$\theta_t = \theta_{t-1} + w_t, \quad w_t \sim \mathcal{N}(0, W), \quad (2.4.1.1)$$

$$Y_t = \theta_t + v_t, \quad v_t \sim \mathcal{N}(0, V). \quad (2.4.1.2)$$

Es claro que en el caso de una *caminata aleatoria más un ruido*  $F_t = G_t = 1$  y  $W$ ,  $V$  son constantes conocidas así como  $\mu_0$  y  $C_0$ .

El motivo por el que este primer modelo recibe su nombre es porque  $\{\theta_t\}$  es una caminata aleatoria<sup>3</sup> y  $\{Y_t\}$  es una medición imprecisa de este proceso, por lo que se dice que  $\{Y_t\}$  es una caminata aleatoria más un error o ruido.

El ejemplo que discutiremos a continuación está basado en el modelo de la de una caminata aleatoria, con la diferencia de que en este caso permitiremos que la varianza de la ecuación del sistema y la de la ecuación de observaciones dependan del tiempo, esto es, sustituiremos  $W$  por  $W_t$  y  $V$  por  $V_t$ .

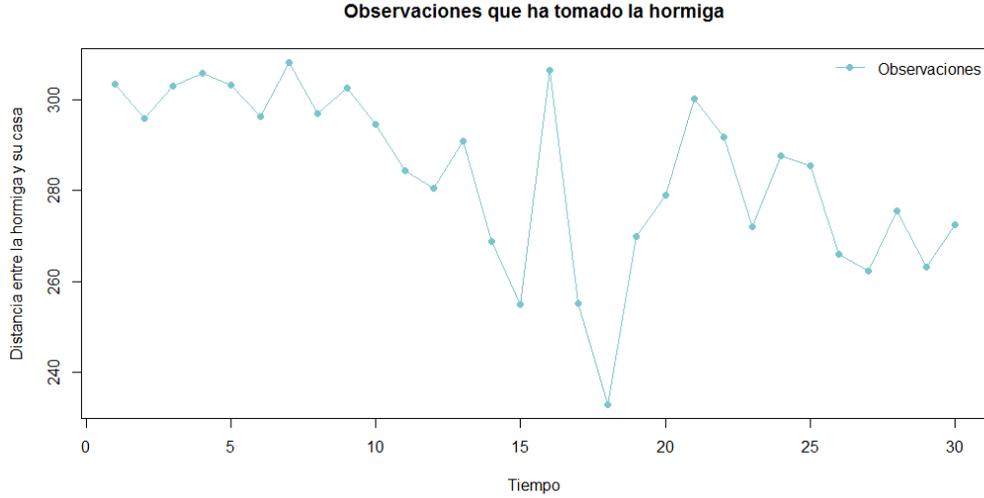
Volvamos a pensar por un momento en la hormiga que desea llevar un pedazo de hoja a su hogar, recordemos que la hormiga quiere saber aproximadamente a que distancia se encuentra de su hogar para administrar su energía. Esta vez, debido al fuerte viento que la hace desplazarse involuntariamente, la distancia entre la hormiga y su destino se comporta como una caminata aleatoria  $\theta_{t+1} = \theta_t + w_t$  donde  $w_t \sim \mathcal{N}(0, W_t)$ . Recordemos también que la hormiga no alcanza a ver su hogar todo el tiempo, solamente lo logra hacer en ciertos momentos. A cada tiempo  $t$  que la hormiga logra visualizar su casa, toma una medida empírica y imprecisa de la distancia que la separa de su casa  $Y_t = \theta_t + v_t$  donde  $v_t \sim \mathcal{N}(0, V_t)$ .

Supongamos que la hormiga ha logrado tomar 30 observaciones representadas en la siguiente gráfica donde para cada tiempo  $t$  la observación  $Y_t$  representa la distancia que la hormiga supone entre ella y su destino:

---

<sup>3</sup>Para aquellos que no esten familiarizados con el término, una caminata aleatoria es un proceso estocástico  $\{X_t\}$  que si al tiempo  $t - 1$ ,  $X_{t-1} = x$  con  $x \in \mathbb{R}$  entonces al siguiente tiempo  $X_t = x + \varepsilon_t$  donde  $\varepsilon_t$  es una variable aleatoria tal que  $\mathbb{P}[\varepsilon_t = y] = \mathbb{P}[\varepsilon_t = -y]$  para toda  $y \in \mathbb{R}$ . En otras palabras la caminata aleatoria describe el movimiento de un objeto que a cada tiempo camina un paso hacia adelante o hacia atrás de tamaño  $y$  con la misma probabilidad.

Figura 2.4.1.1: Observaciones tomadas por la hormiga



De nuevo, basándonos en estas observaciones nos interesa poder estimar la distancia real entre la hormiga y su destino al tiempo  $t = 30$  que es la distancia actual que separa a la hormiga de su casa.

Como se vió en los ejemplos 2.1 y 2.2 este problema puede ser resuelto mediante un algoritmo recursivo, mismo que se vió formalmente para el caso general de un MEE en la página 39. Para un MDL a este algoritmo recursivo de filtración se le llama *Filtro de Kalman* mismo que se revisará con cuidado más adelante, por ahora veremos este algoritmo recursivo para el caso particular de un MDL univariado (es decir,  $\theta_t \in \mathbb{R}$  así como  $Y_t \in \mathbb{R}$  para toda  $t$ ).

**Proposición 2.4.1.1** (Filtro de Kalman para Modelos Dinámicos Lineales univariados).  
*Para un MDL univariado definido en la página 43 existe  $t$  tal que*

$$\theta_{t-1}|y_{1:t-1} \sim \mathcal{N}(\mu_{t-1}, C_{t-1})$$

*y por lo tanto las siguientes afirmaciones se cumplen:*

(i) *La distribución predictiva de un paso adelante de  $\theta_t|y_{1:t-1} \sim \mathcal{N}(a_t, R_t)$  con parámetros:*

$$a_t = \mathbb{E}[\theta_t|y_{1:t-1}] = G_t \mu_{t-1},$$

$$R_t = \text{Var}(\theta_t|y_{1:t-1}) = G_t^2 C_{t-1} + W_t.$$

(ii) *La distribución predictiva de un paso adelante de  $Y_t|y_{1:t-1} \sim \mathcal{N}(f_t, Q_t)$  con parámetros:*

$$f_t = \mathbb{E}[Y_t|y_{1:t-1}] = F_t a_t,$$

$$Q_t = \text{Var}[Y_t|y_{1:t-1}] = F_t^2 R_t + V_t.$$

(iii) La distribución de filtración de  $\theta_t|y_{1:t} \sim \mathcal{N}(\mu_t, C_t)$  con parámetros:

$$\begin{aligned}\mu_t &= \mathbb{E}[\theta_t|y_{1:t}] = a_t + \frac{F_t R_t}{Q_t}(Y_t - F_t a_t), \\ C_t &= \text{Var}(\theta_t|y_{1:t}) = R_t - \frac{F_t^2 R_t^2}{Q_t}.\end{aligned}$$

**Demostración:**

Es muy fácil revisar que en efecto existe  $t$  tal que  $\theta_{t-1} \sim \mathcal{N}(\mu_{t-1}, C_{t-1})$ , basta fijarnos en la definición de un MDL en la ecuación 2.4.0.1 donde se especifica que  $\theta_0 \sim \mathcal{N}(\mu_0, C_0)$ .

Inciso (i):

Se sabe que

$$\theta_t|y_{1:t-1} = G_t \theta_{t-1}|y_{1:t-1} + w_t.$$

Como  $\theta_{t-1}|y_{1:t-1}$  y  $w_t$  son variables aleatorias independientes con distribución Normal y  $G_t \in \mathbb{R}$  entonces es claro que  $\theta_t|y_{1:t-1} \sim \mathcal{N}(a_t, R_t)$  con parámetros:

$$\begin{aligned}a_t &= \mathbb{E}[\theta_t|y_{1:t-1}] \\ &= \mathbb{E}[G_t \theta_{t-1} + w_t|y_{1:t-1}] \\ &= G_t \mathbb{E}[\theta_{t-1}|y_{1:t-1}] + \mathbb{E}[w_t] \\ &= G_t \mu_{t-1}\end{aligned}$$

y

$$\begin{aligned}R_t &= \text{Var}(\theta_t|y_{1:t-1}) \\ &= \text{Var}(G_t \theta_{t-1} + w_t|y_{1:t-1}) \\ &= G_t^2 \text{Var}(\theta_{t-1}|y_{1:t-1}) + \text{Var}(w_t) \\ &= G_t^2 C_{t-1} + W_t.\end{aligned}$$

Inciso (ii):

Por la independencia de  $\theta_t$  y  $v_t$  se sabe que

$$Y_t|y_{1:t-1} = F_t \theta_t|y_{1:t-1} + v_t$$

En la demostración del inciso anterior se vió que  $\theta_t|y_{1:t-1} \sim \mathcal{N}(a_t, R_t)$ , también se conoce que  $v_t$  tiene una distribución Normal y que es independiente de  $\theta_t|y_{1:t-1}$  esto significa que de igual manera  $Y_t|y_{1:t-1} \sim \mathcal{N}(f_t, Q_t)$  con parámetros:

$$\begin{aligned}
f_t &= \mathbb{E}[Y_t|y_{1:t-1}] \\
&= \mathbb{E}[F_t\theta_t + v_t|y_{1:t-1}] \\
&= F_t\mathbb{E}[\theta_t|y_{1:t-1}] + \mathbb{E}[v_t] \\
&= F_t a_t
\end{aligned}$$

y

$$\begin{aligned}
Q_t &= \text{Var}(Y_t|y_{1:t-1}) \\
&= \text{Var}(F_t\theta_t + v_t|y_{1:t-1}) \\
&= F_t^2 \text{Var}(\theta_t|y_{1:t-1}) + \text{Var}(v_t) \\
&= F_t^2 R_t + V_t.
\end{aligned}$$

Inciso (iii):

Para demostrar este último inciso basta fijarnos en que  $F_t\theta_t|y_{1:t-1} \sim \mathcal{N}(F_t a_t, F_t^2 R_t)$  y  $Y_t|F_t\theta_t \sim \mathcal{N}(F_t\theta_t, V_t)$ , haciendo uso de la proposición 1.4.2.1 se obtiene que  $F_t\theta_t|y_{1:t}$  tiene una distribución Normal con parámetros

$$\begin{aligned}
\mathbb{E}[F_t\theta_t|y_{1:t}] &= \frac{F_t^2 R_t}{F_t^2 R_t + V_t} Y_t + \frac{V_t}{V_t + F_t^2 R_t} F_t a_t \\
&= \frac{F_t^2 R_t}{F_t^2 R_t + V_t} Y_t + \left(1 - \frac{F_t^2 R_t}{F_t^2 R_t + V_t}\right) F_t a_t \\
&= \frac{F_t^2 R_t}{Q_t} Y_t + \left(1 - \frac{F_t^2 R_t}{Q_t}\right) F_t a_t \\
&= F_t a_t + \frac{F_t^2 R_t}{Q_t} (Y_t - F_t a_t)
\end{aligned}$$

y

$$\begin{aligned}
\text{Var}[F_t\theta_t|y_{1:t}] &= \frac{V_t F_t^2 R_t}{V_t + F_t^2 R_t} \\
&= \frac{V_t F_t^2 R_t}{Q_t} \\
&= \frac{(Q_t - F_t^2 R_t) F_t^2 R_t}{Q_t} \\
&= F_t^2 R_t - \frac{(F_t^2 R_t)^2}{Q_t}.
\end{aligned}$$

De todo esto se obtiene que

$$\theta_t|y_{1:t} \sim \mathcal{N}(\mu_t, C_t)$$

donde

$$\begin{aligned} \mu_t &= \mathbb{E}[\theta_t|y_{1:t}] \\ &= \frac{1}{F_t} \mathbb{E}[F_t \theta_t|y_{1:t}] \\ &= a_t + \frac{F_t R_t}{Q_t} (Y_t - F_t a_t) \end{aligned}$$

y

$$\begin{aligned} C_t &= \text{Var}(\theta_t|y_{1:t}) \\ &= \frac{1}{F_t^2} \text{Var}(F_t \theta_t|y_{1:t}) \\ &= R_t - \frac{F_t^2 R_t^2}{Q_t}. \end{aligned}$$

□

Volvamos al ejemplo de la hormiga donde las observaciones que va tomando se comportan como una caminata aleatoria más un ruido.

Vamos a proponer un primer modelo donde asignaremos a la media inicial  $\mu_0$  el valor de la primera observación  $Y_1$  y en cuanto a la varianza inicial supongamos  $C_0 = 2$ . Como sabemos que se trata de una caminata aleatoria es claro que  $F_t = G_t = 1$ . Lo único que queda por especificar son las varianzas de la ecuación de observación  $V_t$  y de la ecuación del sistema  $W_t$ , las cuales vamos a suponer que ambas son iguales a 2.

Por lo que el primer modelo que de la siguiente manera

$$\theta_0 \sim \mathcal{N}(Y_1, 2)$$

y

$$\begin{aligned} \theta_t|\theta_{t-1} &\sim \mathcal{N}(\theta_{t-1}, 2), \\ Y_t|\theta_t &\sim \mathcal{N}(\theta_t, 2). \end{aligned}$$

Supongamos que, a continuación, además de la información que ya teníamos, se nos informa que la hormiga calcula que su distancia inicial es de 300 unidades, pero no está tan segura por lo que proponemos una varianza inicial un poco más grande que la anterior  $C_0 = 5$ . De manera similar se nos informa que en condiciones normales la precisión de los estados es el doble que la precisión de las observaciones (es decir  $\frac{1}{W_t} = \frac{2}{V_t}$ ). Se sabe además que al momento de tomar las observaciones de la 11 a la 19 el viento sopla más fuerte,

lo cual se traduce en que para los tiempos  $t = 11 : 19$ ,  $W_t$  será mayor. De igual manera se nos informa que mientras tomaba la vigésima observación gotas de agua cayeron cerca de la hormiga salpicándola y afectando su visión, por lo que las observaciones  $Y_{20:30}$  fueron menos precisas que las demás.

Tomando en cuenta esta nueva información proponemos un segundo modelo. Asignamos los siguientes valores a este segundo modelo  $\mu_0 = 300$ ,  $C_0 = 5$ ,  $W_t = 1$  para  $t = 1 : 10$  y  $t = 20 : 30$ ,  $W_t = 5$  para  $t = 11 : 19$ ,  $V_t = 2$  para  $t = 1 : 19$  y  $V_t = 5$  para  $t = 20 : 30$ . Por lo que el modelo número 2 queda como sigue

$$\theta_0 \sim \mathcal{N}(300, 5)$$

y

$$\begin{aligned} \theta_t | \theta_{t-1} &\sim \mathcal{N}(\theta_{t-1}, W_t), & \text{donde } W_{1:10} = W_{20:30} = 1, W_{11:19} = 5, \\ Y_t | \theta_t &\sim \mathcal{N}(\theta_t, 2), & \text{donde } V_{1:19} = 2, V_{20:30} = 5. \end{aligned}$$

Se han programado en el software estadístico R, unas funciones para especificar un MDL, obtener una muestra de este y dada una serie de tiempo y un MDL (ambos univariados) calcular el filtro de Kalman. El código de estas funciones es el que sigue:

---

#### Código en R

---

```
#Función para especificar un MDL
mdl <- function(m0, C0, g, f, W, V)
{
  lg <- length(g)
  lf <- length(f)
  lW <- length(W)
  lV <- length(V)
  n <- min(c(lg,lf,lW,lV))
  return(list(n=n, m0=m0, C0=C0, g=g, f=f, W=W, V=V))
}

#Función para obtener una realización del MDL univariado
mdl.realic <- function(mdl)
{
  X0 <- rnorm(1, mean = mdl$m0, sd = sqrt(mdl$C0))

  X <- numeric(mdl$n)
  Y <- numeric(mdl$n)

  X[1] <- X0*mdl$g[1] + rnorm(1, mean = 0, sd = sqrt(mdl$W[1]))
  Y[1] <- X[1]*mdl$f[1] + rnorm(1, mean = 0, sd = sqrt(mdl$V[1]))
}
```

```

for(i in 2:mdl$n)
{
  X[i] <- X[i-1]*mdl$g[i] + rnorm(1, mean = 0, mdl$W[i]^2)
  Y[i] <- X[i]*mdl$f[i] + rnorm(1, mean = 0, mdl$V[i]^2)
}
return(list(X=X,Y=Y))
}

```

#Función que calcula el filtro de Kalman de un MDL univariado

```

mdl.filt <- function(Y, mdl)
{
  if(length(Y) > mdl$n)
  {
    print("El modelo no está completamente especificado")
  }
  else
  {
    n <- length(Y)

    a <- numeric(n)
    R <- numeric(n)
    m <- numeric(n)
    C <- numeric(n)
    f <- numeric(n)
    Q <- numeric(n)
    e <- numeric(n)
    e1 <- numeric(n)

    a[1] <- mdl$g[1]*mdl$m0
    R[1] <- mdl$g[1]*mdl$g[1]*mdl$C0 + mdl$W[1]

    f[1] <- mdl$f[1]*a[1]
    Q[1] <- mdl$f[1]*mdl$f[1]*R[1] + mdl$V[1]

    e[1] <- Y[1]-f[1]
    e1[1] <- e[1]/sqrt(Q[1])

    m[1] <- a[1] + (mdl$f[1]*R[1]/Q[1])*(e[1])
    C[1] <- R[1] - ((mdl$f[1]*R[1])^2/Q[1])

    for(i in 2:n)
    {
      a[i] <- mdl$g[i]*m[i-1]

```

```

R[i] <- mdl$g[i]*mdl$g[i]*C[i-1] + mdl$W[i]

f[i] <- mdl$f[i]*a[i]
Q[i] <- mdl$f[i]*mdl$f[i]*R[i] + mdl$V[i]

e[i] <- Y[i]-f[i]
e1[i] <- e[i]/sqrt(Q[i])

m[i] <- a[i] + ((mdl$f[i]*R[i])/Q[i])*(e[i])
C[i] <- R[i] - ((mdl$f[i]*R[i])^2/Q[i])
}

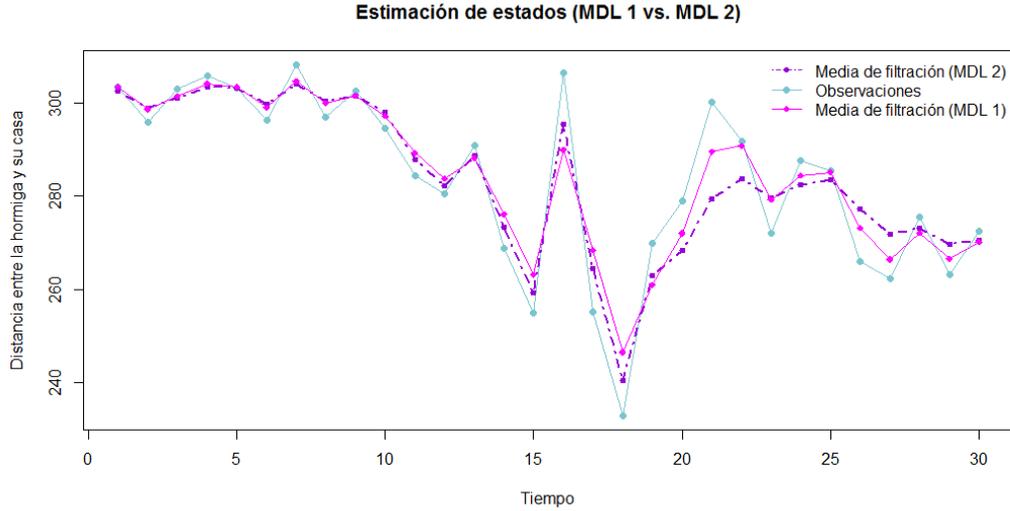
return(list(a=a,R=R,f=f,Q=Q,m=m,C=C,e=e,e1=e1,Y=Y))
}
}

```

---

Haciendo uso de estas funciones, para cada modelo de nuestro ejemplo, obtenemos las distribuciones predictivas de un paso adelante tanto para las observaciones como para los estados y la distribución de filtración. Con esta información podemos, para cada modelo, proponer una probable trayectoria de la distancia que separa a la hormiga de su casa desde el tiempo 1 hasta el 30. En la siguiente gráfica se presenta para cada tiempo la observación que obtuvo la hormiga así como la estimación de la distancia entre la hormiga y su casa dada la información recopilada hasta ese momento según cada modelo.

Figura 2.4.1.2: Comparación entre las estimaciones de estados del Modelo 1 y el Modelo 2



Dado que la distribución de filtración es una distribución Normal, la estimación de la distancia al tiempo  $t$  dada las observaciones tomadas hasta dicho tiempo es la media de filtración  $\mu_t = \mathbb{E}[\theta_t | y_{1:t}]$ , la varianza de la distribución de filtración  $C_t$  nos dice que tanta incertidumbre se tiene acerca de la predicción acabamos de elegir.

Antes de continuar, haremos un parentesis para hacer notar que la media de la distribución de filtración para un MDL univariado (calculada en la proposición anterior) se puede escribir como sigue

$$\begin{aligned}
 \mu_t &= a_t + \frac{F_t R_t}{Q_t} (Y_t - F_t a_t) \\
 &= a_t \left(1 - \frac{F_t^2 R_t}{Q_t}\right) + \frac{F_t R_t}{Q_t} Y_t \\
 &= \frac{V_t}{F_t^2 R_t + V_t} a_t + \frac{F_t R_t}{F_t^2 R_t + V_t} Y_t \\
 &= \frac{V_t}{F_t^2 (G_t^2 C_{t-1} + W_t) + V_t} a_t + \frac{F_t (G_t^2 C_{t-1} + W_t)}{F_t^2 (G_t^2 C_{t-1} + W_t) + V_t} Y_t.
 \end{aligned}$$

Considerando que para el caso particular de la caminata aleatoria más un ruido  $F_t = G_t = 1$  se tiene que

$$\begin{aligned}
 \mu_t &= \frac{V_t}{(C_{t-1} + W_t) + V_t} a_t + \frac{C_{t-1} + W_t}{(C_{t-1} + W_t) + V_t} Y_t \\
 &= \frac{V_t}{(C_{t-1} + W_t) + V_t} \mu_{t-1} + \frac{C_{t-1} + W_t}{(C_{t-1} + W_t) + V_t} Y_t.
 \end{aligned}$$

Esta expresión nos indica que la media de filtración no es más que la suma de  $\mu_{t-1} = \mathbb{E}[\theta_{t-1}|y_{1:t-1}] = \mathbb{E}[\theta_t|y_{1:t-1}] = a_t$  multiplicada por un factor  $j_t$  más la observación  $Y_t$  multiplicada por un factor  $k_t = 1 - j_t$ . Estos factores  $j_t$  y  $k_t$  nos indicarán que tanto depende de la media  $\mu_t$  de la estimación del estado al tiempo anterior  $\mu_{t-1}$  y que tanto depende de la observación  $Y_t$ . Notemos que estos factores son determinados por la varianza de filtración de un tiempo anterior  $C_{t-1}$ , por la varianza de la ecuación de sistema  $W_t$  y por la varianza de la ecuación de observación  $V_t$ . Si  $V_t$  es grande con respecto a  $W_t + C_{t-1}$  entonces  $\mu_t$  dependerá menos de la observación tomada al tiempo  $t$  y más de  $\mu_{t-1}$ , por el contrario, si  $V_t$  es pequeña con respecto a  $W_t + C_{t-1}$  entonces  $\mu_t$  va depender mucho de  $Y_t$  y poco de la estimación del estado al tiempo anterior.

Dada la explicación anterior, es claro que si se considera que el proceso observable  $\{Y_t\}$  es una medición más o menos precisa del proceso inobservable  $\{\theta_t\}$ , asignar un valor pequeño a  $V_t$  (con respecto al valor que asignaremos a  $W_t$ ) en nuestro modelo bastará para hacer que la estimación de  $\{\theta_t\}$  se vea fuertemente influenciada por las observaciones. Por el contrario, si se considera que las observaciones son muy imprecisas y se desea que la estimación del proceso inobservable no se vea tan influenciada por ellas entonces asignamos un valor grande a  $V_t$  con respecto al valor de  $W_t$ .

En la gráfica anterior los puntos morados representan la estimación según el segundo modelo, mientras que los puntos rosas representan la estimación según el primer modelo. En el intervalo de tiempo  $[1, 10]$  podemos ver que ambas estimaciones son bastantes similares, con la ligera diferencia de que las estimaciones dadas por el primer modelo están más influenciadas por las observaciones. A partir del tiempo  $t = 11$  y hasta el tiempo  $t = 19$  podemos ver que las estimaciones del modelo 2 se ven más afectadas por las observaciones que las estimaciones del primer modelo. En los tiempos restantes, las estimaciones del modelo 1 se ven considerablemente más influenciadas por las observaciones que las estimaciones hechas por el modelo 2.

La razón de este comportamiento, como se explicó en los párrafos previos, es precisamente la asignación que se hizo de las varianzas  $W_t$  y  $V_t$ .

Para responder la pregunta de la hormiga de ¿cuál es la distancia más probable que la separa de su destino al tiempo  $t = 30$ ?. Tenemos que fijarnos en las estimaciones de  $\theta_{30}$  del primero y del segundo modelo. Ambas estimaciones son muy parecidas, indican que la distancia de la hormiga al tiempo 30 es de alrededor de 270 unidades, por lo que la hormiga no tiene un gran conflicto en decidir a que modelo debe hacer caso. Sin embargo, hay casos donde no sucede esto y uno debe elegir que modelo seguir, para esto se hace un análisis de los errores de predicción  $e_t = Y_t - f_t = Y_t - \mathbb{E}[Y_t|y_{1:t-1}]$  de cada modelo para determinar cual es más adecuado, regresaremos a esto más adelante. Por ahora, debido a que el segundo modelo se construyó teniendo más información que el primero, pensaremos que este segundo modelo resulta más adecuado.

Antes de abandonar este ejemplo, me parece importante resaltar un par aspectos de los Modelos Dinámicos Lineales que se alcanzan a ver muy claramente en este caso. El primero es que los MDL pueden variar en el tiempo, como se vió en la construcción del segundo modelo las varianzas de las observaciones y de los estados no eran constantes en el tiempo, si no que para distintos intervalos de tiempo éstas varianzas fueron modificadas y el cálculo de las estimaciones resultó ser igual de sencillo. El segundo aspecto es que, además de lo previamente mencionado, al hacer la propuesta del modelo no. 2 quedó claro que es muy sencillo incorporar la información empírica que se tiene acerca del suceso mediante la elección de los valores de las varianzas, de la media inicial y de  $F_t$  y  $G_t$ .

### 2.4.2. Filtro de Kalman

Anteriormente se revisó el filtro de Kalman para un MDL donde  $\theta_t \in \mathbb{R}$  y  $Y_t \in \mathbb{R}$  para toda  $t$ . Ahora revisaremos este filtro para un MDL general definido en la página 43.

Antes de continuar, demostremos la siguiente proposición (generalización al caso multivariado de la proposición 1.4.2.1) que nos será muy útil al hacer el cálculo del filtro de Kalman.

**Proposición 2.4.2.1.** *Sea  $Y$  un vector aleatorio tal que  $Y|\theta \sim \mathcal{N}_m(F\theta, \Sigma)$  donde  $F$  es una matriz conocida de  $m \times p$  y  $\Sigma$  es la matriz de covarianza, si  $\theta \sim \mathcal{N}_p(a, R)$ , entonces*

$$\theta|Y \sim \mathcal{N}_p(\mu, C)$$

donde

$$\mu = C(F^T \Sigma^{-1} y + R^{-1} a)$$

y

$$C = (F^T \Sigma^{-1} F + R^{-1})^{-1}.$$

**Demostración:**

Sabemos que la distribución posterior es proporcional a la verosimilitud por la distribución inicial, por lo que escribimos:

$$\begin{aligned}
\pi(\theta|Y) &\propto \pi(y|\theta)\pi(\theta) \\
&\propto \exp\left\{-\frac{1}{2}(y - F\theta)^T \Sigma^{-1}(y - F\theta)\right\} \exp\left\{-\frac{1}{2}(\theta - a)^T R^{-1}(\theta - a)\right\} \\
&\propto \exp\left\{-\frac{1}{2}(\theta^T F^T \Sigma^{-1} F \theta - \theta^T F^T \Sigma^{-1} y - y^T \Sigma^{-1} F \theta)\right\} \times \\
&\quad \times \exp\left\{-\frac{1}{2}(\theta^T R^{-1} \theta - a^T R^{-1} \theta - \theta^T R^{-1} a)\right\} \\
&\propto \exp\left\{-\frac{1}{2}(\theta^T F^T \Sigma^{-1} F \theta - 2\theta^T F^T \Sigma^{-1} y)\right\} \exp\left\{-\frac{1}{2}(\theta^T R^{-1} \theta - 2\theta^T R^{-1} a)\right\} \dots \text{Nota}^4 \\
&\propto \exp\left\{-\frac{1}{2}[\theta^T (F^T \Sigma^{-1} F + R^{-1}) \theta - 2\theta^T (F^T \Sigma^{-1} y + R^{-1} a)]\right\}.
\end{aligned}$$

Sean

$$C = (F^T \Sigma^{-1} F + R^{-1})^{-1}, \quad \mu = C(F^T \Sigma^{-1} y + R^{-1} a),$$

entonces

$$\begin{aligned}
\pi(\theta|Y) &\propto \pi(y|\theta)\pi(\theta) \\
&\propto \exp\left\{-\frac{1}{2}(\theta^T C^{-1} \theta - 2\theta^T C^{-1} \mu)\right\} \\
&\propto \exp\left\{-\frac{1}{2}(\theta - \mu)^T C^{-1}(\theta - \mu)\right\},
\end{aligned}$$

donde reconocemos el Kernel de una distribución Gaussiana con vector de medias  $\mu$  y matriz de covarianza  $C$ . □

**Proposición 2.4.2.2** (Filtro de Kalman). *Considera un MDL dado por las ecuaciones 2.4.0.1, 2.4.0.2 y 2.4.0.3. Existe  $t$  tal que*

$$\theta_{t-1}|y_{1:t-1} \sim \mathcal{N}_p(\mu_{t-1}, C_{t-1})$$

por lo que las siguientes afirmaciones son ciertas:

---

<sup>4</sup>Tanto  $\Sigma^{-1}$  como  $R^{-1}$  son matrices simétricas, por lo que  $y^T \Sigma^{-1} F \theta = (\theta^T F^T \Sigma^{-1} y)^T$  y  $a^T R^{-1} \theta = (\theta^T R^{-1} a)^T$ . Fijandonos en las dimensiones de las matrices y vectores involucrados es evidente que  $\theta^T F^T \Sigma^{-1} y$  y  $\theta^T R^{-1} a$  son escalares por lo que  $y^T \Sigma^{-1} F \theta = (\theta^T F^T \Sigma^{-1} y)^T = \theta^T F^T \Sigma^{-1} y$  así como  $a^T R^{-1} \theta = (\theta^T R^{-1} a)^T = \theta^T R^{-1} a$ .

(i) La distribución predictiva de un paso adelante de  $\theta_t$  dado  $y_{1:t-1}$  es  $\mathcal{N}_p(a_t, R_t)$  con parámetros:

$$\begin{aligned} a_t &= \mathbb{E}[\theta_t | y_{1:t-1}] = G_t \mu_{t-1}, \\ R_t &= \text{Var}(\theta_t | y_{1:t-1}) = G_t C_{t-1} G_t^T + W_t. \end{aligned}$$

(ii) La distribución predictiva de un paso adelante de  $Y_t$  dado  $y_{1:t-1}$  es  $\mathcal{N}_m(f_t, Q_t)$  con parámetros:

$$\begin{aligned} f_t &= \mathbb{E}[Y_t | y_{1:t-1}] = F_t a_t, \\ Q_t &= \text{Var}(Y_t | y_{1:t-1}) = F_t R_t F_t^T + V_t. \end{aligned}$$

(iii) La distribución de filtración de  $\theta_t$  dado  $y_{1:t}$  es  $\mathcal{N}_p(m_t, C_t)$  con parámetros:

$$\begin{aligned} \mu_t &= \mathbb{E}[\theta_t | y_{1:t}] = a_t + R_t F_t^T Q_t^{-1} e_t, \\ C_t &= \text{Var}(\theta_t | y_{1:t}) = R_t - R_t F_t^T Q_t^{-1} F_t R_t. \end{aligned}$$

donde  $e_t = Y_t - f_t$  es el error de predicción.

### **Demostación:**

Es evidente que existe  $t$  tal que  $\theta_{t-1} | y_{1:t-1} \sim N_p(\mu_{t-1}, C_{t-1})$  ya que  $\theta_0 \sim N_p(\mu_0, C_0)$ .

Antes de demostrar cada uno de los incisos notemos que por propiedades de la Normal Multivariada y por la definición de un MDL se sabe que  $(\theta_0, \theta_1, \dots, \theta_t, Y_1, Y_2, \dots, Y_t)$  es un vector aleatorio con distribución Gaussiana para toda  $t \geq 1$ . Esto significa que todos sus subvectores son Gaussianos así como sus distribuciones condicionales.

Sea  $t \geq 1$  tal que  $\theta_{t-1} | y_{1:t-1} \sim N_p(\mu_{t-1}, C_{t-1})$ .

Inciso(i):

Se sabe que  $\theta_t | y_{1:t-1}$  es un vector Gaussiano, por lo que basta con calcular su media y su varianza para especificar su distribución. Sea

$$\begin{aligned} a_t &= \mathbb{E}[\theta_t | y_{1:t-1}] \\ &= \mathbb{E}[G_t \theta_{t-1} + w_t | y_{1:t-1}] \\ &= G_t \mathbb{E}[\theta_{t-1} | y_{1:t-1}] + \mathbb{E}[w_t] \\ &= G_t \mu_{t-1} \end{aligned}$$

y

$$\begin{aligned} R_t &= \text{Var}(\theta_t | y_{1:t-1}) \\ &= \text{Var}(G_t \theta_{t-1} + w_t | y_{1:t-1}) \\ &= G_t \text{Var}(\theta_{t-1} | y_{1:t-1}) G_t^T + \text{Var}(w_t) \\ &= G_t C_{t-1} G_t^T + W_t, \end{aligned}$$

entonces  $\theta_t|y_{1:t-1} \sim \mathcal{N}_p(a_t, R_t)$ .

Inciso(ii):

Sabemos que  $Y_t|y_{1:t-1}$  es un vector Gaussiano de  $m \times 1$ , calculemos su vector de medias y su matriz de covarianza para conocer su distribución. Sea

$$\begin{aligned} f_t &= \mathbb{E}[Y_t|y_{1:t-1}] \\ &= \mathbb{E}[F_t\theta_t + v_t|y_{1:t-1}] \\ &= F_t\mathbb{E}[\theta_t|y_{1:t-1}] + \mathbb{E}[v_t] \\ &= F_t a_t \end{aligned}$$

y

$$\begin{aligned} Q_t &= \text{Var}(Y_t|y_{1:t-1}) \\ &= \text{Var}(F_t\theta_t + v_t|y_{1:t-1}) \\ &= F_t\text{Var}(\theta_t|y_{1:t-1})F_t^T + \text{Var}(v_t) \\ &= F_t R_t F_t^T + V_t, \end{aligned}$$

entonces  $Y_t|y_{1:t-1} \sim \mathcal{N}_p(f_t, Q_t)$ .

Inciso(iii):

Para demostrar este inciso hacemos uso de la Proposición 2.4.2.1 tomando como distribución inicial  $\pi(\theta_t|y_{1:t-1})$  y como verosimilitud  $\pi(y_t|\theta_t)$ , donde además se conoce que  $\theta_t|y_{1:t-1} \sim \mathcal{N}_p(a_t, R_t)$  y  $Y_t|\theta_t \sim \mathcal{N}_m(F_t\theta_t, V_t)$ . De aquí obtenemos que  $\theta_t|y_{1:t} \sim \mathcal{N}_p(\mu_t, C_t)$  donde

$$\mu_t = (F_t^T V_t^{-1} F_t + R_t^{-1})^{-1} (F_t^T V_t^{-1} y_t + R_t^{-1} a_t)$$

y

$$C_t = (F_t^T V_t^{-1} F_t + R_t^{-1})^{-1}.$$

Ahora, notemos que  $(F_t^T V_t^{-1} F_t + R_t^{-1})^{-1} = R_t - R_t F_t^T (F_t^T R_t F_t + V_t)^{-1} F_t R_t$ .

La manera más sencilla de verificar esto es revisar que  $R_t - R_t F_t^T (F_t^T R_t F_t + V_t)^{-1} F_t R_t$  es la matriz inversa de  $F_t^T V_t^{-1} F_t + R_t^{-1}$  :

$$\begin{aligned} &(F_t^T V_t^{-1} F_t + R_t^{-1})(R_t - R_t F_t^T (F_t^T R_t F_t + V_t)^{-1} F_t R_t) \\ &= I + F_t^T V_t^{-1} F_t R_t - (F_t^T V_t^{-1} F_t + R_t^{-1})(R_t F_t^T)(F_t^T R_t F_t + V_t)^{-1} F_t R_t \\ &= I + F_t^T V_t^{-1} F_t R_t - (F_t^T V_t^{-1} F_t R_t F_t^T + F_t^T)(F_t^T R_t F_t + V_t)^{-1} F_t R_t \\ &= I + F_t^T V_t^{-1} F_t R_t - F_t^T V_t^{-1} (F_t R_t F_t^T + V_t)(F_t^T R_t F_t + V_t)^{-1} F_t R_t \\ &= I + F_t^T V_t^{-1} F_t R_t - F_t^T V_t^{-1} F_t R_t = I. \end{aligned}$$

Con esto concluimos que efectivamente  $(F_t^T V_t^{-1} F_t + R_t^{-1})^{-1} = R_t - R_t F_t^T (F_t^T R_t F_t + V_t)^{-1} F_t R_t$ , por lo que podemos reescribir la matriz de covarianza posterior como:

$$\begin{aligned}
C_t &= R_t - R_t F_t^T (F_t^T R_t F_t + V_t)^{-1} F_t R_t \\
&= R_t - R_t F_t^T Q_t^{-1} F_t R_t.
\end{aligned}$$

Con esta información reescribimos el vector de medias posterior de la siguiente manera:

$$\begin{aligned}
\mu_t &= (R_t - R_t F_t^T Q_t^{-1} F_t R_t) (F_t^T V_t^{-1} y_t + R_t^{-1} a_t) \\
&= a_t - R_t F_t^T Q_t^{-1} F_t a_t + R_t F_t^T V_t^{-1} y_t + R_t F_t^T Q_t^{-1} F_t R_t F_t^T V_t^{-1} y_t \\
&= a_t - R_t F_t^T Q_t^{-1} F_t a_t + R_t F_t^T (I + Q_t^{-1} F_t R_t F_t^T) V_t^{-1} y_t \\
&= a_t - R_t F_t^T Q_t^{-1} F_t a_t + R_t F_t^T Q_t^{-1} (Q + F_t R_t F_t^T) V_t^{-1} y_t \\
&= a_t - R_t F_t^T Q_t^{-1} F_t a_t + R_t F_t^T Q_t^{-1} V_t V_t^{-1} y_t \\
&= a_t - R_t F_t^T Q_t^{-1} (y_t - F_t a_t) \\
&= a_t - R_t F_t^T Q_t^{-1} e_t,
\end{aligned}$$

donde  $e_t = (y_t - \mathbb{E}[Y_t | y_{1:t-1}]) = (y_t - f_t) = (y_t - F_t a_t)$  es el error de predicción.  $\square$

Esta manera de expresar a al vector de medias de filtración  $\mu_t$  demuestra que esta es una corrección de la estimación de  $\theta_t$  al tiempo  $t - 1$ , incorporando la información que proporciona la nueva observación  $y_t$ , al sumarle de error de predicción  $e_t$  multiplicado por un factor  $K_t = R_t F_t^T Q_t^{-1}$ .

Es importante hacer notar que el filtro de Kalman sufre de inestabilidad numérica. Esto es, a pesar de que las matrices del modelo propuesto sean todas definidas positivas, aplicar el filtro de Kalman de manera recursiva, puede producir matrices de covarianza no simétricas e incluso definidas negativas al trabajar con valores muy pequeños sobre todo de  $V_t$ . Para contrarrestar este problema existe un filtro alternativo, numéricamente más estable, que se basa en la descomposición singular de matrices abreviada por *SVD*<sup>5</sup>.

Esta descomposición se refiere a lo siguiente. Sea  $A$  una matriz de covarianza simétrica y definida positiva, entonces podemos escribir a  $A$  de la siguiente forma

$$A = U D U^T,$$

donde  $U$  es una matriz ortogonal<sup>6</sup> y  $D$  se una matriz diagonal. A esta descomposición se le llama la SVD de  $A$ .

El filtro basado en la SVD lo que hace es ir actualizando, en lugar de las matrices de covarianza, la SVD de estas. En este documento no se revisarán los detalles de esta versión alternativa de filtración, sin embargo se hace la referencia a Oshman and Bar-Itzhack

<sup>5</sup>Singular Value decomposition

<sup>6</sup>Esto es,  $U^T = U^{-1}$ . Ver la sección de matrices en la parte de anexos para más información.

(1986) y Wang et al. (1992) para más información acerca del tema.

Recordemos que dado un MEE y dadas las observaciones  $y_{1:t}$ , podemos reconocer tres tipos de problemas referentes a la estimación de estados, todos resuleven calculando  $\pi(\theta_s|y_{1:t})$ .

- El primero es el calcular  $\pi(\theta_t|y_{1:t})$ , y se resuelve mediante el algoritmo recursivo de filtración.
- En el segundo problema se pretende calcular  $\pi(\theta_s|y_{1:t})$  para  $s < t$  y existe un algoritmo recursivo, de suavización, mediante el cual es posible solucionarlo.
- El tercer y último problema referente a la estimación de estados es el de predecir algún valor futuro del vector de estados a través de calcular  $\pi(\theta_s|y_{1:t})$  para  $s > t$ .

Estos tres problemas se han discutido para el caso general de un Modelo Espacial de Estados. En los párrafos anteriores se revisó el algoritmo de filtración para el caso particular de un MDL. A continuación estudiaremos el algoritmo de suavización y de predicción para un MDL.

### 2.4.3. Suavizador de Kalman

**Proposición 2.4.3.1** (Suavizador de Kalman). *Para un Modelo Dinámico Lineal dado por las ecuaciones 2.4.0.1, 2.4.0.2 y 2.4.0.3, si se sabe que para alguna  $t < r$   $\theta_{t+1}|y_{1:r} \sim \mathcal{N}_p(s_{t+1}, S_{t+1})$  entonces*

$$\theta_t|y_{1:r} \sim \mathcal{N}_p(s_t, S_t)$$

donde

$$s_t = \mu_t + C_t G_{t+1}^T R_{t+1}^{-1} (s_{t+1} - a_{t+1})$$

y

$$S_t = C_t - C_t G_{t+1}^T R_{t+1}^{-1} (R_{t+1} - S_{t+1}) R_{t+1}^{-1} G_{t+1} C_t.$$

#### Demostración:

Supongamos que para alguna  $1 \leq t < r$  se tiene que  $\theta_{t+1}|y_{1:r} \sim \mathcal{N}_p(s_{t+1}, S_{t+1})$ . Sabemos por propiedades de la Normal Multivariada (revisar Anexo A.3.1) que  $(\theta_0, \theta_1, \dots, \theta_r, Y_1, Y_2, \dots, Y_r)$  es un vector Gaussiano y que por lo tanto las distribuciones marginales y condicionales involucradas son gaussianas, en particular  $\theta_t|y_{1:r}$  tiene distribución Gaussiana, por lo que para conocer su distribución basta calcular su vector de medias y su matriz de covarianza.

Por propiedades de esperanza y varianza condicional se sabe que

$$\mathbb{E}[\theta_t|y_{1:r}] = \mathbb{E}[\mathbb{E}[\theta_t|y_{1:r}, \theta_{t+1}]|y_{1:r}]$$

y

$$Var(\theta_t|y_{1:r}) = Var(\mathbb{E}[\theta_t|y_{1:r}, \theta_{t+1}]|y_{1:r}) + \mathbb{E}[Var(\theta_t|y_{1:r}, \theta_{t+1})|y_{1:r}].$$

Ahora, notemos que  $\theta_t$  es independiente de  $y_{t+1:r}$  dado  $\theta_{t+1}$  por lo que  $\pi(\theta_t|y_{1:r}, \theta_{t+1}) = \pi(\theta_t|y_{1:t}, \theta_{t+1})$ . Esta densidad la podemos calcular utilizando el teorema de Bayes si tomamos  $\pi(\theta_t|y_{1:t})$  como distribución inicial y  $\pi(\theta_{t+1}|\theta_t)$  como verosimilitud. Esto porque se sabe que  $\theta_{t+1}|\theta_t \sim N_p(G_{t+1}\theta_t, W_{t+1})$  y  $\pi(\theta_t|y_{1:t})$  la podemos calcular a través del filtro del Kalman de donde se obtiene que  $\theta_t|y_{1:t} \sim \mathcal{N}_p(\mu_t, C_t)$ .

Haciendo uso de la proposición 2.4.2.1 resulta que

$$\begin{aligned}\mathbb{E}[\theta_t|y_{1:t}, \theta_{t+1}] &= (G_{t+1}^T W_{t+1}^{-1} G_{t+1} + C_t^{-1})(G_{t+1}^T W_{t+1}^{-1} \theta_{t+1} + C_t^{-1} \mu_t) \\ &= \mu_t + C_t G_{t+1}^T (G_{t+1} C_t G_{t+1}^T + W_{t+1})^{-1} (\theta_{t+1} - G_{t+1} \mu_t) \\ &= \mu_t + C_t G_{t+1}^T R_{t+1}^{-1} (\theta_{t+1} - a_{t+1})\end{aligned}$$

y

$$\begin{aligned}\text{Var}(\theta_t|y_{1:t}, \theta_{t+1}) &= (G_{t+1}^T W_{t+1}^{-1} G_{t+1} + C_t^{-1})^{-1} \\ &= C_t - C_t G_{t+1}^T (G_{t+1} C_t G_{t+1}^T + W_{t+1})^{-1} G_{t+1} C_t \\ &= C_t - C_t G_{t+1}^T R_{t+1}^{-1} G_{t+1} C_t.\end{aligned}$$

**Nota:** En la demostración del Filtro de Kalman se verifica que las últimas dos igualdades son ciertas tanto para el vector de medias posterior como para la matriz de covarianza posterior.

Ahora, sean

$$\begin{aligned}s_t &= \mathbb{E}[\theta_t|y_{1:r}] \\ &= \mathbb{E}[\mathbb{E}[\theta_t|y_{1:r}, \theta_{t+1}]|y_{1:r}] \\ &= \mathbb{E}[\mu_t + C_t G_{t+1}^T R_{t+1}^{-1} (\theta_{t+1} - a_{t+1})|y_{1:r}] \\ &= \mu_t + C_t G_{t+1}^T R_{t+1}^{-1} (s_{t+1} - a_{t+1})\end{aligned}$$

y

$$\begin{aligned}S_t &= \text{Var}[\theta_t|y_{1:r}] \\ &= \text{Var}(\mathbb{E}[\theta_t|y_{1:r}, \theta_{t+1}]|y_{1:r}) + \mathbb{E}[\text{Var}(\theta_t|y_{1:r}, \theta_{t+1})|y_{1:r}] \\ &= \text{Var}(\mu_t + C_t G_{t+1}^T R_{t+1}^{-1} (\theta_{t+1} - a_{t+1})|y_{1:r}) + \mathbb{E}[C_t - C_t G_{t+1}^T R_{t+1}^{-1} G_{t+1} C_t|y_{1:r}] \\ &= (C_t G_{t+1}^T R_{t+1}^{-1} S_{t+1} R_{t+1}^{-1} G_{t+1} C_t) + (C_t - C_t G_{t+1}^T R_{t+1}^{-1} G_{t+1} C_t) \\ &= C_t - C_t G_{t+1}^T R_{t+1}^{-1} R_{t+1} R_{t+1}^{-1} G_{t+1} C_t + C_t G_{t+1}^T R_{t+1}^{-1} S_{t+1} R_{t+1}^{-1} G_{t+1} C_t \\ &= C_t - C_t G_{t+1}^T R_{t+1}^{-1} (R_{t+1} - S_{t+1}) R_{t+1}^{-1} G_{t+1} C_t.\end{aligned}$$

Por lo tanto se concluye que

$$\theta_t|y_{1:r} \sim \mathcal{N}_p(s_t, S_t)$$

□

Notemos que para utilizar el suavizador de Kalman, debemos encontrar  $1 < t \leq r$  para la cual se conozca  $\theta_t|y_{1:r}$ . Al tener esta información, es evidente que podemos calcular de manera recursiva  $\theta_s|y_{1:r}$  para toda  $s < t$ . Lo ideal sería entonces poder calcular  $\theta_r|y_{1:r}$ , pero démonos cuenta que esta densidad ya la hemos podido calcular mediante el Filtro de Kalman, por lo que cada vez que se busque estimar algún estado  $\theta_t$  dadas las observaciones  $y_{1:r}$  con  $1 \leq t < r$  lo primero que se hace es calcular la densidad de filtración  $\pi(\theta_r|y_{1:r})$  para después aplicar consecutivamente el Suavizador de Kalman hasta obtener la densidad deseada.

Al igual que para el Filtro de Kalman, para este algoritmo hay una versión alternativa de suavizador (numéricamente más estable) que trabaja con la SVD de las matrices de covarianza involucradas.

#### 2.4.4. Algoritmo recursivo de predicción para un MDL

**Proposición 2.4.4.1.** *Supóngase que para un MDL dado por las ecuaciones 2.4.0.1, 2.4.0.2 y 2.4.0.3, se sabe que  $\theta_{t+(k-1)}|y_{1:t} \sim \mathcal{N}_p(a_t(k-1), R_t(k-1))$  para alguna  $1 \leq k$ . Entonces las siguientes afirmaciones son ciertas.*

(i)  $\theta_{t+k}|y_{1:t} \sim \mathcal{N}_p(a_t(k), R_t(k))$  donde

$$a_t(k) = G_{t+k}a_t(k-1)$$

y

$$R_t(k) = G_{t+k}R_t(k-1)G_{t+k}^T + W_{t+k}.$$

(ii)  $Y_{t+k}|y_{1:t} \sim \mathcal{N}_p(f_t(k), Q_t(k))$  donde

$$f_t(k) = F_{t+k}a_t(k)$$

y

$$Q_t(k+1) = F_{t+k}R_t(k)F_{t+k}^T + V_{t+k}.$$

#### Demostración:

Como ya se ha mencionado anteriormente, todas las distribuciones condicionales del vector  $(\theta_{0:R}, Y_{1:R})$  son Gaussianas para cualquier  $R \geq 1$ . Por lo que para conocer explícitamente  $\pi(\theta_{t+k}|y_{1:t})$  así como  $\pi(Y_{t+k}|y_{1:t})$  basta con calcular el vector de medias y la matriz de covarianza de ambos vectores aleatorios.

Inciso(i):

Sean

$$\begin{aligned} a_t(k) &= \mathbb{E}[\theta_{t+k}|y_{1:t}] \\ &= \mathbb{E}[\mathbb{E}[\theta_{t+k}|y_{1:t}, \theta_{t+(k-1)}]|y_{1:t}] \\ &= \mathbb{E}[G_{t+k}\theta_{t+(k-1)}|y_{1:t}] \\ &= G_{t+k}\mathbb{E}[\theta_{t+(k-1)}|y_{1:t}] \\ &= G_{t+k}a_t(k-1) \end{aligned}$$

y

$$\begin{aligned}
R_t(k) &= \text{Var}(\theta_{t+k}|y_{1:t}) \\
&= \text{Var}(\mathbb{E}[\theta_{t+k}|y_{1:t}, \theta_{t+(k-1)}]|y_{1:t}) + \mathbb{E}[\text{Var}(\theta_{t+k}|y_{1:t}, \theta_{t+(k-1)})|y_{1:t}] \\
&= \text{Var}(G_{t+k}\theta_{t+(k-1)}|y_{1:t}) + \mathbb{E}[W_{t+k}|y_{1:t}] \\
&= G_{t+k}R_t(k-1)G_{t+k}^T + W_{t+k}.
\end{aligned}$$

Por lo tanto  $\theta_{t+k}|y_{1:t} \sim \mathcal{N}_p(a_t(k), R_t(k))$ .

Inciso(ii):  
Sean

$$\begin{aligned}
f_t(k) &= \mathbb{E}[Y_{t+k}|y_{1:t}] \\
&= \mathbb{E}[\mathbb{E}[Y_{t+k}|y_{1:t}, \theta_{t+k}]|y_{1:t}] \\
&= \mathbb{E}[F_{t+k}\theta_{t+k}|y_{1:t}] \\
&= F_{t+k}\mathbb{E}[\theta_{t+k}|y_{1:t}] \\
&= F_{t+k}a_t(k)
\end{aligned}$$

y

$$\begin{aligned}
Q_t(k) &= \text{Var}(Y_{t+k}|y_{1:t}) \\
&= \text{Var}(\mathbb{E}[Y_{t+k}|y_{1:t}, \theta_{t+k}]|y_{1:t}) + \mathbb{E}[\text{Var}(Y_{t+k}|y_{1:t}, \theta_{t+k})|y_{1:t}] \\
&= \text{Var}(F_{t+k}\theta_{t+k}|y_{1:t}) + \mathbb{E}[V_{t+k}|y_{1:t}] \\
&= F_{t+k}R_t(k)F_{t+k}^T + V_{t+k}.
\end{aligned}$$

Con lo que se concluye que  $Y_{t+k}|y_{1:t} \sim \mathcal{N}_p(f_t(k), Q_t(k))$ .

□

Al igual que en el problema de suavización, si se quisiera conocer  $\pi(\theta_s|y_{1:t})$  ó  $\pi(Y_s|y_{1:t})$  para alguna  $s > t$ , lo primero que se hace es calcular la densidad de filtración  $\pi(\theta_t|y_{1:t})$  para después aplicar el algoritmo recursivo de predicción consecutivamente hasta obtener la densidad requerida.

### 2.4.5. Un último ejemplo: El Modelo de Crecimiento Lineal

Antes de continuar discutamos de manera breve un ejemplo de un MDL que no es univariado. El modelo que discutiremos se llama *Modelo de crecimiento lineal* y se rige por las siguientes ecuaciones:

$$Y_t = \beta_t + v_t, \quad v_t \sim \mathcal{N}(0, V_t), \quad (2.4.5.1)$$

$$\beta_t = \beta_{t-1} + \gamma_{t-1} + w_t, \quad w_{t,1} \sim \mathcal{N}(0, \sigma_{\beta_t}^2), \quad (2.4.5.2)$$

$$\gamma_t = \gamma_{t-1} + u_t, \quad w_{t,2} \sim \mathcal{N}(0, \sigma_{\gamma_t}^2). \quad (2.4.5.3)$$

donde  $v_t, w_{t,1}$  y  $w_{t,2}$  son independientes entre sí. Antes de continuar notemos que este es un MDL donde  $\{Y_t\}$  toma valores en  $\mathbb{R}$ ,  $\{\theta_t\}$  toma valores en  $\mathbb{R}^2$ ,

$$\theta_t = \begin{pmatrix} \beta_t \\ \gamma_t \end{pmatrix}, \quad w_t = \begin{pmatrix} w_{t,1} \\ w_{t,2} \end{pmatrix}, \quad W_t = \begin{pmatrix} \sigma_{\beta_t}^2 & 0 \\ 0 & \sigma_{\gamma_t}^2 \end{pmatrix}, \quad G = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad F = \begin{pmatrix} 1 & 0 \end{pmatrix}.$$

Revisar que esto es cierto es muy fácil, basta ver que

$$\begin{aligned} Y_t &= F_t \theta_t + v_t \\ &= \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} \beta_t \\ \gamma_t \end{pmatrix} + v_t \\ &= \beta_t + v_t, \end{aligned}$$

además

$$\begin{aligned} \theta_t &= G_t \theta_{t-1} + w_t \\ \iff \begin{pmatrix} \beta_t \\ \gamma_t \end{pmatrix} &= \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_{t-1} \\ \gamma_{t-1} \end{pmatrix} + \begin{pmatrix} w_{t,1} \\ w_{t,2} \end{pmatrix} \\ &= \begin{pmatrix} \beta_{t-1} + \gamma_{t-1} + w_{t,1} \\ \gamma_{t-1} + w_{t,2} \end{pmatrix}, \end{aligned}$$

donde  $F_t = F$  y  $G_t = G$  para toda  $t$ . La matriz de covarianza  $W_t$  de  $w_t$ , descrita anteriormente, muestra que  $Cov(w_{t,1}, w_{t,2}) = 0$  o lo que se traduce en que estas dos variables son independientes<sup>7</sup>.

Consideremos la siguiente situación, supongamos que se está estudiando el crecimiento de una población de peces Koi que habitan en un estanque. Para esto, año con año se desea estimar el número real de peces que habitan dicho estanque, este número no se puede conocer explícitamente, sin embargo, es posible tomar una medición no exacta de él. Llamemos  $\beta_t$  al número real de Kois en el estanque al año  $t$  y sea  $Y_t$  la medición imprecisa que se tomó de dicho número. Supongamos que en adición se tiene el conocimiento de que el número de peces al año  $t$  será igual al número de peces que había en el estanque al año pasado más dos factores:

1.  $w_{t,1}$  que representa a la diferencia entre el número de peces que nacieron menos el número de peces que murieron a lo largo del año, y
2.  $\gamma_{t-1}$  que se refiere a la diferencia entre el número de peces que inmigraron al estanque menos el número de peces que emigraron del estanque a lo largo del año pasado.

---

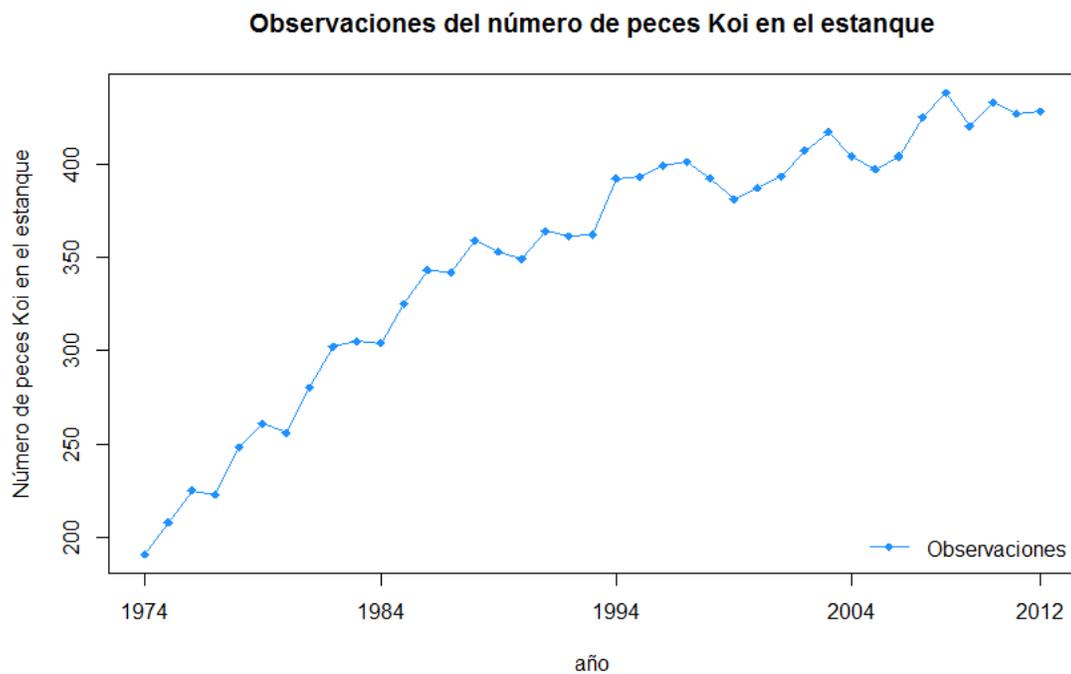
<sup>7</sup>Esto debido a que tienen una distribución Gaussiana

Sobre  $w_{t,1}$  se sabe que la probabilidad de que nazca un pez es igual a la probabilidad de que un pez perezca por lo que  $\mathbb{E}[w_{t,1}] = 0$ . En cuanto al otro factor,  $\gamma_t$ , se espera que el total de peces que migraron de esta población a lo largo del año  $t - 1$  sea igual al número de peces que migraron a lo largo del año anterior, es decir  $\mathbb{E}[\gamma_t | \gamma_{t-1}] = \gamma_{t-1}$  y que la probabilidad de que este número crezca en una unidad es la misma que la probabilidad de que este número decrezca en una unidad.

Este problema lo podríamos plantear como un modelo de crecimiento lineal, previamente presentado. Ahora, es claro que no puede haber una población de 40.75 peces, para solucionar este problema lo que se hará será redondear el número para cuestiones informativas, más no redondearemos este número para hacer el cálculo. Es decir, si para cierta  $t$  estimamos que  $\beta_{t-1} = 98.6$  y  $\gamma_t = 5.13$  calcularemos  $\beta_t = 98.6 + 5.13 + w_{t,1}$  sin embargo diremos que el número estimado de peces que había en la población al año  $t - 1$  fue de 98 y el número estimado de peces que migró a la población a lo largo de dicho año fue de 5.

Imaginemos que se nos dan las mediciones que se han tomado del número de peces en el estanque para los años 1974 al 2012. Por simplicidad llamaremos  $Y_1$  a la primera observación que tenemos,  $Y_2$  a la segunda observación que está disponible y así consecutivamente.

Figura 2.4.5.1: Observaciones tomadas del número de peces Koi en el estanque



Dadas las observaciones anteriores se nos pide que estimemos el número real de peces Koi en el estanque para los años 2012, 2013, 2014, 2015, 2016 y 2017. Además se nos pide que hagamos un análisis en retrospectiva y estimemos el número de peces en el estanque para los años 1973 al 2011.

Supongamos que en adición se nos informa que se piensa que en el año 1973 había alrededor de 200 peces en el estanque y que se piensa que a lo largo de ese año el número de peces aumentó en 6 debido al factor de migración. Con esta información proponemos el siguiente modelo:

$$\theta_0 \sim \mathcal{N}_2(\mu_0, C_0),$$

donde

$$\mu_0 = \begin{pmatrix} 200 \\ 6 \end{pmatrix} \quad y \quad C_0 = \begin{pmatrix} 3 & 0 \\ 0 & 2 \end{pmatrix},$$

$$\begin{aligned} Y_t &= F\theta_t + v_t & v_t &\sim \mathcal{N}(0, V), \\ \theta_t &= G\theta_{t-1} + w_t & w_t &\sim \mathcal{N}_2(0, W), \end{aligned}$$

donde

$$F = \begin{pmatrix} 1 & 0 \end{pmatrix}, \quad V = 30 \quad G = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad W = \begin{pmatrix} 5 & 0 \\ 0 & 3 \end{pmatrix}.$$

**Nota:** Fijemonos que el modelo propuesto es invariante en el tiempo ya que  $V_t = V$  y  $W_t = W$  para toda  $t$ .

Como ya lo hemos visto, lo primero que debemos hacer es calcular la densidad de filtración  $\theta_T|y_{1:T}$ , donde  $y_T$  es la última observación disponible, para después poder calcular las densidades de suavización y de predicción deseadas.

Existe una librería en  $R$  mediante la cual se puede hacer el análisis de un MDL dada una serie de tiempo. En el siguiente capítulo se aboradará con detalle esta librería, sin embargo para efectos de este ejemplo, simplemente la introduciremos y discutiremos algunas de sus funciones:

- Esta librería se carga con el comando `library(dlm)`, una vez cargada la librería podemos acceder y hacer uso de sus funciones.
- `dlm()`: Esta función recibe como parámetros el vector de medias y matriz de covarianza iniciales (`m0` y `C0`), las matrices constantes de las ecuaciones de observaciones y de estados `FF` y `GG`, y las matrices de covarianza `V` y `W` de  $v_t$  y  $w_t$  respectivamente. Lo que regresa esta función es un objeto de clase `dlm` que representa a un MDL especificado.

**Nota:** Dada esta manera de especificar los parámetros en la función, solamente se pueden introducir modelos invariantes en el tiempo. Para efectos de este ejemplo

esto no va a ser un problema, sin embargo para casos donde el modelo varia en el tiempo habrá que hacer modificaciones del retorno de la función (esto se revisará más adelante).

- **d1mFilter()**: Recibe como parámetros una serie de tiempo (posiblemente multivariada) así como un objeto de clase **d1m**. Esta función calcula la densidad de filtración así como las densidades de predicción de un paso adelante para los estados y para las observaciones (para cada tiempo que sea posible calcularlas<sup>8</sup>), y regresa un objeto de clase **d1mFiltered** que contiene:
  - (i) **mod**: El modelo.
  - (ii) **y**: Un vector o un conjunto ordenado de vectores de tamaño  $T$  que representa a la serie de tiempo  $(y_1, y_2, \dots, y_T)$ .
  - (iii) **m**: Una familia de tamaño  $T+1$  de los vectores de medias de filtración  $(\mu_0, \mu_1, \dots, \mu_T)$ .
  - (iv) **a**: Una familia de tamaño  $T$  de los vectores de medias de predicción de un paso adelante para el proceso de estados  $(a_1, a_2, \dots, a_T)$ .
  - (v) **f**: Una familia de tamaño  $T$  de los vectores de medias de predicción de un paso adelante para el proceso observable  $(f_1, f_2, \dots, f_T)$ .
  - (vi) **U.C**: Una familia de tamaño  $T + 1$  de matrices ortogonales.
  - (vii) **D.C**: Una familia de tamaño  $T + 1$  de vectores.

Tales que **U.C** conjuntamente con **D.C** se refieren a los conjuntos ordenados  $\{U_{C_t}\}_{0 \leq t \leq T}$  y  $\{D_{C_t}\}_{0 \leq t \leq T}$  respectivamente, que cumplen

$$U_{C_t} \text{diag}(D_{C_t}^2) U_{C_t}^T = C_t.$$

Es decir  $U_{C_t}$  y  $\text{diag}(D_{C_t}^2)$ <sup>9</sup> son las matrices que especifican la SVD de la matriz de covarianza de filtración  $(C_t)$ .

- (viii) **U.R**: Una familia de tamaño  $T$  de matrices ortogonales.
- (ix) **D.R**: Una familia de tamaño  $T$  de vectores.

De manera análoga a como sucede con **U.C** y **D.C**, **U.R** conjuntamente con **D.R** son las matrices y vectores que especifican la SVD de las matrices de covarianza de predicción de estados  $(R_1, R_2, \dots, R_T)$ .

- **d1mSmooth()**: El parámetro que recibe esta función es un objeto de clase **d1mFiltered** y basado en este objeto calcula las densidades de suavización de los estados. El retorno de esta función es una lista que contiene:

---

<sup>8</sup>Es decir, calcula las densidades de  $\theta_t|y_{1:t}$ ,  $\theta_t|y_{1:t-1}$  y  $Y_t|y_{1:t-1}$  para toda  $t \leq T$  donde  $y_T$  se refiere a la última observación disponible en  $y$

<sup>9</sup> $\text{diag}(D_{C_t}^2)$  se refiere a la matriz diagonal tal que los elementos de su diagonal son los componentes del vector  $D_{C_t}$  elevados al cuadrado

- (i) **s**: La familia de tamaño  $T+1$  de vectores de medias de suavización  $(s_0, s_1, \dots, s_T)$ , donde  $s_T = \mathbb{E}[\theta_T|y_{1:T}] = \mu_T$  es más bien el vector de medias de filtración al tiempo  $T$ .
- (ii) **U.S**: Una familia de matrices ortogonales de tamaño  $T + 1$
- (iii) **U.S**: Una familia de vectores de tamaño  $T + 1$ .

Donde **U.S**: y **D.S**: conjuntamente son las matrices y vectores que especifican la SVD de las matrices de covarianza de suavización  $(S_0, S_1, \dots, S_T)$ .

- **d1mForecast()**: Recibe como parámetros un objeto de clase **d1mFiltered**, un entero **nAhead** (se refiere al número de tiempos posteriores para los que se requiere hacer una predicción ya sea de estados o de observaciones), y un entero **sampleNew** (que se refiere al número de muestras que se requiere que generemos de los estados y/o de las observaciones para los próximos **nAhead** tiempos). Esta función regresa una lista de 6 objetos:
  - (i) **a**: La familia, con  $k = \mathbf{nAhead}$  elementos, de los vectores de medias de predicción de 1,2,...,k pasos adelante del vector de estados  $(a_T(1), a_T(2), \dots, a_T(k))$ .
  - (ii) **f**: La familia, con  $k = \mathbf{nAhead}$  elementos, de los vectores de medias de predicción de 1,2,...,k pasos adelante de las observaciones  $(f_T(1), f_T(2), \dots, f_T(k))$ .
  - (iii) **R**: La familia, con  $k = \mathbf{nAhead}$  elementos, de las matrices de covarianza de predicción de 1,2,...,k pasos adelante de los estados  $(R_T(1), R_T(2), \dots, R_T(k))$ .
  - (iv) **Q**: La familia, con  $k = \mathbf{nAhead}$  elementos, de las matrices de covarianza de predicción de 1,2,...,k pasos adelante de las estados observaciones  $(R_T(1), R_T(2), \dots, R_T(k))$ .
  - (v) **newStates**: Una familia de  $n = \mathbf{sampleNew}$  elementos donde cada elemento es una muestra del proceso de estados para los siguientes  $k = \mathbf{nAhead}$  tiempos.
  - (vi) **newObs**: Una familia de  $n = \mathbf{sampleNew}$  elementos donde cada elemento es una muestra del proceso observable para los siguientes  $k = \mathbf{nAhead}$  tiempos.
- **d1mSvd2var()**: Mediante esta función se pueden recuperar las matrices de covarianza ya sean de filtración, de suavización o de predicción. Esta función recibe como parámetros la familia de matrices **U.A** y la familia de vectores **D.A** y regresa la familia de matrices, llamemosla **A**, tal que cada elemento  $A_{[i]} = (U.A_{[i]})diag(D.A_{[i]})^2(U.A_{[i]})^T$ .

Regresemos, entonces, al problema de estimar el número de peces Koi en el estanque. Para esto supongamos que la serie de tiempo ya la tenemos en **R** y la hemos nombrado **tsKoi**

---

### Código en R

---

```
#Cargamos la librería
library(d1m)
```

```

#Especificamos el modelo
dlmkoi <- dlm(m0 = c(200,6), C0 = diag(c(3,2)), FF = matrix(c(1,0),nr=1),
V = 30, GG = matrix(c(1,0,1,1),nr=2), W = diag(c(5,3)) )

#Sea tsKoi el vector en R que contiene la serie de tiempo

#Calculamos las densidades de filtración (así como las de un paso adelante)
koiFilt <- dlmFilter(tsKoi,dlmkoi)

#Calculamos la densidad de suavización
koiSmooth <- dlmSmooth(koiFilt)

#Calculamos las densidades de predicción de k pasos adelante para k = 1,2,...,5
#y generamos 10 muestras del vector de estados y de las observaciones
#para los siguientes 5 tiempos
koiFore <- dlmForecast(koiFilt, nAhead = 5, sampleNew = 10)

#Construimos el vector formado por las estimaciones del número real
#de peces Koi en el estanque para los años 1973 al 2017
KoiState <- round(c(koiSmooth$s[,1],koiFore$a[,1]))

#Por último graficamos la estimación de los estados para cada año
#así como las observaciones
tsKoi.plot <- c(NA,tsKoi)
plot(KoiState, type = 'o', pch = 20, lty = 2, lwd = 1, col = "chocolate1",
xaxt = "n", xlab = "año", ylab = "Número de peces Koi en el estanque",
main = "Estimación del número de peces Koi en el estanque")
lines(tsKoi.plot, type = 'o', pch = 18, lty = 4, lwd = 1, col = "dodgerblue")
leg <- c("Estimación del número real de peces Koi", "Observaciones")
legend("bottomright", legend = leg,
col = c("chocolate1", "dodgerblue"),
lty = c(2, 4), pch = c(18, 20), bty = "n")
axis(1, at = c(1,11,21,31,41), labels = c(1973,1983,1993,2003,2013))

```

---

Antes de presentar resultados notemos que a cada tiempo  $0 \leq t \leq T$

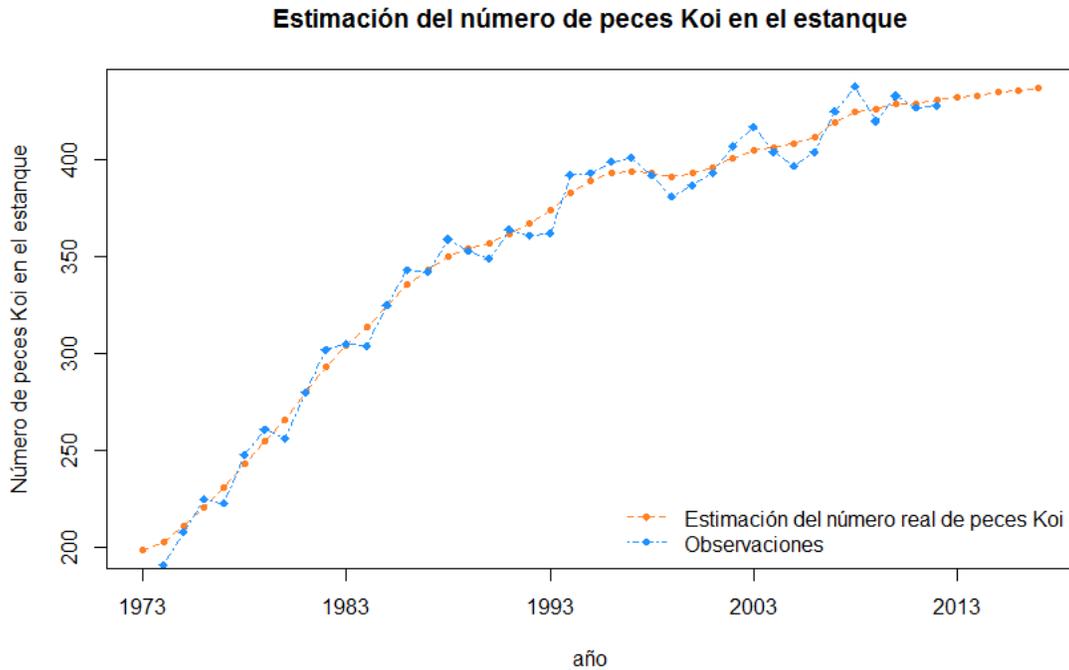
$$\mu_t = \mathbb{E}[\theta_t | y_{1:t}] = \begin{pmatrix} \mathbb{E}[\beta_t | y_{1:t}] \\ \mathbb{E}[\gamma_t | y_{1:t}] \end{pmatrix}$$

y análogamente sucede para  $s_t = \mathbb{E}[\theta_t | y_{1:T}]$ ,  $a_t = \mathbb{E}[\theta_t | y_{1:t-1}]$  y  $a_T(k) = \mathbb{E}[\theta_{T+k} | y_{1:T}]$ .

Por lo tanto, al estimar el número real de peces en el estanque ( $\beta_t$ ) a cada tiempo  $t$ , se toma la primera entrada (redondeada al entero más cercano) del vector de medias

correspondiente. Para los años 1973 al 2011 se toma la media de suavización<sup>10</sup> como estimador, para el año 2012 se toma la media de filtración y para los años 2013 a 2017 se toma la media de predicción del número de pasos adelante que corresponda. Con el código anteriormente presentado se ha generado la siguiente gráfica que resume los resultados obtenidos.

Figura 2.4.5.2: Estimación del número de peces Koi en el estanque (1973-2017)



Con este ejemplo damos por concluida la parte de estimación de estados y predicción de observaciones futuras para un MDL, a continuación se presenta una manera de verificar si el modelo propuesto para cierta serie de tiempo es adecuado.

#### 2.4.6. El proceso de inovaciones y la verificación del modelo

Hasta ahora, dadas algunas observaciones y dado un MDL, ya hemos revisado como hacer estimaciones del proceso inobservable (de estados) y como hacer predicciones del mismo así como de observaciones futuras. Es importante revisar si el modelo que se a propuesto es adecuado, ya que en base a este se obtienen todos los resultados, por lo que el proponer un MDL erroneo podría derivar en estimaciones y predicciones equivocadas. La verificación del modelo se hace mediante un análisis de los errores de predicción.

<sup>10</sup>La razón por la que se ha elegido tomar como estimador la media de suavización y no la media de filtración es porque la densidad de suavización se calcula tomando en cuenta una mayor cantidad de información que la media de filtración, por lo que es más factible que arroje un mejor estimador.

Recordemos que dado un MDL y las observaciones  $y_{1:T}$ , es posible para cada tiempo  $1 \leq t < T$ , hacer una predicción de cual será la observación actual dadas las observación que se han recopilado hasta el tiempo anterior esto es, calcular  $\pi(Y_t|y_{1:t-1})$  y basándonos en esta densidad elegir el valor de  $Y_t$  que es más factible obtener. Debido a que  $Y_t|y_{1:t-1}$  tiene una distribución Gaussiana, como ya se revisó anteriormente, la predicción que haremos será

$$f_t = \mathbb{E}[Y_t|y_{1:t-1}] = F_t a_t = F_t \mathbb{E}[\theta_t|y_{1:t-1}].$$

Se define entonces al *error de predicción* como

$$e_t = Y_t - f_t.$$

En ocasiones a los errores de predicción también se les llama *residuales*.

**Definición 2.4.6.1.** Dado un MDL se le llama *proceso de innovación* a la familia de variables aleatorias  $\{e_t\}_{t \geq 1}$  donde  $e_t = Y_t - f_t$  es el error de predicción al tiempo  $t$ .

**Proposición 2.4.6.1.** Sea  $\{e_t\}_{t \geq 1}$  el proceso de innovación de algún Modelo Dinámico Lineal. Entonces este proceso tiene las siguientes propiedades.

- (i)  $\mathbb{E}[e_t] = 0$ .
- (ii)  $\text{Var}(e_t) = Q_t$  donde  $Q_t = \text{Var}(Y_t|y_{1:t-1})$ .
- (iii) Para cualquier función  $g$  de  $Y_1, Y_2, \dots, Y_{t-1}$ , se tiene que  $\text{Cov}(e_t, g(Y_{1:t-1})) = 0$ .
- (iv) Para cualquier  $s < t$  se tiene que  $\text{Cov}(e_t, Y_s) = 0$ .
- (v) Para cualquier  $s < t$  se tiene que  $\text{Cov}(e_t, e_s) = 0$ .
- (vi)  $e_t$  es una función lineal de  $Y_1, \dots, Y_t$ .
- (vii)  $\{e_t\}_{t \geq 1}$  es un proceso Gaussiano.

para toda  $t \geq 1$ .

**Demostración:**

Inciso(i):

$$\begin{aligned} \mathbb{E}[e_t] &= \mathbb{E}[\mathbb{E}[e_t|y_{1:t-1}]] \\ &= \mathbb{E}[\mathbb{E}[Y_t - f_t|y_{1:t-1}]] \\ &= \mathbb{E}[\mathbb{E}[Y_t - \mathbb{E}[Y_t|y_{1:t-1}]|y_{1:t-1}]] \\ &= \mathbb{E}[\mathbb{E}[Y_t|y_{1:t-1}] - \mathbb{E}[Y_t|y_{1:t-1}]] = 0. \end{aligned}$$

**Nota:** 0 se refiere al vector de mismas dimensiones de  $Y_t$  tal que todos sus componentes son el escalar 0.

Inciso(ii):

$$\begin{aligned}
\text{Var}(e_t) &= \text{Var}(\mathbb{E}[e_t|y_{1:t-1}]) + \mathbb{E}[\text{Var}(e_t|y_{1:t-1})] \\
&= \text{Var}(\mathbb{E}[Y_t - f_t|y_{1:t-1}]) + \mathbb{E}[\text{Var}(Y_t - f_t|y_{1:t-1})] \\
&= \text{Var}(\mathbb{E}[Y_t|y_{1:t-1}] - f_t) + \mathbb{E}[\text{Var}(Y_t|y_{1:t-1})] \\
&= \mathbb{E}[Q_t] = Q_t.
\end{aligned}$$

Inciso(iii):

Sea  $Z = g(Y_1, \dots, Y_{t-1})$ , notemos que  $Z$  puede ser un vector o un escalar. Entonces

$$\begin{aligned}
\text{Cov}(e_t, Z) &= \mathbb{E}[(e_t - \mathbb{E}[e_t])(Z - \mathbb{E}[Z])^T] \\
&= \mathbb{E}[e_t Z^T - \mathbb{E}[e_t] Z^T - e_t \mathbb{E}[Z]^T + \mathbb{E}[e_t] \mathbb{E}[Z]^T] \\
&= \mathbb{E}[e_t Z^T] - \mathbb{E}[e_t] \mathbb{E}[Z]^T \\
&= \mathbb{E}[e_t Z^T] \\
&= \mathbb{E}[\mathbb{E}[e_t Z^T | y_{1:t-1}]] \\
&= \mathbb{E}[\mathbb{E}[e_t | y_{1:t-1}] Z^T] \\
&= 0.
\end{aligned}$$

Inciso(iv):

Considera la función  $g_{(s)}$  de  $Y_1, \dots, Y_{t-1}$  tal que  $g_{(s)}(Y_1, \dots, Y_{t-1}) = Y_s$  para toda  $s < t$ . Por el inciso anterior se tiene que

$$\text{Cov}(e_t, Y_s) = \text{Cov}(e_t, g_{(s)}(Y_1, \dots, Y_{t-1})) = 0.$$

Inciso(v):

Considera la función  $g_{(s)}$  de  $Y_1, \dots, Y_{t-1}$  tal que

$$g_{(s)}(Y_1, \dots, Y_{t-1}) = Y_s - \mathbb{E}[Y_s | y_{1:s-1}] = Y_s - f_s = e_s,$$

para toda  $s < t$ . Por el inciso (iii) se tiene que

$$\text{Cov}(e_t, e_s) = \text{Cov}(e_t, g_{(s)}(Y_1, \dots, Y_{t-1})) = 0.$$

Inciso(vi):

Como  $(Y_1, Y_2, \dots, Y_t)$  es un vector Gaussiano entonces  $f_t = \mathbb{E}[Y_t | y_{1:t-1}]$  es una función lineal de  $(Y_1, \dots, Y_{t-1})$ . Lo que significa que  $e_t = Y_t - f_t$  es una función lineal de  $Y_1, \dots, Y_t$ .

Inciso(vii):

El el inciso anterior se demostró que  $e_s$  es un función lineal de  $Y_{1:s}$  para toda  $1 \leq s$  lo que significa que podemos escribir

$$\begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_t \end{pmatrix} = A \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_t \end{pmatrix} + a$$

donde  $A$  es una matriz constante y  $a$  es un vector constante tales que la expresión anterior está bien definida.

Como se sabe que  $(Y_1, \dots, Y_t)$  es un vector Gaussiano, esto significa que la distribución de  $(e_1, \dots, e_t)$  y por lo tanto la de cualquiera de sus vectores es Gaussiana, además, como esto sucede para toda  $1 \leq t$  podemos concluir que  $\{e_t\}_{t \geq 1}$  es un proceso Gaussiano.  $\square$

Estas propiedades que acabamos de observar acerca del proceso de innovación pueden ser usadas para realizar una revisión de los supuestos que se hicieron al proponer el Modelo Dinámico Lineal.

Si las observaciones son univariadas, podemos definir a las innovaciones (ó residuales) estandarizadas como  $\tilde{e}_t = \frac{e_t}{\sqrt{Q_t}}$ . Por los incisos (i), (ii) y (v) Se sabe que  $\tilde{e}_1, \tilde{e}_2, \dots, \tilde{e}_t$  son variables aleatorias con distribución Normal Estándar independientes,<sup>11</sup> esto significa que al hacer el cálculo de las innovaciones estandarizadas, estas deberían verse como una muestra de tamaño  $t$  de una familia de variables aleatorias independientes e idénticamente distribuidas con distribución  $\mathcal{N}(0, 1)$ . En esta parte hay que tener cuidado, ya que si el tamaño de la muestra no es suficientemente grande, es posible que los residuales estandarizados no sugieran una distribución que corresponda al proceso mencionado a pesar de que el modelo sea correcto.

Por el otro lado, si las observaciones son multivariadas (digamos  $p$ -variadas), lo que se puede hacer es calcular  $\tilde{e}_t = B_t e_t$  donde  $B_t$  es tal que  $B_t Q_t B_t^T = I$ . Es claro que  $\mathbb{E}[\tilde{e}_t] = \mathbb{E}[B_t e_t] = 0$  y que  $Var(\tilde{e}_t) = Var(B_t e_t) = B_t Q_t B_t^T = I$ . Esto implica que los componentes del error estandarizado  $\tilde{e}_{t_1}, \tilde{e}_{t_2}, \dots, \tilde{e}_{t_p}$  son variables aleatorias Normales independientes, además por el inciso (v) de la proposición 2.4.6.1 se puede deducir que  $\tilde{e}_t$  es no correlacionado y por lo tanto independiente de  $\tilde{e}_s$  para toda  $t \neq s$ . Con esta información es posible realizar la revisión del modelo propuesto al igual que se hizo cuando las observaciones eran univariadas.

La librería en `d1m` en R incluye algunas funciones que nos permiten realizar dicho análisis.

- `residuals()`: Esta función extrae las innovaciones estandarizadas de un MDL dadas las observaciones  $y_1, y_2, \dots, y_T$ . Recibe como parámetro un objeto de clase `d1mFiltered` y retorna una lista con 2 componentes:
  - (i) `res`: Un vector o conjunto de vectores de tamaño  $T$  con los residuales estandarizados  $(\tilde{e}_1, \tilde{e}_2, \dots, \tilde{e}_T)$ .

---

<sup>11</sup>La independencia se debe a que en el caso de distribuciones Gaussianas la no correlación implica independencia

(ii) **sd**: Un vector o conjunto de vectores de tamaño  $T$  que incluye la desviación estándar de cada uno de los residuales, en el caso de que las observaciones sean univariadas regresa el vector  $(\sqrt{Q_1}, \sqrt{Q_2}, \dots, \sqrt{Q_T})$ .

- **tsdiag()**: Recibe como parámetro un objeto de clase `dlmFiltered`. El retorno de esta función son tres gráficas donde en la primera presentan los residuales estandarizados, en la segunda se presenta la gráfica de la Función de Auto-correlación (AFC), y en la tercer gráfica se muestran los p-values.

Adicionalmente, se puede utilizar la función `qqnorm()`, pasándole como parámetro las inovaciones estandarizaaos que se extrajeron mediante la función `residuals()`, para obtener una gráfica comparativa de los cuantiles de una distribución Normal Estándar contra los cuantiles de los residuales estandarizados.

Así mismo, puede resultar de interés, contrastar dos o más Modelos Dinámicos Lineales propuestos para una misma serie de tiempo. Existen varias medidas de precisión para comparar estos modelos y poder determinar cual se ajusta mejor a la serie de tiempo. Algunas de estas medidas de precisión y criterios de desempate son:

- **MAD**. Las siglas vienen de su nombre en inglés *Mean Absolute Deviation*. Este criterio calcula la media del valor absoluto de los errores de predicción  $e_t$ . De forma que si  $n$  es el número de observaciones, entonces

$$MAD = \frac{1}{n} \sum_{t=1}^n |e_t| = \frac{1}{n} \sum_{t=1}^n |y_t - f_t|. \quad (2.4.6.1)$$

- **MSE**. Estas siglas son la abreviación de su nombre en inglés *Mean Square Error*. Este criterio calcula la media del cuadrado de los errores de predicción  $e_t$ . Es decir, sea  $n$  el número de observaciones disponibles, se tiene entonces que

$$MSE = \frac{1}{n} \sum_{t=1}^n e_t^2 = \frac{1}{n} \sum_{t=1}^n (y_t - f_t)^2. \quad (2.4.6.2)$$

- **MAPE**. El tercer y último criterio que revisaremos recibe su nombreal abreviar *Mean Absolute Percentage Error*. Este criterio calcula la media del valor absoluto de los errores  $e_t$  divididos entre la observación correspondiente  $y_t$ . Esto es

$$MAPE = \frac{1}{n} \sum_{t=1}^n \frac{|e_t|}{y_t} = \frac{1}{n} \sum_{t=1}^n \frac{|y_t - f_t|}{y_t}. \quad (2.4.6.3)$$

donde nuevamente  $n$  es el número de observaciones disponibles.

Evidentemente, entre mejor se ajuste un modelo a una serie de tiempo los errores de predicción serán menores, por lo que entre más pequeño se el resultado de los criterio MAP, MSE y MAPE más adecuado será el modelo.

En los siguiente capítulos se ejemplifica la verificación del modelo asi como la utilización estas medidas de precisión como criterio de desempate entre dos o más modelos.

### 3. Especificación del modelo

En el capítulo anterior se discutió como realizar estimaciones y predicciones tanto del proceso inobservable de estados como de las observaciones dado un MDL y una serie de observaciones. En la práctica especificar este modelo puede ser una tarea difícil.

En esta parte del documento discutiremos el problema de la especificación del modelo para aquellos casos en los que las observaciones son univariadas.

#### 3.1. Modelos dinámicos lineales con observaciones univariadas

Los Modelos Dinámicos Lineales tienen una estructura aditiva, la cual discutiremos a continuación, que permite pensar a la serie observada como la suma de distintos componentes. De esta manera podemos abordar el problema de asignarle un MDL a la serie observada en partes, asignándole un MDL a cada componente y en base a esto construir el MDL de la serie observada.

Veamos a lo que nos referíamos en el párrafo anterior. Consideremos una serie de tiempo  $\{Y_t\}$ . Podemos asumir que esta serie se puede expresar de la siguiente forma.

$$Y_t = Y_{1t} + Y_{2t} + \dots + Y_{mt}, \quad (3.1.0.4)$$

donde  $Y_{it}$  es independiente de  $Y_{jt}$  para toda  $i \neq j$ .

De esta manera,  $Y_{1t}$  podría representar un componente de tendencia,  $Y_{2t}$  un componente de ciclicidad,  $Y_{3t}$  podría simbolizar a un componente de regresión y así consecutivamente. Consideremos que cada componente  $Y_{it}$  está descrito un MDL de la siguiente forma:

$$\begin{aligned} Y_{it} &= F_{it}\theta_{it} + v_{it}, & v_{it} &\sim \mathcal{N}(0, V_{it}), \\ \theta_{it} &= G_{it}\theta_{i,t-1} + w_{it}, & w_{it} &\sim \mathcal{N}_{p_i}(0, W_{it}). \end{aligned}$$

donde  $(Y_{it}, \theta_{it})$  es independiente de  $(Y_{jt}, \theta_{jt})$  para toda  $i \neq j$ , de tal forma que la serie de tiempo  $Y_t = \sum_{i=1}^m Y_{it}$  está descrita por el siguiente modelo.

$$\begin{aligned} Y_t &= F_t\theta_t + v_t, & v_t &\sim \mathcal{N}(0, V_t), \\ \theta_t &= G_t\theta_{t-1} + w_t, & w_t &\sim \mathcal{N}_p(0, W_t). \end{aligned}$$

donde

$$\theta_t = \begin{pmatrix} \theta_{1t} \\ \theta_{2t} \\ \vdots \\ \theta_{mt} \end{pmatrix}, \quad p = \sum_{i=1}^m p_i, \quad F_t = (F_{1t}, F_{2t}, \dots, F_{mt}), \quad V_t = \sum_{i=1}^m V_{it},$$

$$G_t = \begin{pmatrix} G_{1_t} & 0 & \cdots & 0 \\ 0 & G_{2_t} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & G_{m_t} \end{pmatrix} \quad y \quad W_t = \begin{pmatrix} W_{1_t} & 0 & \cdots & 0 \\ 0 & W_{2_t} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & W_{m_t} \end{pmatrix}.$$

Esto es muy fácil de corroborar, se debe a la independencia entre  $(Y_{i_t}, \theta_{i_t})$  y  $(Y_{j_t}, \theta_{j_t})$  para toda  $i \neq j$ .

### 3.1.1. Modelos de Tendencia

Para describir la tendencia de una serie de tiempo  $\{Y_t\}$ , los modelos más comunmente usados son los Modelos Polinómicos, los cuales reciben su nombre del hecho de que modelo de este tipo se cumple que para  $k \geq 1$

$$f_t(k) = \mathbb{E}[Y_{t+k}|y_{1:t}] = c_{0_t} + c_{1_t}k + \dots + c_{(n-1)_t}k^{n-1}, \quad (3.1.1.1)$$

donde  $c_{1_t}, c_{2_t}, \dots, c_{(n-1)_t}$  son funciones lineales de la media de filtración  $m_t = \mathbb{E}[\theta_t|y_{1:t}]$  e independientes de  $k$ . Esto es, la tendencia esperada al tiempo  $t$  (que puede pensarse como el comportamiento esperado de  $Y_{t+k}$  para  $k \geq 1$ ) dada la información disponible hasta ese momento es un polinomio de orden  $n - 1$  en  $k$ . Se sabe que casi cualquier forma, que tenga la tendencia esperada de la serie, puede aproximarse mediante un polinomio, por esta razón estos modelos resultan sumamente útiles al querer describir un MDL con tendencia.

En el capítulo pasado se discutieron dos modelos, la caminata aleatoria más un ruido y el modelo de crecimiento lineal. Ambos son modelos de tendencia tales que  $n = 1$  y  $n = 2$  respectivamente, (es decir  $f_t(k)$  es un polinomio en  $k$  de orden 0 y 1 respectivamente). Discutamos en principio estos dos modelos que conocemos mejor par después pasar a discutir el caso general.

#### El modelo de nivel local

El modelo de nivel local, o caminata aleatoria está descrito, como ya se ha visto, por las siguientes ecuaciones:

$$\begin{aligned} Y_t &= \theta_t + v_t & v_t &\sim \mathcal{N}(0, V_t), \\ \theta_t &= \theta_{t-1} + w_t & w_t &\sim \mathcal{N}(0, W_t). \end{aligned}$$

Es claro que aquí  $F_t = 1 = G_t$  para toda  $t$ . Se tiene entonces que

$$\begin{aligned}
f_t(k) &= \mathbb{E}[Y_{t+k}|y_{1:t}] = F_{t+k}a_t(k) \\
&= a_t(k) = \mathbb{E}[\theta_{t+k}|y_{1:t}] = G_{t+k}a_t(k-1) \\
&= a_t(k-1) \\
&\vdots \\
&= a_t(0) = \mu_t.
\end{aligned}$$

Esto significa que para un tiempo fijo  $f_t(k) = \mu_t$  es una función constante para  $k$  (un polinomio de grado 0). Esto es,  $\mathbb{E}[Y_{t+1}|y_{1:t}] = \mathbb{E}[Y_{t+2}|y_{1:t}] = \dots = \mathbb{E}[Y_{t+k}|y_{1:t}] = \mu_t$ , es decir, lo que se espera de los valores futuros de  $\{Y_t\}$  es que no presenten tendencia a la alta ni a la baja, o en otras palabras, que mantengan su nivel (el nivel de la media de filtración  $\mu_t$ ).

Notemos que aquí la dimensión del vector de estados es 1. Por esta razón a este modelo se le llamará Modelo Polinómico de 1<sup>er</sup> orden.

### Modelo de tendencia lineal local

Este modelo también es conocido por el nombre de Modelo de Crecimiento Lineal, y está descrito, como ya se ha visto anteriormente por las siguientes ecuaciones:

$$\begin{aligned}
Y_t &= F\theta_t + v_t, & v_t &\sim \mathcal{N}(0, V_t) \\
\theta_t &= G\theta_{t-1} + w_t, & w_t &\sim \mathcal{N}_2(0, \text{diag}(\sigma_\beta^2, \sigma_\gamma^2)),
\end{aligned}$$

donde

$$\theta_t = \begin{pmatrix} \beta_t \\ \gamma_t \end{pmatrix}, \quad G = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad y \quad F = (1, 0).$$

O escrito de otra manera

$$\begin{aligned}
Y_t &= \beta_t + v_t, & v_t &\sim \mathcal{N}(0, V_t), \\
\beta_t &= \beta_{t-1} + \gamma_{t-1} + w_{t,1}, & w_{t,1} &\sim \mathcal{N}(0, \sigma_{\beta_t}^2), \\
\gamma_t &= \gamma_{t-1} + w_{t,2}, & w_{t,2} &\sim \mathcal{N}(0, \sigma_{\gamma_t}^2).
\end{aligned}$$

A  $\beta_t$  se le interpreta como el nivel local y a  $\gamma_t$  como la tasa de crecimiento local.

Revisemos entonces como se comporta  $f_t(k) = \mathbb{E}[Y_{t+k}|y_{1:t}]$  como función de  $k$  para este modelo.

$$\begin{aligned}
f_t(k) &= \mathbb{E}[Y_{t+k}|y_{1:t}] = F_{t+k}a_t(k) \\
&= Fa_t(k) = F\mathbb{E}[\theta_{t+k}|y_{1:t}] \\
&= (1, 0) \begin{pmatrix} \mathbb{E}[\beta_{t+k}|y_{1:t}] \\ \mathbb{E}[\gamma_{t+k}|y_{1:t}] \end{pmatrix} \\
&= \mathbb{E}[\beta_{t+k}|y_{1:t}] \\
&= \mathbb{E}[\beta_{t+(k-1)}|y_{1:t}] + \mathbb{E}[\gamma_{t+(k-1)}|y_{1:t}] \\
&= \mathbb{E}[\beta_{t+(k-1)}|y_{1:t}] + \mathbb{E}[\gamma_{t+(k-2)} + w_{t+(k-1),2}|y_{1:t}] \\
&= \mathbb{E}[\beta_{t+(k-1)}|y_{1:t}] + \mathbb{E}[\gamma_{t+(k-2)}|y_{1:t}] + \mathbb{E}[w_{t+(k-1),2}] \\
&= \mathbb{E}[\beta_{t+(k-1)}|y_{1:t}] + \mathbb{E}[\gamma_{t+(k-2)}|y_{1:t}] \\
&\quad \vdots \\
&= \mathbb{E}[\beta_{t+(k-1)}|y_{1:t}] + \mathbb{E}[\gamma_t|y_{1:t}] \\
&\quad \vdots \\
&= \mathbb{E}[\beta_t|y_{1:t}] + k\mathbb{E}[\gamma_t|y_{1:t}] \\
&= \widehat{\beta}_t + k\widehat{\gamma}_t.
\end{aligned}$$

Si escribimos a  $f_t(k)$  de la siguiente forma

$$f_t(k) = (1, 0)\mu_t + (0, 1)\mu_t k = c_{0t} + c_{1t}k.$$

Es evidente que  $f_t(k)$  es una función lineal (ó un polinomio de 1<sup>er</sup> grado) de  $k$  donde  $c_{0t}$  y  $c_{1t}$  son funciones lineales de  $\mu_t$ . Esto significa que la tendencia esperada al tiempo  $t$  de la serie de tiempo es una recta con pendiente  $\widehat{\gamma}_t$  y que cuando  $k = 0$  la recta se encuentra en  $\widehat{\beta}_t$ . Esto nos proporciona una explicación de porque  $\beta_t$  y  $\gamma_t$  son interpretados como el nivel local y la tasa de crecimiento local de la serie respectivamente.

Como en este caso la dimensión del vector de estados es 2, este modelo recibe su tercer nombre de Modelo Polinómico de 2<sup>do</sup> orden.

### El Modelo Polinómico de n-ésimo orden

La caminata aleatoria así como el Modelo de Crecimiento Lineal son casos particulares de este modelo con  $n = 1$  y  $n = 2$  respectivamente. Para el caso general se tiene que el modelo es el siguiente:

$$\begin{aligned}
Y_t &= F\theta_t + v_t, & v_t &\sim \mathcal{N}(0, V_t), \\
\theta_t &= G\theta_{t-1} + w_t, & w_t &\sim \mathcal{N}_n(0, W_t),
\end{aligned}$$

donde

$$\theta_t = \begin{pmatrix} \theta_t^{(1)} \\ \theta_t^{(2)} \\ \vdots \\ \theta_t^{(n)} \end{pmatrix}, \quad G = \begin{pmatrix} 1 & 1 & 0 & \cdots & 0 \\ 0 & 1 & 1 & 0 & \cdots & 0 \\ \vdots & & & \ddots & \vdots & \\ 0 & \cdots & 0 & 1 & 1 \\ 0 & \cdots & & 0 & 1 \end{pmatrix},$$

$$F = (1, 0, \dots, 0), \quad y \quad W_t = \text{diag}(W_t^{(1)}, W_t^{(2)}, \dots, W_t^{(n)})$$

En términos de sus componentes el modelo puede ser, equivalentemente, escrito como

$$\begin{aligned} Y_t &= \theta_t^{(1)} + v_t, & v_t &\sim \mathcal{N}(0, V_t), \\ \theta_t^{(p)} &= \theta_{t-1}^{(p)} + \theta_{t-1}^{(p+1)} + w_t^{(p)}, & w_t^{(p)} &\sim \mathcal{N}(0, W_t^{(p)}), \\ \theta_t^{(n)} &= \theta_{t-1}^{(n)} + w_t^{(n)}, & w_t^{(n)} &\sim \mathcal{N}(0, W_t^{(n)}), \end{aligned}$$

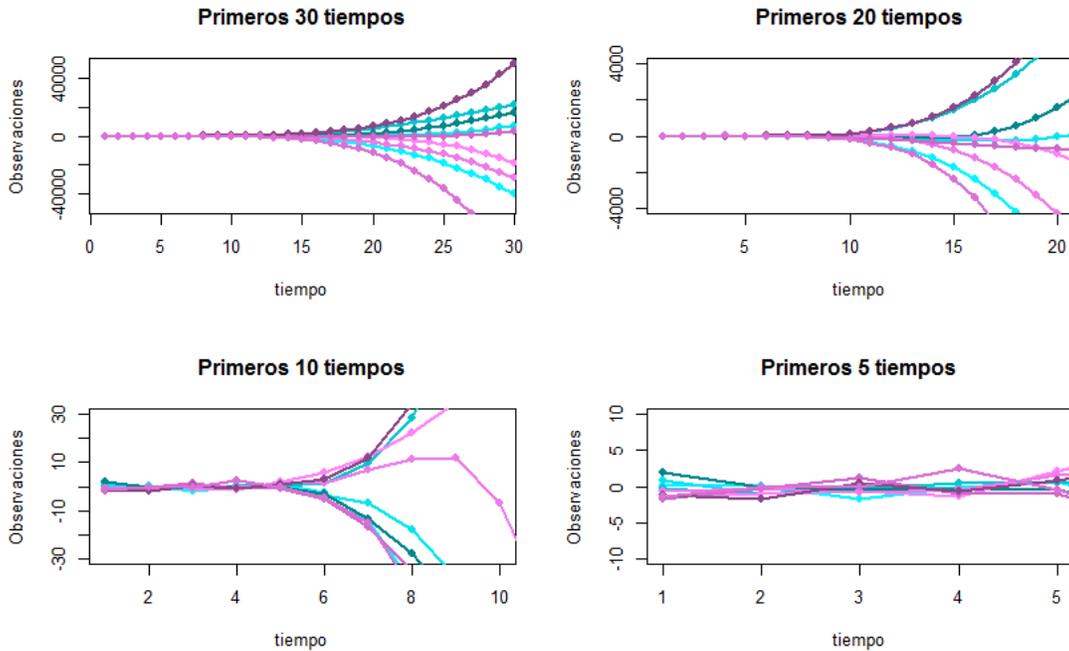
donde  $p = 1, 2, \dots, n - 1$

Al igual que con los modelos polinómicos de 1<sup>er</sup> y 2<sup>do</sup> orden, se puede demostrar que  $f_t(k)$  es un polinomio de grado  $n - 1$  en  $k$  tal que todos sus coeficientes son funciones lineales de  $\mu_t = \mathbb{E}[\theta_t | y_{1:t}]$ . La prueba de esto para el caso general es similar a la prueba para el caso particular del Modelos de Crecimiento Lineal.

Si en un modelo polinómico de n-ésimo orden hacemos  $W_t^{(1)} = \dots = W_t^{(n-1)} = 0$ , el modelo que resulta se conoce como *Caminata aleatoria integrada*.

Para fines ilustrativos se ha simulado una caminata aleatoria integrada de orden 5, donde  $W_t^{(5)} = 1$ ,  $V_t = 1$ ,  $\mu_0 = (0, \dots, 0)$  y  $C_0 = \text{diag}(0, \dots, 0, 1)$ . A continuación se presenta una gráfica de una muestra de tamaño 9 de este proceso.

Figura 3.1.1.1: fig:9 Simulaciones de una Caminata aleatoria integrada de orden 5



Las gráficas anteriores muestran distintas partes del recorrido de cada una de las 9 simulaciones de la caminata aleatoria integrada de orden 5 para poder ilustrar mejor el comportamiento de este proceso.

En la librería `d1m` en R existe una función mediante la cual se pueden especificar modelos polinómicos de manera más sencilla que si se hiciera a través de la función `d1m()`. Dicha función se describe a continuación.

- `d1mModPoly()`: Esta función recibe como parámetros:
  - (i) `order`: El orden del polinomio. En caso de no ser especificado la función lo supone igual a 2.
  - (ii) `dV`: La varianza de el ruido o error de la ecuación de observaciones. En caso de no ser especificada se supone igual a 1.
  - (iii) `dW`: Un vector de tamaño  $n = \text{order}$  que representa a los elementos diagonales de la matriz de covarianza del ruido de la ecuación de estados. En caso de que no se especifique se supondrá que  $dW = (0, \dots, 0, 1)$ .
  - (iv) `m0`: El vector de medias de la distribución inicial del vector de estados, es un vector de tamaño  $n = \text{order}$  que en caso de no ser especificado se supone igual a  $(0, \dots, 0)$ .

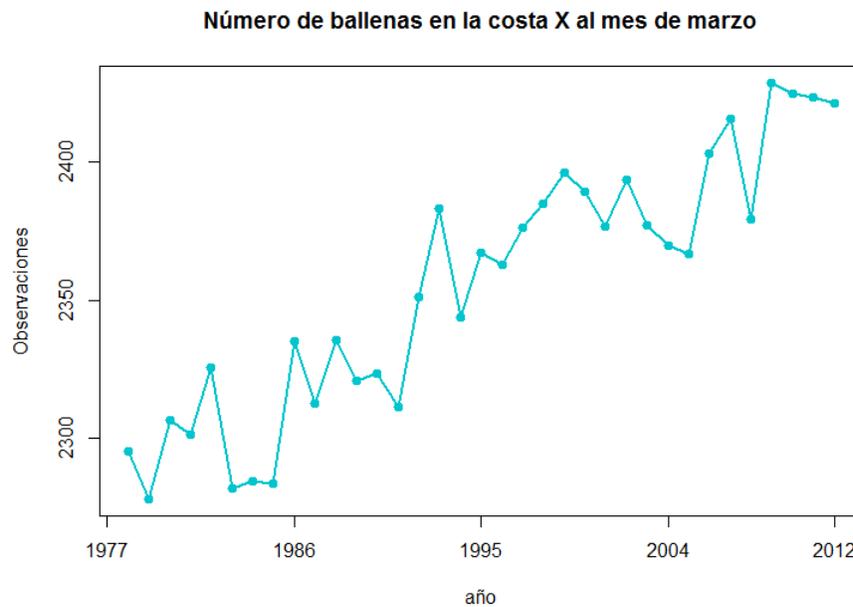
- (v) **C0**: La matriz de covarianza de la distribución inicial del vector de estados, es un vector de tamaño  $n = \text{order}$  que representa a la diagonal de la matriz correspondiente. En caso de no ser especificada se supondrá que es la matriz  $\text{diag}(10^7, \dots, 10^7)$ .

La salida de esta función es un objeto de clase `d1m`, el cual, junto con alguna serie de tiempo (a la que se le propuso el modelo), se puede pasar como entrada a la función `d1mFilter()` para calcular las densidades de filtración y de predicción de un paso adelante. A continuación, si se deseara hacer análisis en retrospectiva o predicción de observaciones y/o estados futuros, con la salida de la función `d1mFilter()` y utilizando las funciones `d1mSmooth()` y `d1mForecast()` se pueden calcular las densidades de suavización y de predicción.

### Un ejemplo

Supongamos que un grupo de biólogos marinos está estudiando al número de ballenas que migran a una pequeña costa, durante los meses de enero a abril. Ellos están trabajando fuerte para que cada vez esta costa sea un mejor lugar para que las ballenas tengan sus crías. Se nos han proporcionado observaciones que se han venido tomando a lo largo del tiempo y se nos pide que ajustemos un modelo adecuado a la serie de tiempo así como que hagamos una predicción (dada la información disponible) de cual será la observación para cada uno de los siguientes 10 años. La información que se nos ha proporcionado se muestra a continuación.

Figura 3.1.1.2: Observaciones del número de ballenas en la costa al mes de marzo



Al ver la serie de tiempo nos damos cuenta que tiene una tendencia que aparenta ser lineal, sin embargo, se proponen estos 3 modelos:

1. El primer modelo al que llamaremos `mod1` es un Modelo de Crecimiento Lineal, en particular una caminata aleatoria integrada de orden 2. Para el cual se va a suponer  $\mu_0 = (2300, 0)$ ,  $V_t = 20$  y el resto de los parámetros se tomarán las que propone la función `d1mModPoly()` en caso de no ser especificados.
2. El segundo modelo se va a referir a una caminata aleatoria integrada de orden 3. Para este modelo se supondrá  $\mu_0 = (2300, 0, 0)$ ,  $V_t = 20$  y en cuanto a los parámetros restantes vamos a respetar los que propuestos por la función `d1mModPoly()` . A este modelo lo llamaremos `mod2`.
3. El tercer y último modelo que propondremos será un Modelo de nivel local, es decir una caminata aleatoria más un ruido, dando continuidad, vamos a nombrar a este modelo `mod3` . De igual manera que en los dos modelos anteriores, la mayor parte de los parámetros que tomaremos serán los propuestos por la función `d1mModPoly()` , con excepción de la media inicial que será  $\mu_0 = 2, 300$  y la varianza de la ecuación de observaciones que será  $V_t = 20$ .

Mediante el siguiente código de R, se han propuesto estos tres modelos. Para cada uno de los modelos se ha calculado el filtro y se ha hecho la predicción de las observaciones para los siguiente 10 tiempos. Además a través de los criterio MAD, MSE y MAPE se ha decidido cual modelo se ajusta mejor a la serie de tiempo.

Sea  $Y$  el vector en R que representa a la serie de tiempo proporcionada.

---

### Código en R

---

```
#Cargamos la librería
library(dlm)

#Se crean las funciones calculan los criterios MAD, MSE y MAPE
MAD <- function(u,v)
{
  r <- mean(abs(u-v))
  return(r)
}
MSE <- function(u,v)
{
  r <- mean((u-v)^{2})
  return(r)
}
MAPE <- function(u,v)
{
```

```

    r <- mean(abs(u-v)/v)
    return(r)
}

#Se propone y se calcula el filtro para cada modelo
mod1 <- dlmModPoly(2, dV = 20, m0 = c(2300,0))
mod1Filt <- dlmFilter(Y,mod1)

mod2 <- dlmModPoly(3, dV = 20, m0 = c(2300,0,0))
mod2Filt <- dlmFilter(Y,mod2)

mod3 <- dlmModPoly(1, dV = 20, m0 = 2300)
mod3Filt <- dlmFilter(Y,mod3)

#Se calcula para cada modelo la media de predicción de la observaciones
#para cada uno de los siguiente 10 años
mod1Fore <- dlmForecast(mod1Filt, nAhead = 10)
mod2Fore <- dlmForecast(mod2Filt, nAhead = 10)
mod3Fore <- dlmForecast(mod3Filt, nAhead = 10)

#Se hace una gráfica donde se muestran las observaciones asi como la
#predicción de un paso adelante para cada tiempo t
plot(Y, type = "o", pch = 19, lwd = 2, col = "turquoise3",
xlim = c(1,45), ylim = c(2250,2480), xlab = "tiempo", xaxt = "n",
ylab = "Observaciones y predicciones de un paso adelante",
main = "Número de ballenas en la costa X al mes de marzo" )
lines(mod1Filt$f, type = "o", pch = 16, lwd = 1, lty = 4, col = "orchid1")
lines(mod2Filt$f, type = "o", pch = 16, lwd = 1, lty = 4, col = "orchid3")
lines(mod3Filt$f, type = "o", pch = 16, lwd = 1, lty = 4, col = "orchid4")
axis(1, at = c(0,9,18,27,36,45), labels = c(1977,1986,1995,2004,2013,2022))
leg <- c("Las observaciones  $Y(t) = y(t)$ ", "E[Y(t)|y(1:t-1)] según el modelo 1",
"E[Y(t)|y(1:t-1)] según el modelo 2", "E[Y(t)|y(1:t-1)] según el modelo 3")
legend("bottomright", legend = leg,
col = c("turquoise3", "orchid1", "orchid3", "orchid4"),
lty = c(1,4,4,4), lwd = c(2,1,1,1), pch = c(19,16,16,16), bty = "n")

#Se calculan los criterio MAD, MSE y MAPE para cada modelo
print(MAD(mod1Filt$f,Y))
print(MSE(mod1Filt$f,Y))
print(MAPE(mod1Filt$f,Y))

print(MAD(mod2Filt$f,Y))
print(MSE(mod2Filt$f,Y))

```

```

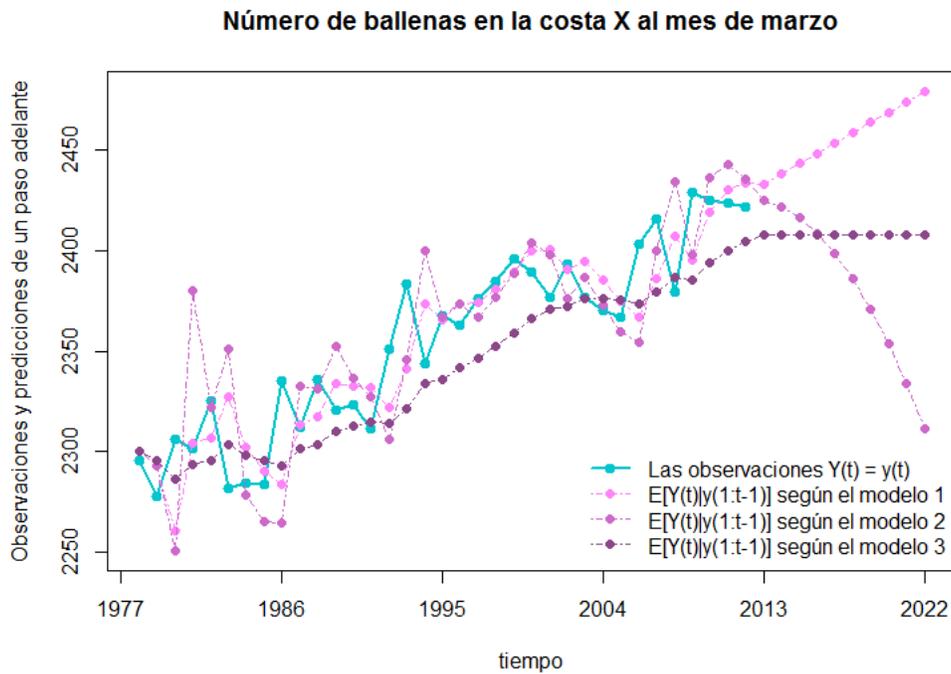
print(MAPE(mod2Filt$f,Y))

print(MAD(mod3Filt$f,Y))
print(MSE(mod3Filt$f,Y))
print(MAPE(mod3Filt$f,Y))

```

Del código anterior se obtiene la siguiente gráfica

Figura 3.1.1.3: Predicción del número de ballenas tomando en cuenta la información disponible de años anteriores



**Nota:** Para los años 2014 en adelante la gráfica correspondiente a cada modelo representa  $\mathbb{E}[Y_t|Y_{t-k}]$  donde  $2 \leq k \leq 10$  y a cada tiempo  $Y_{t-k}$  es la observación correspondiente al año 2012.

y los criterios MAD, MSE y MAPE para cada modelo

	MAD	MSE	MAPE
mod1	18.006	520.2255	0.0077
mod2	24.3266	1058.249	0.0103
mod3	21.5213	653.7183	0.0091

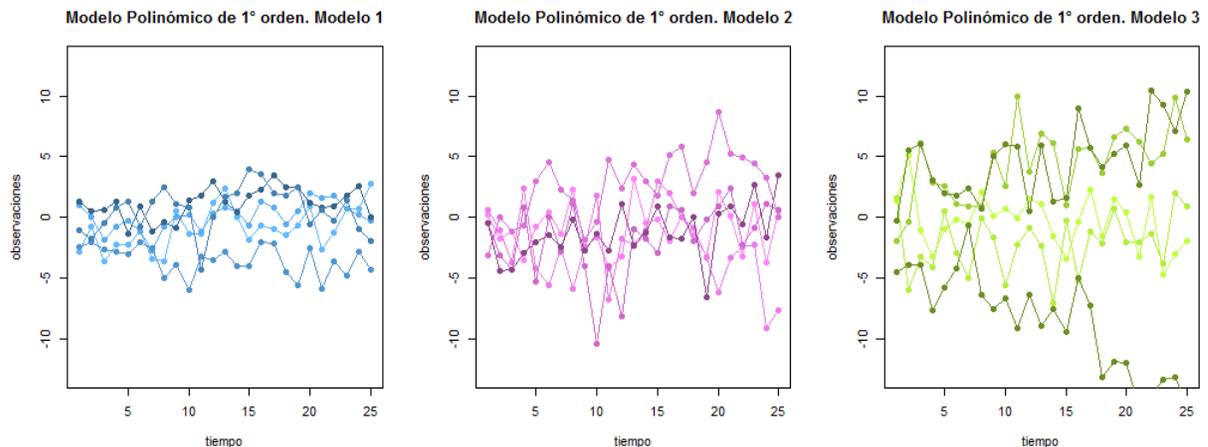
Analizando la gráfica, así como cada uno de los criterios, se puede concluir que el modelo que se ajusta mejor es el modelo de crecimiento lineal, mod1, por lo que las predicciones para los 10 siguientes años será las obtenidas a través de este modelo.

Para cerciorarse de que efectivamente el modelo es adecuado, se podría realizar la revisión del modelo explicado en el capítulo pasado. Por ahora continuemos con la especificación del modelo.

Antes de concluir esta sección cabe mencionar que los modelos de tendencia más usados son aquellos que ajustan tendencias más suaves, es decir, los de orden menor. Resulta un buen ejercicio, para conocer mejor el comportamiento de cada uno de estos modelos, hacer simulaciones de ellos variando cada uno de los parámetros.

Por ejemplo, tomemos un Modelo Polinómico de orden 1 y propongamos 3 variantes, a la primera propuesta se asignará  $V_t = 1$ , a la segunda  $V_t = 5$  y la tercera  $V_t = 10$ . Las tres variantes tendrán en común  $C_0 = 1$ ,  $W_0 = 1$ , y el resto de los parámetros serán los propuestos por la función `d1mModPoly()`. De esta manera estamos modificando la varianza observacional dejando fijos el resto de los parámetros. Para cada una de estas variantes se han generado 5 muestras las cuales se grafican a continuación.

Figura 3.1.1.4: Simulaciones de un Modelo Polinómico de orden 1



Podemos notar que el Modelo 1 (en el que  $V_t = 1$ ) las observaciones no se dispersan o varían tanto mientras que para el Modelo 3 (en el que  $V_t = 10$ ) las observaciones se dispersan mucho. Los tres modelos aparentan no tener una tendencia creciente ni decreciente, más bien conservan un nivel local en 0.

Amplíemos un poco el panorama proponiendo 3 variantes de un Modelo de Polinómico de orden 2 y 3 variantes de uno de orden 3.

Todos los Modelos de Crecimiento Lineal tendrán  $C_0 = \text{diag}(10, 10) = W_t$  y el resto de los parámetros (con excepción de  $V_t$ ), serán los propuestos por la función `d1mModPoly()`.

Para el Modelo 1 de orden 2  $V_t = 10$ , para el Modelo 2  $V_t = 50$  y para el Modelo 3  $V_t = 100$ .

Análogamente, los Modelos Polinómicos de orden 3 serán tales que  $C_0 = W_t = \text{diag}(10, 10, 10)$ , y los parámetros (excepto  $V_t$ ) serán los propuestos por la función `d1mModPoly()`. En esta ocasión los Modelos 1,2 y 3 tendrán  $V_t = 100$ ,  $V_t = 10000$  y  $V_t = 1000000$  respectivamente.

Al igual que para los modelos de nivel local, para cada uno de estos modelos se han tomado 5 muestras de las que se muestra la representación gráfica a continuación

Figura 3.1.1.5: Simulaciones de un Modelo Polinómico de orden 2

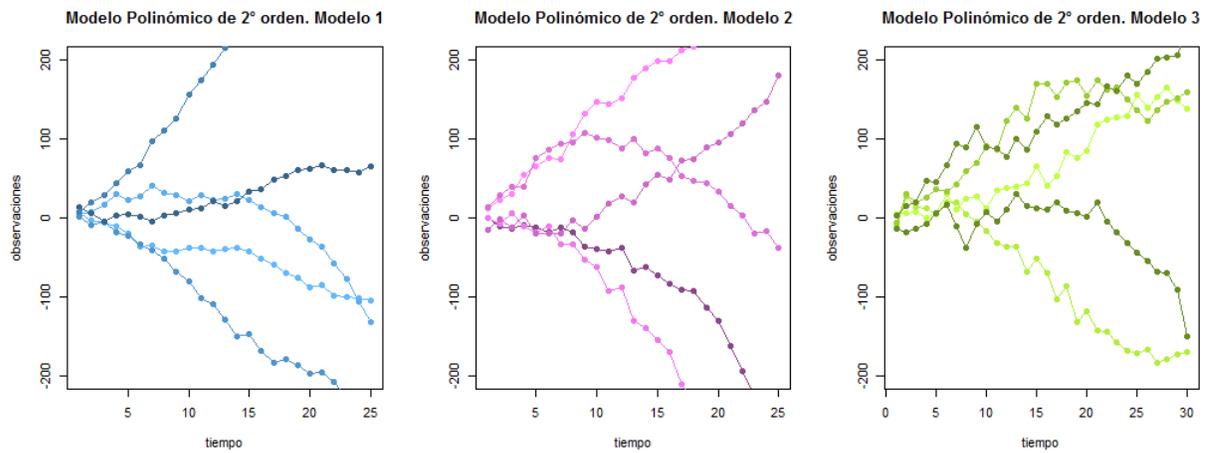
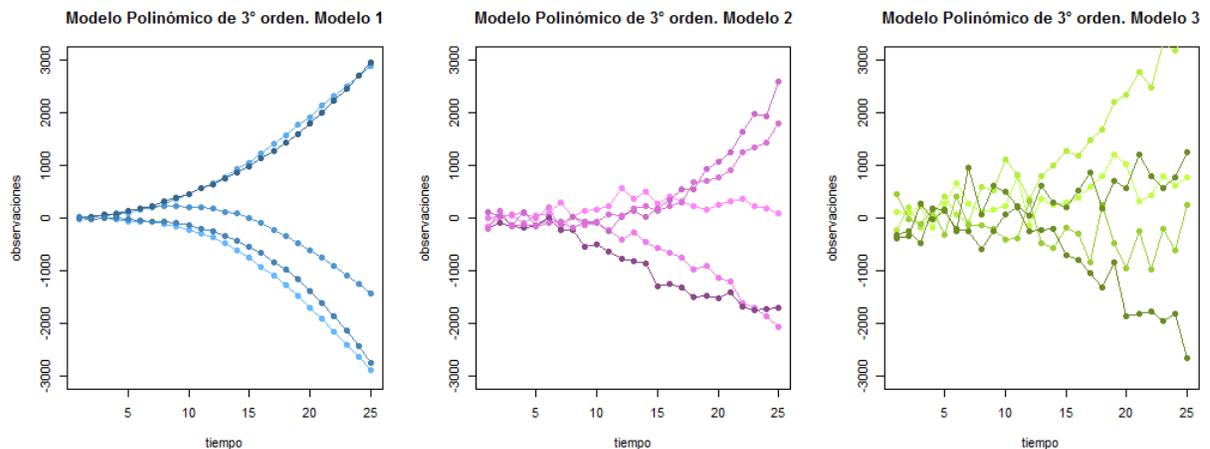


Figura 3.1.1.6: Simulaciones de un Modelo Polinómico de orden 3



Notemos que los modelos polinómicos de orden 2 en general aparentan crecer o decrecer de una manera constante una vez más, el hecho de que la varianza observacional sea mayor lo que hace es que las observaciones se dispersen más de su media. En cuanto a los modelos polinómicos de orden 3 podemos ver que estos crecen o decrecen de una manera mucho más rápida que los modelos de orden menor, de la misma forma, conforme va pasando el tiempo la velocidad de crecimiento (o decrecimiento) es mayor, esto resulta muy claro en la gráfica del Modelo 1 de tercer grado.

Al observar estas gráficas se puede concluir que la entre más grande sea  $V_t$ , las observaciones presentarán una desviación mayor de su media esperada, es decir el ruido” será mayor.

El estudiar de esta manera, simulando distintos Modelos Dinámicos Lineales nos ayuda mucho a conocer mejor el porque del comportamiento de cada uno, y de esta forma al tener una serie de tiempo que requiere ser analizada, la experiencia adquirida a través de este estudio será de mucha utilidad.

Por ahora abandonemos a los Modelos de tendencia y enfoquemonos en aquellos que explican los ciclos en las series de tiempo.

### 3.1.2. Modelos de comportamiento cíclico

En este documento se presentarán dos alternativas de modelar una serie de tiempo que presenta una conducta cíclica. El primer modelo que se analizará se llama Modelo de factores estacionales, el segundo recibe el nombre de Modelo estacional con forma de Fourier.

#### Modelos de factores estacionales

Supongamos que se tiene una serie de tiempo  $\{Y_t\}$ , que muestra un comportamiento cíclico, es decir tiene una fluctuación en forma de onda alrededor de la tendencia. Por simplicidad supongamos que la serie de tiempo tiene media cero, en caso de tener una media distinta de cero o de tener algún tipo de tendencia, esta se puede representar mediante un componente de tendencia ajeno al componente de ciclicidad.

Veamos un caso particular antes de presentar el modelo general. Imaginemos que se tiene una serie de tiempo cíclica,  $\{Y_t\}$ , tal que se toman observaciones (equidistantes) 4 veces en el año (supongamos sin pérdida de generalidad que se toman el primer día de cada una de las 4 estaciones). De tal forma que si  $Y_1$  se refiere a la observación correspondiente a la primavera, entonces  $Y_2$  se referirá al verano,  $Y_3$  al otoño,  $Y_4$  al invierno y evidentemente  $Y_5$  de nuevo a la primavera. De esta manera, todas las observaciones tomadas en primavera serán de la forma  $Y_{4k+1}$ , todas las referentes al verano tendrán la forma  $Y_{4k+2}$ , las referentes al otoño serán de la forma  $Y_{4k+3}$ , y por último las observaciones de la forma  $Y_{4k}$  se habrán tomado en invierno, todo esto para alguna  $k \in \mathbb{N}$ . Supongamos que existen desviaciones de la media expresadas por los coeficientes  $\tau_1, \tau_2, \tau_3$  y  $\tau_4$  (cada uno correspondiente a una estación del año), tales que  $Y_1 = \tau_1 + v_1$  donde  $v_1$  es un error aleatorio. Y si para alguna  $t$

$$Y_t = \tau_i + v_t \quad (3.1.2.1)$$

entonces

$$Y_{t+1} = \tau_{i+1} + v_{t+1}, \quad (3.1.2.2)$$

donde  $i = 1, 2, 3, 4$  y  $\tau_5 = \tau_1$ .

Este modelo puede ser expresado como un MDL de la siguiente manera: Sean

$$\theta_t = \begin{pmatrix} \tau_{t,1} \\ \tau_{t,2} \\ \tau_{t,3} \\ \tau_{t,4} \end{pmatrix}, \quad F_t = F = (1, 0, \dots, 0), \quad y \quad G_t = G = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix},$$

donde  $\tau_{t,1}$  se refiere a la desviación correspondiente a la observación tomada al tiempo  $t$ , sean además las ecuaciones de observaciones y del sistema como siempre:

$$\begin{aligned} Y_t &= F\theta_t + v_t, & v_t &\sim \mathcal{N}(0, V_t), \\ \theta_t &= G\theta_{t-1} + w_t, & w_t &\sim \mathcal{N}_4(0, W_t). \end{aligned}$$

Notemos entonces que a cada tiempo la matriz  $F\theta_t$  es igual a la primera entrada del vector  $\theta_t$  ( $\tau_{t,1}$ ), además  $G\theta_{t-1}$  da como resultado una permutación del vector  $\theta_{t-1}$  tal que si  $\theta_{t-1} = (\tau_{t-1,1}, \tau_{t-1,2}, \tau_{t-1,3}, \tau_{t-1,4})^T$ , entonces

$$\begin{aligned} \theta_t &= G\theta_{t-1} + w_t \\ &= (\tau_{t-1,2}, \tau_{t-1,3}, \tau_{t-1,4}, \tau_{t-1,1})^T + w_t \\ &= (\tau_{t,1}, \tau_{t,2}, \tau_{t,3}, \tau_{t,4})^T + w_t, \end{aligned}$$

es decir la matriz  $G$  va rotando los elementos del vector de estados.

Además para que se cumplan las ecuaciones 3.1.2.1 y 3.1.2.2 para toda  $t$  entonces  $W_t = W = 0_{4 \times 4}$ . En este caso el modelo se llama Modelo estacional estático. En general, si  $W_t \neq 0$  entonces las 4 desviaciones de la media  $\tau_1, \tau_2, \tau_3$  y  $\tau_4$  van a ir cambiando en el tiempo, por esta razón hay que ser cuidadosos al especificar a la matriz  $W_t$  de tal manera que los procesos, que generan cada una de las desviaciones de la media,  $\{\tau_n^{(i)}\}_{n \geq 1}$  donde  $\tau_n^{(i)} = \tau_{i+(n-1)*s}$  para  $i = 1, 2, 3, 4$  se comporten de la manera deseada.

En los párrafos anteriores analizamos el modelo para el caso en que el periodo de la serie es  $s = 4$ . Para el caso general en el que el periodo es  $s$ , la serie de tiempo puede modelarse a través de un MDL (con las mismas ecuaciones de observaciones y del sistema) tal que el vector de desviaciones estacionales  $\theta_t \in \mathbb{R}^s$ , y  $G$  es una matriz de permutación de  $s \times s$  como sigue:

$$\theta_t = \begin{pmatrix} \tau_{t,1} \\ \tau_{t,2} \\ \vdots \\ \tau_{t,s} \end{pmatrix}, \quad G_t = G = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & & \ddots & \ddots & & \vdots \\ 0 & 0 & \cdots & 0 & 1 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 \\ 1 & 0 & \cdots & & 0 & 0 \end{pmatrix}.$$

(Es decir,  $\theta_t$  es un vector  $s$ -dimensional tal que su primera entrada es el factor estacional correspondiente a la observación  $Y_t$ , y  $G = (g_{i,j})_{i=1:s,j=1:s}$  es la matriz de dimensión  $s \times s$  tal que  $g_{s,1} = 1 = g_{i,i+1}$  para toda  $i = 1, 2, \dots, s-1$  y el resto de las entradas es 0.)

La matriz  $F_t$  es un vector de  $1 \times s$  constante en el tiempo  $F = (1, 0, \dots, 0)$ . Por lo general se exigen restricciones de identificabilidad sobre los factores estacionales, comunmente se propone  $\sum_{i=1}^s \tau_i = 0$ . Esto tiene sentido, ya que significa que si la suma de los factores que están por arriba de la media es de  $x$  unidades, entonces la suma de los factores que están por abajo de la media deberá de ser  $-x$ , (esto ya que estamos suponiendo que la media es 0). Está restricción implica que hay  $s-1$  factores estacionales libres, ya que si los primeros  $s-1$  factores está determinados el último factor será  $\tau_s = -\sum_{i=1}^{s-1} \tau_i$ . Esto sugiere un modelo alternativo equivalente que utiliza un vector de estados  $s-1$  dimensional y una matriz  $G_t = G$  de  $(s-1) \times (s-1)$ .

$$\theta_t = \begin{pmatrix} \tau_{t,1} \\ \tau_{t,2} \\ \vdots \\ \tau_{t,s-1} \end{pmatrix}, \quad G_t = G = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & & \ddots & \ddots & & \vdots \\ 0 & 0 & \cdots & 0 & 1 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 \\ -1 & -1 & \cdots & & -1 & -1 \end{pmatrix}.$$

Con  $G$  definida de esta manera se tiene que para cualquier tiempo  $t$

$$\begin{aligned} \theta_{t+1} &= G\theta_t + w_t = G(\tau_{t,1}, \tau_{t,2}, \dots, \tau_{t,s-1})^T + w_t \\ &= (\tau_{t,2}, \tau_{t,3}, \dots, \tau_{t,s-1}, -\sum_{i=1}^{s-1} \tau_{t,i})^T + w_t \\ &= (\tau_{t,2}, \tau_{t,3}, \dots, \tau_{t,s-1}, \tau_{t,s})^T + w_t \\ &= (\tau_{t+1,1}, \tau_{t+1,2}, \dots, \tau_{t+1,s-1})^T. \end{aligned}$$

Además la matriz  $F_t$  para este modelo es un vector  $s-1$  dimensional, nuevamente constante en el tiempo,  $F = (1, 0, \dots, 0)$ .

Una variación dinámica en las desviaciones estacionales puede ser introducida al hacer que la matriz de covarianza sea  $W_t \neq 0$ , en la práctica una de las más comunmente propuestas es  $W_t = \text{diag}(\sigma^2, 0, \dots, 0)$  donde claramente  $\sigma^2$  es una constante mayor que 0.

En la librería `d1m` de R la función descrita a continuación ajusta modelos de factores estacionales.

■ `d1mModSeas()`: Los parámetros de esta función son:

- (i) `frequency`: El número de periodos de la serie.
- (ii) `dV`: La varianza de la ecuación de observaciones. En caso de no ser especificada se supone 1.
- (iii) `dW`: Un vector de tamaño `frequency-1` que representa a la diagonal de la matriz de covarianza de la ecuación del sistema. De no ser especificado se toma igual a  $(1, 0, \dots, 0)$ .
- (iv) `m0`: Es un vector de dimensión `frequency-1`, se refiere a la vector de medias inicial de las  $s - 1$  (donde  $s = \text{frequency}$ ) desviaciones libres del modelo <sup>12</sup>. En caso de no ser especificado se supone es igual a  $(0, \dots, 0)$ . Este parámetro es necesario especificarlo (distinto a  $(0, \dots, 0)$ ) para que el modelo tenga sentido (de otra manera no tiene sentido definir desviaciones de la media que son iguales a la media).
- (v) `C0`: Se refiere a una matriz de tamaño  $(s - 1) \times (s - 1)$  (con  $s = \text{frequency}$ ) que representa a la matriz de covarianza inicial del sistema. en caso de no ser especificada se supone  $C_0 = \text{diag}(10^7, \dots, 10^7)$ .

La salida de esta función es un objeto de clase `d1m` que representa al componente cíclico, de periodo (o frecuencia) `frequency`, de una serie de tiempo.

No está de más hacer notar que esta función trabaja con una matriz  $G$  que rota los elementos del vector de estados en la otra dirección de la especificada anteriormente, es decir si para algún tiempo  $\theta_t = (\tau_{t,1}, \tau_{t,2}, \dots, \tau_{t,s-1})^T$  entonces

$$\begin{aligned} \theta_{t+1} &= G\theta_t = (\tau_{t,1}, \tau_{t,2}, \dots, \tau_{t,s-1})^T + w_t \\ &= \left(-\sum_{i=1}^{s-1} \tau_{t,i}, \tau_{t,1}, \tau_{t,2}, \dots, \tau_{t,s-2}\right)^T + w_t \\ &= (\tau_{t,s}, \tau_{t,1}, \tau_{t,2}, \dots, \tau_{t,s-2})^T + w_t \\ &= (\tau_{t+1,1}, \tau_{t+1,2}, \dots, \tau_{t+1,s-1})^T, \end{aligned}$$

de forma que la matriz  $G$  que rota los estados de esta manera es

$$G = \begin{pmatrix} -1 & -1 & \dots & -1 & -1 \\ 1 & 0 & & 0 & 0 \\ 0 & 1 & & 0 & 0 \\ & & \ddots & & \\ 0 & 0 & & 1 & 0 \end{pmatrix}.$$

---

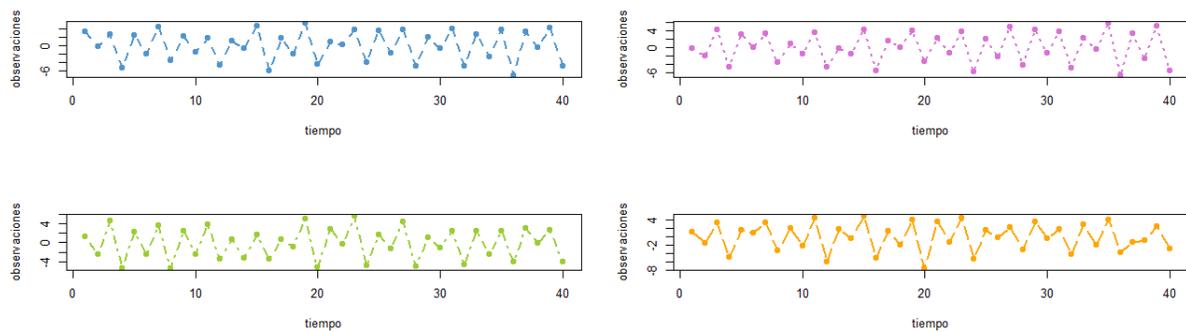
<sup>12</sup>la media de la desviación restante está determinada por  $\mathbb{E}[\tau_s] = -\sum_{i=1}^{s-1} \mathbb{E}[\tau_i]$

Notemos que de esta manera hay que tener cuidado al definir el vector de estados, de forma que si antes hubieramos definido  $\theta_0 = (\tau_s, \tau_1, \dots, \tau_{s-2})$  ahora lo haríamos de esta manera  $\theta_0 = (\tau_s, \tau_{s-1}, \dots, \tau_2)$ , de tal forma que al permmutar el vector de estados, de ambas maneras, a cada tiempo, se tenga en la primera entrada la desviación adecuada.

Al igual que como se hizo con los modelos de tendencia, con el propósito de ejemplificar y conocer mejor estos modelos, se han propuesto 4 modelos y par cada uno se han generado 4 muestras.

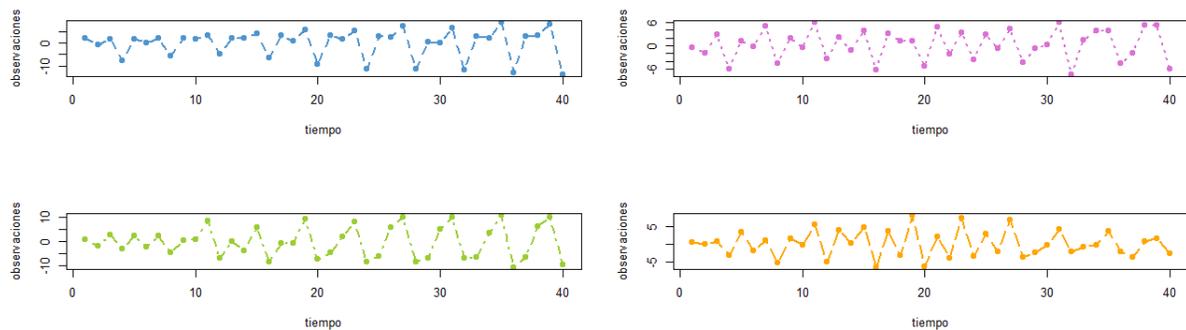
El primero será un modelo estacional estático tal que  $s = \text{frequency} = 4$ ,  $\tau_1 = 2$ ,  $\tau_2 = -1$ ,  $\tau_3 = 4$  y  $\tau_4 = -5$ . Es decir, el primer modelo será el definido por `dlmModSeas(frequency = 4, m0 = c(-5,4,-1), dV = 1, C0 = diag(c(0,0,0)), dW = c(0,0,0))`.

Figura 3.1.2.1: Modelo 1. Modelo estacional estático de periodo 4



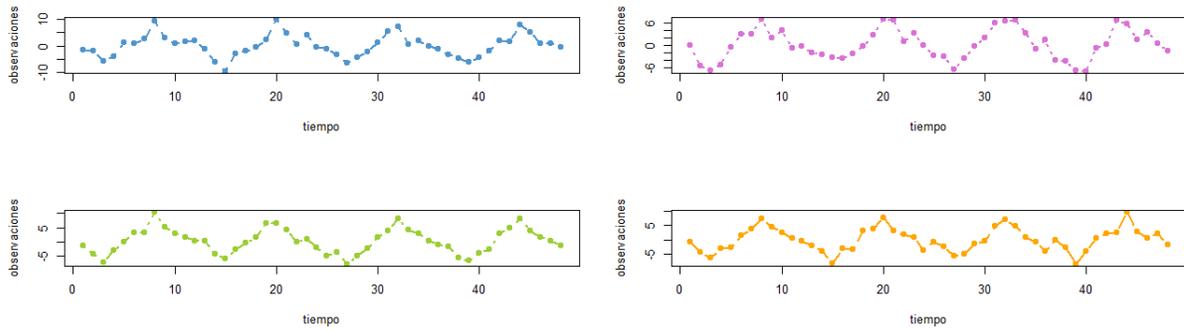
El segundo modelo será una variante no estática del primer modelo. Es decir será igual al primer modelo con la diferencia que de  $W_t = \text{diag}(1, 0, 0)$ .

Figura 3.1.2.2: Modelo 2. Modelo estacional no estático de periodo 4



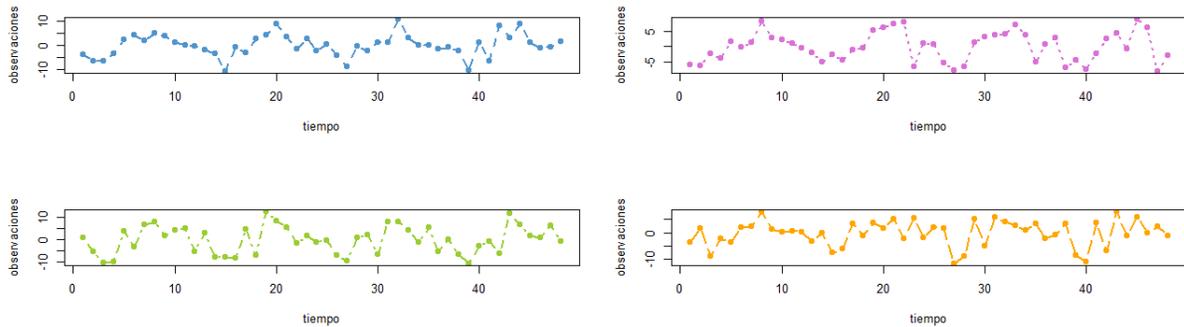
El tercer modelo será un modelo estacional estático con  $s = \text{frequency} = 12$ , y quedará definido por el especificado por `dlmModSeas(frequency = 12, m0 = c(-1,1,2,4,7,4,1,-1,-4,-7,-4), dV = 2, C0 = diag(rep(0,11)), dW = rep(0,11))`.

Figura 3.1.2.3: Modelo 3. Modelo estacional estático de periodo 12



El cuarto y último modelo será una variante no estática del tercer modelo. Esto es, será igual al modelo anterior con la diferencia que de  $W_t = \text{diag}(2, 0, \dots, 0)$ .

Figura 3.1.2.4: Modelo 4. Modelo estacional no estático de periodo 12



Estas gráficas nos muestran el comportamiento de los Modelos de Factores Estacionales. En las gráficas 1 y 3 se ilustra de manera muy clara y sencilla que cada 12 tiempos el nivel de la serie tiende a repetirse. Es decir  $Y_t \stackrel{d}{=} Y_{t+k}$ .

Por otro lado, notese que en los modelos 2 y 4 a comparación de los modelos 1 y 3 respectivamente, se puede apreciar de manera muy clara que en el modelo estático las desviaciones se mantiene constantes en el tiempo y las observaciones son una medición imprecisa de estas, mientras que en el modelo no estático estas desviaciones tienden a

modificarse en el tiempo. Por ejemplo en el caso de  $s = 4$ , en el Modelo 1 podemos ver que la serie forma una secuencia de "M"s más o menos uniformes, mientras que en el caso no estático estas "M"s aparecen en alguna parte pero tienden a distorcionarse e incluso a cambiar por completo su forma.

Se puede aprender mucho más de estos modelos si seguimos simulando y variando cada uno de los parámetros del modelo. En este caso sólo se exploró el periodo de la serie y la matriz de covarianza del sistema. Esto porque los casos más comunes son  $s = 4$  y  $s = 12$  (datos trimestrales y mensuales respectivamente) y resulta además muy ilustrativo y útil ver la manera en la que estos modelos permiten ajustar un ciclo que se modifique en el tiempo.

### Modelos estacionales con forma de Fourier

Considera una función  $g : T \rightarrow \mathbb{R}$  (donde  $T = \{1, 2, \dots\}$  es un conjunto discreto de tiempos) tal que  $g_t = g(t)$  tiene periodo  $s$ , es decir se caracteriza por los valores que toma en  $t = 1, 2, \dots, s$  y es tal que  $g_{i+ks} = g_i = \tau_i$  para toda  $i = 1, 2, \dots, s$  y para toda  $k \in \mathbb{N}$ . Es posible asociarle a esta función un vector  $\tau \in \mathbb{R}^s$  tal que  $\tau = (\tau_1, \tau_2, \dots, \tau_s)^T$ . Ahora si  $u = \{u_1, u_2, \dots, u_s\}$  es una base de  $\mathbb{R}^s$  entonces existe una única combinación lineal de  $u$  que resulta en  $\tau$ . Es decir

$$\tau = \sum_{i=1}^s a_i u_i \quad (3.1.2.3)$$

para escalares únicos  $a_1, a_2, \dots, a_s$ .

Por cuestiones de simplicidad supongamos que  $s$  es un número par. A continuación construiremos una base de  $\mathbb{R}^s$  basándonos en las *frecuencias de fourier* definidas por:

$$w_j = \frac{2\pi j}{s}, \quad j = 1, 2, \dots, s. \quad (3.1.2.4)$$

Considera entonces a los vectores  $s$ -dimensionales siguientes.

$$\begin{aligned} e_0 &= (1, 1, \dots, 1)^T \\ c_1 &= (\cos(w_1), \cos(2w_1), \dots, \cos(sw_1))^T \\ s_1 &= (\sin(w_1), \sin(2w_1), \dots, \sin(sw_1))^T \\ &\vdots \\ c_j &= (\cos(w_j), \cos(2w_j), \dots, \cos(sw_j))^T \\ s_j &= (\sin(w_j), \sin(2w_j), \dots, \sin(sw_j))^T \\ &\vdots \\ c_{s/2} &= (\cos(w_{s/2}), \cos(2w_{s/2}), \dots, \cos(sw_{s/2}))^T. \end{aligned}$$

Es fácil notar, realizando el producto interno (dos a dos) de cada vector, y utilizando identidades trigonométricas, que estos son ortogonales y por lo tanto son linealmente independientes. Además el número de vectores que acabamos de definir es  $1 + 2\left(\frac{s}{2} - 1\right) + 1 = s$ , por lo que  $\beta = \{e_0, c_1, s_1, \dots, c_j, s_j, \dots, c_{s/2}\}$  es un conjunto de  $s$  vectores linealmente independientes, lo que significa que es una base<sup>13</sup> de  $\mathbb{R}^s$ . En otras palabras, para todo vector  $\tau$  de  $\mathbb{R}^s$  existen escalares únicos  $a_0, a_1, \dots, a_{s/2}, b_1, b_{(s/2)-1}$  tales que

$$\tau = a_0 e_0 + \sum_{i=1}^{(s/2)-1} (a_i c_i + b_i s_i) + a_{s/2} c_{s/2}.$$

Asumiremos que  $a_0 = 0$  ya que se ha venido suponiendo que el componente cíclico de la serie de tiempo tiene media 0, si la serie de tiempo tuviera alguna tendencia o media distinta de 0, como ya se ha visto anteriormente, este se puede modelar mediante un componente ajeno. Por lo que podemos escribir

$$\tau = \sum_{i=1}^{s/2} (a_i c_i + b_i s_i) \quad \text{con } b_{s/2} = 0.$$

Esto implica que el  $t$ -ésimo componente de  $\tau$  y por lo tanto  $g_t$ <sup>14</sup> (para  $t = 1, 2, \dots, s$ ) se puede escribir como :

$$g_t = \tau_t = \sum_{i=1}^{s/2} a_i \cos(tw_i) + b_i \sin(tw_i).$$

Definamos entonces a la  $j$ -ésima armónica de  $g_t$  de la siguiente manera

$$S_j(t) = a_j \cos(tw_j) + b_j \sin(tw_j),$$

claramente

$$g_t = \sum_{j=1}^{s/2} S_j(t).$$

Para poder usar esto en el contexto de los Modelos Dinámicos Lineales lo único que resta es encontrar la transformación que hay que hacer a  $S_j(t)$  para obtener  $S_j(t+1) = a_j \cos((t+1)w_j) + b_j \sin((t+1)w_j)$ . Para  $j < s/2$  es fácil darse cuenta de que si no se conocen  $a_j$  y  $b_j$  no se puede determinar  $S_j(t+1)$  a partir de  $S_j(t)$ . Sin embargo, si suponemos que se conoce  $S_j(t)$  así como su conjugado armónico:

$$S_j^*(t) = -a_j \sin(tw_j) + b_j \cos(tw_j),$$

entonces se puede calcular  $S_j(t+1)$  y  $S_j^*(t+1)$  en términos de  $S_j(t)$  y de su conjugado de la siguiente manera

<sup>13</sup>Una base es un conjunto de vectores linealmente independientes que generan al espacio.

<sup>14</sup>donde  $g : \{1, 2, \dots\} \rightarrow \mathbb{R}$  es la función de periodo  $s$  asociada al vector  $\tau$ .

$$\begin{aligned}
S_j(t+1) &= a_j \cos((t+1)w_j) + b_j \operatorname{sen}((t+1)w_j) \\
&= a_j \cos(tw_j + w_j) + b_j \operatorname{sen}(tw_j + w_j) \\
&= a_j [\cos(tw_j)\cos(w_j) - \operatorname{sen}(tw_j)\operatorname{sen}(w_j)] + b_j [\cos(tw_j)\operatorname{sen}(w_j) + \operatorname{sen}(tw_j)\cos(w_j)] \dots \text{Nota}^{15} \\
&= (a_j \cos(tw_j) + b_j \operatorname{sen}(tw_j))\cos(w_j) + (-a_j \operatorname{sen}(tw_j) + b_j \cos(tw_j))\operatorname{sen}(w_j) \\
&= S_j(t)\cos(w_j) + S_j^*(t)\operatorname{sen}(w_j),
\end{aligned}$$

y además

$$\begin{aligned}
S_j^*(t+1) &= -a_j \operatorname{sen}((t+1)w_j) + b_j \cos((t+1)w_j) \\
&= -a_j \operatorname{sen}(tw_j + w_j) + b_j \cos(tw_j + w_j) \\
&= -a_j [\cos(tw_j)\operatorname{sen}(w_j) + \operatorname{sen}(tw_j)\cos(w_j)] + b_j [\cos(tw_j)\cos(w_j) - \operatorname{sen}(tw_j)\operatorname{sen}(w_j)] \\
&= (-a_j \operatorname{sen}(tw_j) + b_j \cos(tw_j))\cos(w_j) - (a_j \cos(tw_j) + b_j \operatorname{sen}(tw_j))\operatorname{sen}(w_j) \\
&= S_j^*(t)\cos(w_j) - S_j(t)\operatorname{sen}(w_j).
\end{aligned}$$

Estos resultados los podemos escribir en forma matricial como

$$\begin{pmatrix} S_j(t+1) \\ S_j^*(t+1) \end{pmatrix} = \begin{pmatrix} \cos(w_j) & \operatorname{sen}(w_j) \\ -\operatorname{sen}(w_j) & \cos(w_j) \end{pmatrix} \begin{pmatrix} S_j(t) \\ S_j^*(t) \end{pmatrix}.$$

Definamos entonces a la matriz

$$H_j = \begin{pmatrix} \cos(w_j) & \operatorname{sen}(w_j) \\ -\operatorname{sen}(w_j) & \cos(w_j) \end{pmatrix}.$$

para  $j = 1, 2, \dots, (s/2) - 1$  y notemos que el caso en que  $j = s/2$  es mucho más simple ya que

$$\begin{aligned}
S_{s/2}(t+1) &= a_{s/2} \cos((t+1)\pi) \\
&= -a_{s/2} \cos(t\pi) \\
&= -S_{s/2}(t).
\end{aligned}$$

Por lo que hacemos  $H_{s/2} = (-1)$ .

Ahora considera MDL tal que  $F_t = F = (1, 0, 1, \dots, 0, 1)$ ,

$$G_t = G = \begin{pmatrix} H_1 & 0 & \dots & 0 \\ 0 & H_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \dots & 0 & H_{s/2} \end{pmatrix}$$

---

<sup>15</sup>Esto se puede demostrar usando la fórmula de Fourier, que nos dice que  $e^{ix} = \cos(x) + i\operatorname{sen}(x)$ . Con esto se tiene que  $e^{i(a+b)} = e^{ia}e^{ib} \implies \cos(a+b) + i\operatorname{sen}(a+b) = [\cos(a) + i\operatorname{sen}(a)][\cos(b) + i\operatorname{sen}(b)] = \cos(a)\cos(b) - \operatorname{sen}(a)\operatorname{sen}(b) + i(\cos(a)\operatorname{sen}(b) + \cos(b)\operatorname{sen}(a))$  Lo que implica que  $\cos(a+b) = \cos(a)\cos(b) - \operatorname{sen}(a)\operatorname{sen}(b)$  y  $\operatorname{sen}(a+b) = \cos(a)\operatorname{sen}(b) + \cos(b)\operatorname{sen}(a)$

y  $\theta_t = (S_1(t), S_1^*(t), S_2(t), S_2^*(t), \dots, S_{s/2}(t))^T$ . De esta manera

$$\begin{aligned} Y_t &= F\theta_t + v_t \\ &= \sum_{j=1}^{s/2} S_j(t) + v_t \\ &= g_t + v_t \end{aligned}$$

y además

$$\begin{aligned} \theta_{t+1} &= G\theta_t + w_t \\ &= G(S_1(t), S_1^*(t), S_2(t), S_2^*(t), \dots, S_{s/2}(t))^T + w_t \\ &= (S_1(t+1), S_1^*(t+1), S_2(t+1), S_2^*(t+1), \dots, S_{s/2}(t+1))^T + w_t. \end{aligned}$$

Démonos cuenta de que si  $W_t = 0$  entonces para cualquier  $k \in \mathbb{N}$   $\theta_{t+ks} = \theta_t$  ya que para toda  $j = 1, 2, \dots, s/2$  se tiene que

$$\begin{aligned} S_j(t+ks) &= a_j \cos((t+ks)w_j) + b_j \sen((t+ks)w_j) \\ &= a_j \cos\left((t+ks)\frac{2\pi j}{s}\right) + b_j \sen\left((t+ks)\frac{2\pi j}{s}\right) \\ &= a_j \cos\left(\frac{2\pi t j}{s} + 2\pi k j\right) + b_j \sen\left(\frac{2\pi t j}{s} + 2\pi k j\right) \\ &= a_j \cos\left(\frac{2\pi t j}{s}\right) + b_j \sen\left(\frac{2\pi t j}{s}\right) = S_j(t). \end{aligned}$$

lo que significa que

$$\theta_{t+ks} = \sum_{j=1}^{s/2} S_j(t+ks) = \sum_{j=1}^{s/2} S_j(t) = \theta_t.$$

Así  $Y_{t+ks}$  para toda  $k \in \mathbb{N}$  resulta ser una medida imprecisa de  $g_t$ . En el caso de que  $W_t \neq 0$ , la serie de tiempo ya no será una medida imprecisa de una función de periodo  $s$ , ya que esta función va a cambiar en el tiempo, es decir, con probabilidad 1  $g_t \neq g_{t+ks}$  si  $k \neq 0$ .

También se puede dar el caso en que  $V_t = 0$  y  $W_t = 0$  en donde  $Y_t = g_t$  con probabilidad 1 para toda  $t$ , por lo que no existe ningún elemento estocástico en el componente cíclico definido por este MDL<sup>16</sup>.

Para cualquier caso

---

<sup>16</sup>Como ya se había mencionado anteriormente, para calcular el filtro de Kalman debe de existir la inversa de  $V_t = 0$  por lo que en este escenario, resulta natural sumarle a este modelo otro MDL (puede ser polinómico, de regresión, etc.) tal que su varianza observacional sea distinta de 0.

$$\theta_0 = (a_1, b_1, \dots, a_{\frac{s}{2}-1}, b_{\frac{s}{2}-1}, a_{\frac{s}{2}})^T.$$

Por la construcción que se hizo, resulta claro que haciendo la elección adecuada de  $\theta_0$ , se puede representar cualquier función periódica de periodo  $s$ .

Recordemos que todo este desarrollo se hizo para el caso en que  $s$  es un número par. El caso en que  $s$  es un número impar se puede tratar de una manera similar con la diferencia de que la última armónica, en este caso  $S_{\frac{s-1}{2}}(t)$  no tiene solamente un componente de coseno, si no que también tiene un componente de seno como el resto de se las armónicas, de manera que:

$$\tau_t = \sum_{j=1}^{\frac{s-1}{2}} S_j(t)$$

(donde  $S_j(t) = a_j \cos(tw_j) + b_j \sin(tw_j)$ ).

Para la representación a través de un MDL se definirá para toda  $j = 1, 2, \dots, \frac{s-1}{2}$

$$H_j = \begin{pmatrix} \cos(w_j) & \sin(w_j) \\ -\sin(w_j) & \cos(w_j) \end{pmatrix},$$

y en base a estas matrices la matriz de evolución

$$G_t = G = \begin{pmatrix} H_1 & 0 & \dots & 0 \\ 0 & H_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & H_{\frac{s-1}{2}} \end{pmatrix}.$$

En la matriz anterior, 0 se refiere a la matriz de tamaño  $2 \times 2$  tal que todos sus componentes son iguales a cero.

La matriz de observación será  $F_t = F = (1, 0, 1, 0, \dots, 1, 0)$  y el vector de estados al tiempo  $t$  será  $\theta_t = (S_1(t), S_1^*(t), S_2(t), S_2^*(t), \dots, S_{\frac{s-1}{2}}(t), S_{\frac{s-1}{2}}^*(t))^T$ . Claramente el vector de estados será el descrito anteriormente para toda  $t$  solamente si  $W_t = 0$ .

Muy comunmente se desea modelar un componente cíclico muy suave<sup>17</sup> para esto lo que se hace es descartar un número de armónicas de alta frecuencia y retener solamente unas pocas de las primeras armónicas que son las que oscilan más suavemente. Para lograr esto basta asignar  $a_j = b_j = 0$  para  $q < j$  donde  $q$  es el número de ármonicas que queramos retener. Notemos que si  $q < s/2$  entonces el modelo puede ser especificado por un vector de estados de dimensión  $2 * q$  sin necesidad de guardar a la parte restante del vector (que

---

<sup>17</sup>Con suave se refiere a que no tenga cambios bruscos, es decir que  $g_t$  y  $g_{t+1}$  sea puntos relativamente cercanos

está compuesta de ceros), este número se debe a que necesitamos a cada paso retener tanto a las  $q$  primeras armónicas como a sus respectivos conjugados.

En la librería `d1m` de R existe una función que ajusta modelos con forma de Fourier, la cual se especifica a continuación.

- `d1mModTrig()`: Los parámetros de esta función son los siguientes
  - (i) `s`: El periodo, en caso de ser entero. Si se especifica este parámetro y no se especifica el número de armónicas que se desea retener, la función lo considera igual a la parte entera de  $s/2$ , es decir  $\lfloor \frac{s}{2} \rfloor$ .
  - (ii) `q`: El número de armónicas que se desea retener.
  - (iii) `tau`: El periodo, en caso de no ser entero. Si se especifica este parámetro es necesario especificar `q`.
  - (iv) `om`: Se refiere a la frecuencia. Para este caso también es necesario especificar el número de parámetros que se necesita retener.
  - (v) `dV`: Se refiere a la varianza observacional. De no ser especificada se toma igual a 1.
  - (vi) `dW`: La matriz de covarianza de la ecuación del sistema. Si no la especificamos se considera igual a 0.
  - (vii) `m0`: Se refiere al vector de medias inicial del vector de estados. Si no se especifica por el usuario la función asigna  $m0 = (0, 0, \dots, 0)$
  - (viii) `C0`: Representa a la matriz de covarianza inicial. De no especificarse se toma le asigna la matriz  $diag(10^7, \dots, 10^7)$ .

En este documento solamente trataremos series de tiempo tales que su componente cíclico tiene periodo entero, por lo que al utilizar la función `d1mModTrig()` solamente trabajaremos especificando `s` en vez de `tau` u `om`.

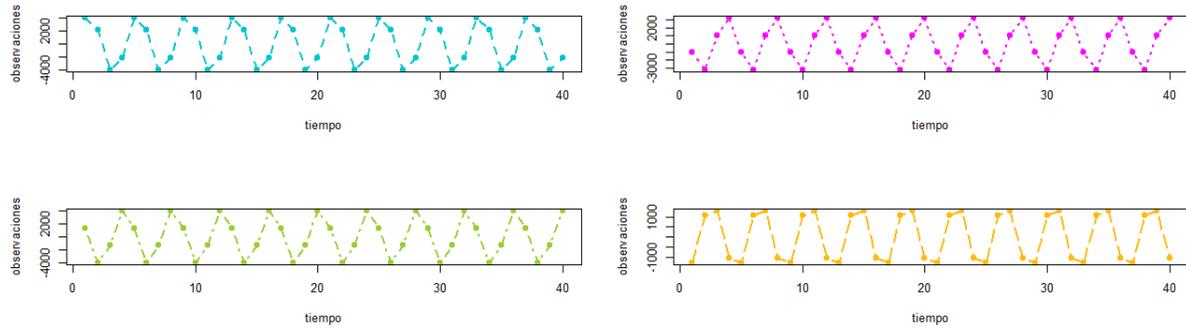
Antes de dar por terminada esta parte de componentes cíclicos, al igual que se hizo para lo modelos de Factores Estacionales y los Modelos Polinómicos, se presentaran unas cuantas simulaciones de algunos procesos con formas de Fourier. Al igual que para los factores estacionales se ha decidido tratar con componentes tales que su periodo es 4 y 12. Esto por dos razones, la primera es para que el lector puede tener una comparación gráfica como se ven los componentes cíclicos de un mismo periodo a través e dos Modelos distintos y la segunda es porque, como ya se ha comentado, en la práctica las series con periodo 4 y 12 se encuentran con mucha frecuencia.

En esta ocasión mantendremos en todos los modelos una matriz de covarianza del sistema idéntica a cero y una varianza de las observaciones muy pequeña. Para el vector de medias inicial y la matriz de covarianza inicial tomaremos las propuestas por el sistema. Jugarémos principalmente con el número de armónicas que retenemos para ilustrar lo que

se platicó acerca como se afecta la suavidad del ciclo en función del numero de armónicas retenidas.

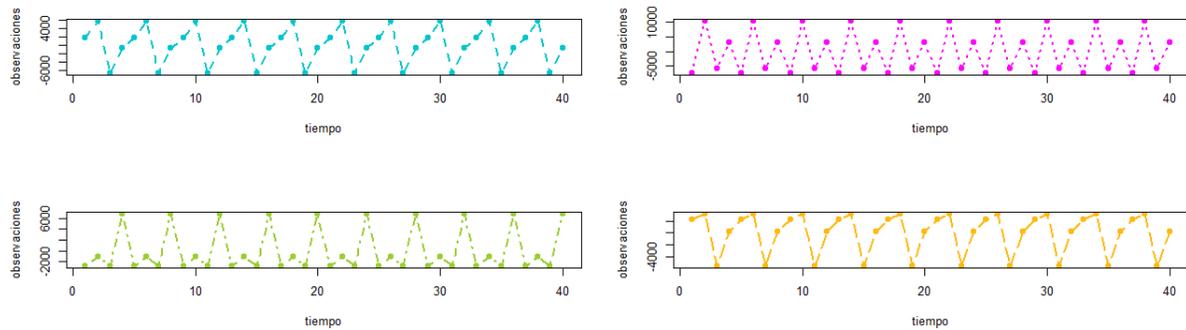
Comencemos entonces con presentar las 4 muestras de un Modelo de periodo 4 en la cual se retiene una armónica

Figura 3.1.2.5: Modelo con Formas de Fourier de periodo 4 reteniendo 1 armónica



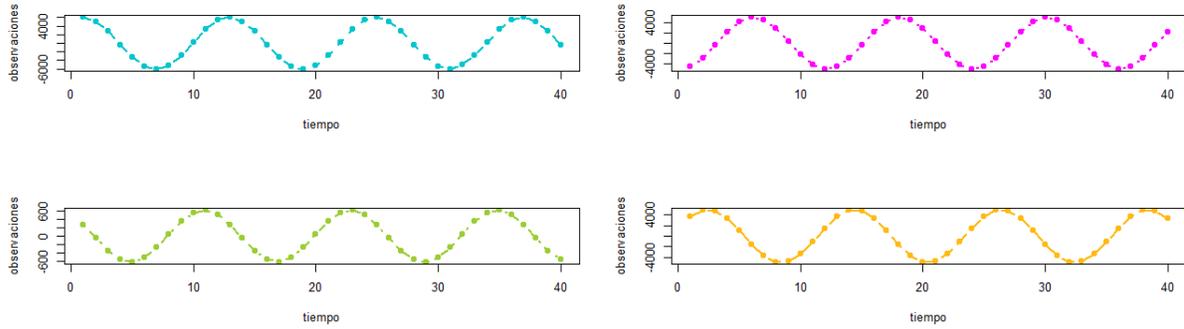
Esta segunda gráfica representa al igual que la anterior 4 muestra de un Modelo con  $s = 4$  con la diferencia de que en esta ocasión se retendrán  $q = 2$  armónicas.

Figura 3.1.2.6: Modelo con Formas de Fourier de periodo 4 reteniendo 2 armónica



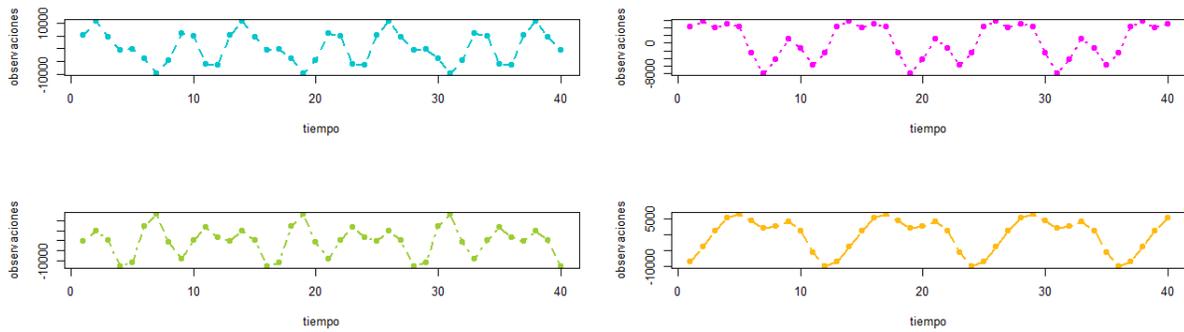
En esta tercer gráfica se representan a las 4 muestras tomadas de un Modelo de periodo 12 que retiene unicamente 1 armónica. Es decir, los ciclos más suaves de periodo 12 los veremos a continuación.

Figura 3.1.2.7: Modelo con Formas de Fourier de periodo 12 reteniendo 1 armónica



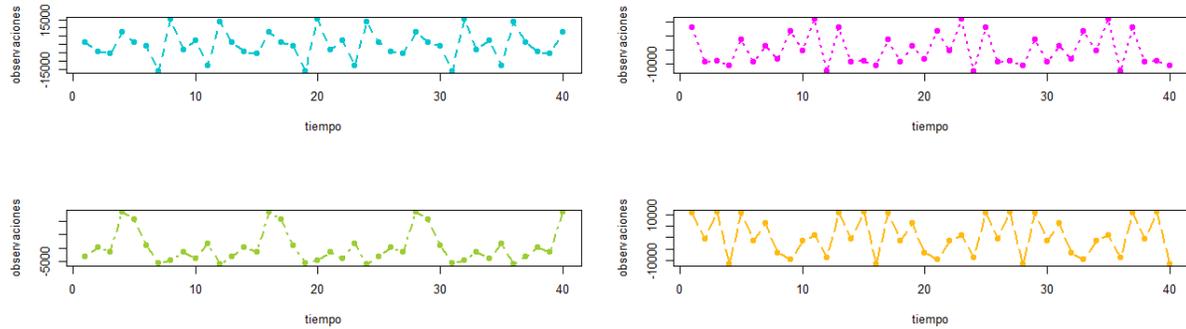
La siguiente gráfica muestra a las muestras de un Modelo con Forma de Fourier tal que  $s = 12$  y  $q = 3$ .

Figura 3.1.2.8: Modelo con Formas de Fourier de periodo 12 reteniendo 3 armónicas



En la quinta y última gráfica encontramos que cada una de las 4 muestra del Modelo con  $s = 12$  y  $q = 6$ , representa a un ciclo de periodo 12 muy poco suave.

Figura 3.1.2.9: Modelo con Formas de Fourier de periodo 12 reteniendo 6 armónica



En todas las gráficas anteriores resulta evidente que cada  $s$  tiempos las observaciones del Modelo se comportan de la misma manera. Además también resulta muy evidente que conforme va incrementando el número de armónicas  $q$  que se retiene, la función se va haciendo más dura y abrupta.

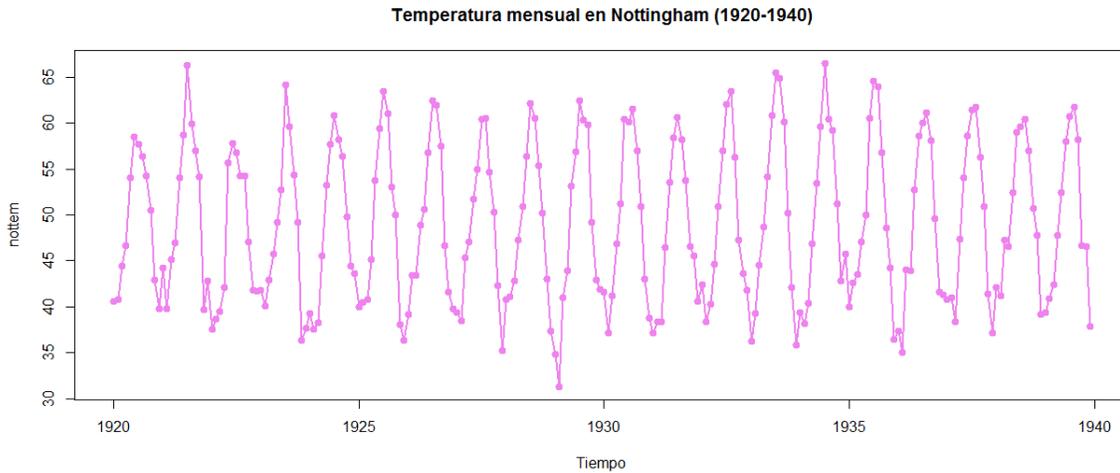
### 3.1.3. La serie `nottem`

Hasta ahora hemos revisado dos tipos de modelos distintos que describen el componente cíclico de una serie de tiempo y también hemos revisado a los modelos polinómicos de  $n$ -ésimo orden que describen el nivel o alguna tendencia que pueda tener una serie de tiempo. Como se mencionó al principio de este capítulo, los MDL tiene una propiedad aditiva que permite pensar a la serie de tiempo como la suma de dos o más series independientes, y describir a cada componenete de la serie mediante un MDL distinto. Una vez descritos todos los componentes de la serie, se puede contruir un MDL para la serie original a partir de los MDLs propuestos para cada uno de sus componentes.

Se ha decidido analizar a la serie de tiempo que representa la temperatura mensual en Nottingham para los años 1920 – 1940 y se encuentra en R como `nottem`. Esta elección se debe a que permite ilustrar la propiedad aditiva de los modelos y comparar la manera de ajustar ciclos mediante un modelo de factores estacionales y un modelo con formas de Fourier.

La serie `nottem` se muestra a continuación

Figura 3.1.3.1: Observaciones de la serie `nottem`



Es evidente que esta serie tiene un componente cíclico con periodo  $s = 12$ , y no presenta tendencia a la alta ni a la baja, solamente un nivel local que claramente es distinto de 0. Por esta razón se ha decidido ajustar un MDL que conste de dos componentes. El primer componente será un Modelo de nivel local y el segundo será un componente cíclico.

Propongamos entonces dos modelos para esta serie de tiempo. El primero tal que el componente cíclico sea un modelo con formas de Fourier y el segundo tal que su componente cíclico sea un modelo de factores estacionales.

### Modelo 1

Primero especifiquemos el componente que describe el nivel local. Para este asignaremos  $\mu_0 = 50$  ya que la media de la serie `nottem` está entre 49 y 50 y elegiremos una varianza inicial algo grande  $C_0 = 100$ . También a este componente asignaremos una varianza de estados igual a cero, es decir  $W_t = 0$  y una varianza observacional  $V_t = 0.1$  para toda  $t$ .

Para este modelo el componente cíclico será un modelo con formas de Fourier tal que  $s = 12$ , el número de armónicas que se retendrá será 6, la varianza observacional  $V_t = 5$  y la matriz de covarianza del sistema será  $W_t = \text{diag}(0, \dots, 0)$ . Además el vector de medias inicial y la varianza inicial serán las propuesta por la función `dlnModTrig`.

### Modelo 2

Para esta segunda propuesta el componente que describe el nivel local será igual que para el modelo 1 con la diferencia de que  $V_t = 5$  para toda  $t$ .

Por su parte el componente cíclico será un modelo de factores estacionales de periodo  $s = 12$ . Las matrices de covarianza del sistema y observacionales serán las propuestas por la función `dlmModSeas()`,  $V_t = 1$  y  $W_t = \text{diag}(1, 0, \dots, 0)$ . El vector de medias inicial será  $\mu_0 = (y_{12}, y_{11}, \dots, y_2)$  donde  $y_i$  se refiere a la  $i$ -ésima observación de la serie `nottem`, y la matriz de covarianza inicial se tomará como  $C_0 = \text{diag}(100, \dots, 100)$ .

En R estos dos modelos se especifica mediante el código de abajo.

---

### Código en R

---

```
#Especificación Modelo 1
mod1t <- dlmModTrig(s = 12, dV = 5, dW = diag(rep(0,11)))
mod1p <- dlmModPoly(order = 1, dV = 0.1, dW = 0, m0 = 50, C0 = 100)
mod1 <- mod1t + mod1p

#Especificación Modelo 2
m02 <- numeric(11)
for(i in 1:11)
{
  m02[i] <- nottem[13-i] - 50
}
mod2s <- dlmModSeas(frequency = 12, m0 = m02, C0 = diag(rep(100,11)))
mod2p <- dlmModPoly(order = 1, dV = 5, dW = 0, m0 = 50, C0 = 100)
mod2 <- mod2s + mod2p
```

---

Una vez especificados ambos modelos, calculemos el filtro de cada uno y presentemos un par de gráficas comparativas donde se presentan la serie `nottem` junto con la predicción de un paso adelante de las observaciones  $f_t = \mathbb{E}[Y_t | y_{1:t-1}]$  según cada uno de los modelos.

---

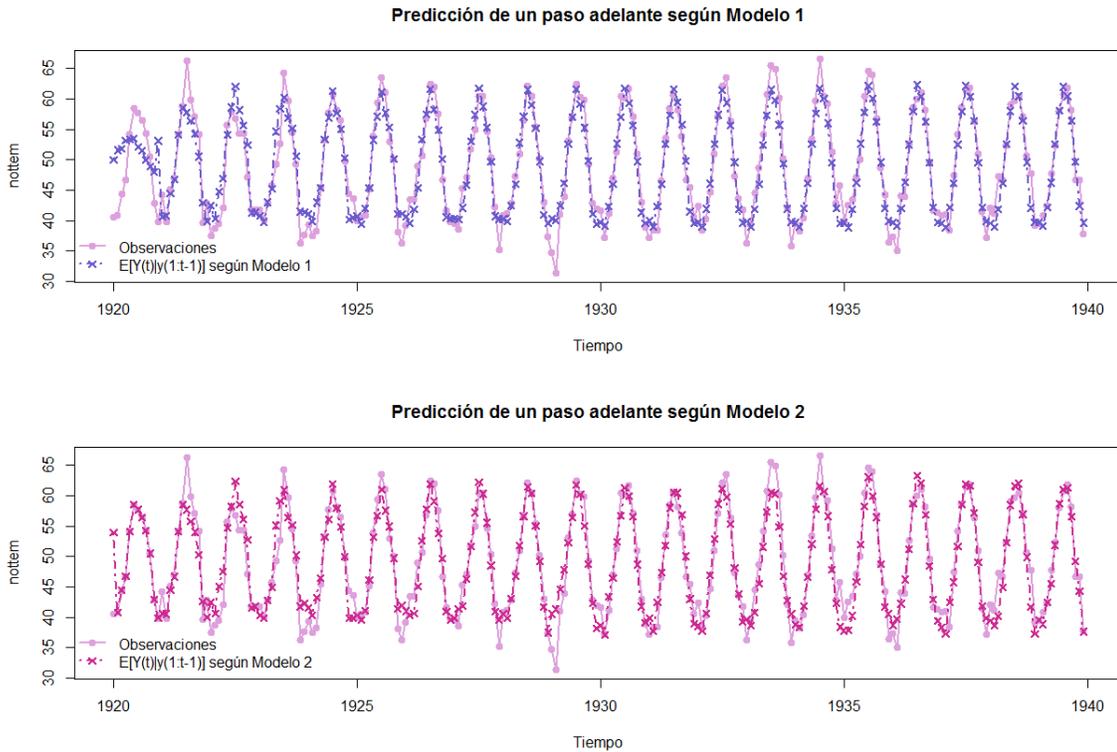
### Código en R

---

```
#Cálculo del filtro para ambos modelos
mod1Filt <- dlmFilter(nottem, mod1)
mod2Filt <- dlmFilter(nottem, mod2)
#Gráfica de las predicciones de un paso adelante
par(mfrow=c(2,1))
plot(nottem, type = 'o', lty = 1, pch = 19, lwd = 2, col = "violet")
lines(mod2Filt$f, type = 'o', pch = 4, lty = 4, lwd = 2, col = "violetred")
plot(nottem, type = 'o', lty = 1, pch = 19, lwd = 2, col = "violet")
lines(mod1Filt$f, type = 'o', pch = 4, lty = 4, lwd = 2, col = "slateblue")
```

---

Figura 3.1.3.2: Predicción de un paso adelante según cada modelo



Podemos ver a simple vista que ambos modelos hacen predicciones bastante buenas, podría hacerse una comparación entre ambos más formal usando los criterios MAD, MSE y MAPE, y además realizar la revisión del modelo para ver que tan factible es cada uno. Por ahora, supongamos que ambos modelos se ajustan suficientemente bien y concentremos en estimar y analizar el vector de estados para cada uno de los modelos con el fin de lograr un conocimiento más profundo de los modelos cíclicos.

Comencemos por concentrarnos en el Modelo 1, el componente cíclico de este es un modelo con formas de Fourier con periodo 12 y en el cual se ha decidido retener todas las armónicas. Para este modelo nos interesa analizar el comportamiento esperado de cada una de las armónicas dada la información disponible. Para esto, recordemos que el vector de estados al tiempo  $t$  es

$$\theta_t = \begin{pmatrix} S_1(t) \\ S_1^*(t) \\ S_2(t) \\ S_2^*(t) \\ \vdots \\ S_6(t) \end{pmatrix}.$$

donde  $S_i(t)$  es la  $i$ -ésima armónica del modelo al tiempo  $t$ .

De esta manera, la media de suavización al tiempo  $t$  y por lo tanto la estimación del vector de estados a dicho tiempo será

$$s_t = \mathbb{E}[\theta_t | y_{1:n}] = \begin{pmatrix} \mathbb{E}[S_1(t) | y_{1:n}] \\ \mathbb{E}[S_1^*(t) | y_{1:n}] \\ \mathbb{E}[S_2(t) | y_{1:n}] \\ \mathbb{E}[S_2^*(t) | y_{1:n}] \\ \vdots \\ \mathbb{E}[S_6(t) | y_{1:n}] \end{pmatrix}.$$

Es claro entonces que la estimación de  $S_j(t)$  será la  $(1 + 2(j - 1))$ -ésima entrada de  $s_t$ .

En R podemos calcular el vector de medias de  $\theta_t | y_{1:n}$  para  $t = 0, 1, \dots, n$  y extraer del vector  $s_t$  la estimación de la  $j$ -ésima armónica  $S_j(t)$  para  $j = 1, 2, \dots, 6$  de la siguiente manera:

---

### Código en R

---

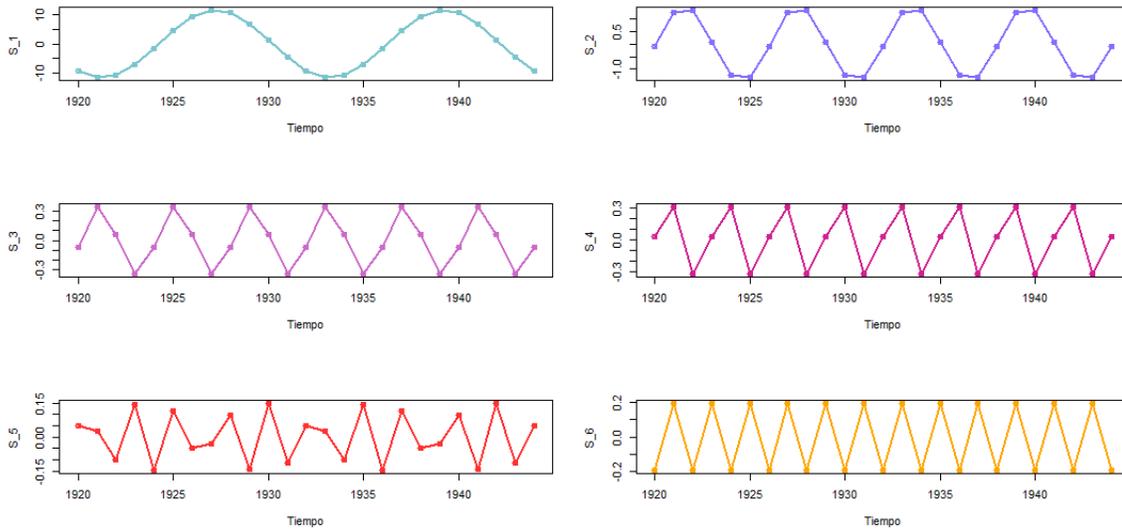
```
#Cálculo del algoritmo recursivo de filtración
mod1Smooth <- dlmSmooth(mod1Filt)

#Estimación de las armónicas
S_1 <- ts(mod1Smooth$s[1:25,1], start = start(nottem-1))
S_2 <- ts(mod1Smooth$s[1:25,3], start = start(nottem-1))
S_3 <- ts(mod1Smooth$s[1:25,5], start = start(nottem-1))
S_4 <- ts(mod1Smooth$s[1:25,7], start = start(nottem-1))
S_5 <- ts(mod1Smooth$s[1:25,9], start = start(nottem-1))
S_6 <- ts(mod1Smooth$s[1:25,11], start = start(nottem-1))

#Grafica de la estimación de las armónicas
par(mfrow=c(3,2))
plot(S_1, type = 'o', xlab = "Tiempo", col = "cadetblue3", lwd = 2)
plot(S_2, type = 'o', xlab = "Tiempo", col = "slateblue1", lwd = 2)
plot(S_3, type = 'o', xlab = "Tiempo", col = "orchid3", lwd = 2)
plot(S_4, type = 'o', xlab = "Tiempo", col = "violetred", lwd = 2)
plot(S_5, type = 'o', xlab = "Tiempo", col = "firebrick1", lwd = 2)
plot(S_6, type = 'o', xlab = "Tiempo", col = "orange", lwd = 2)
```

---

Figura 3.1.3.3: Armónicas del Modelo 1 y la serie `nottem`



En este caso como  $W_t = \text{diag}(0, \dots, 0)$ , entonces  $S_j(t + 12) = S_j(t)$  para toda  $j = 1, 2, \dots, 6$  y  $t \geq 1$ , lo que implica que  $\mathbb{E}[S_j(t + 12)|y_{1:n}] = \mathbb{E}[S_j(t)|y_{1:n}]$ . Por lo que la estimación de la  $j$ -ésima armónica será la misma cada doce tiempos, es por esta razón que en las gráficas anteriores solamente se muestra dicha estimación para los primeros 25 tiempos.

Notemos que en efecto, como se había mencionado con anterioridad, las primeras armónicas son funciones (del tiempo) más suaves y las últimas armónicas son funciones más duras y abruptas.

Así ya hemos logrado encontrar el comportamiento esperado de cada una de las armónicas del modelo tomando en cuenta toda la información que las observaciones proporcionan de estas.

Falta estimar el nivel local, llamémoslo  $m_t$ , como  $W_t = 0$  también para el componente polinómico de orden 1, la estimación de  $m_t$  va a ser igual para todos los tiempos, así que para conocerla explícitamente simplemente escribimos en R.

```
print(mod1Smooth$s[1,12])
```

Consideremos ahora el Modelo 2, y estimemos también para cada tiempo el vector de estados. Recordemos que para este modelo el vector de estados al tiempo  $t$  es tal que su primera entrada se refiere al factor estacional  $\tau_t$  del cual  $Y_t - m_t$  (donde  $m_t$  es el nivel local) es una medición imprecisa, además  $\theta_t$  es tal que guarda la información necesaria para que al siguiente tiempo se pueda especificar  $\tau_{t+1}$  y guardar en la primera entrada de  $\theta_{t+1}$  lo que nos interesa aquí es estimar estos factores estacionales. Para lograr esto estimaremos primero el vector de estados como  $\mathbb{E}[\theta_t|y_{1:n}]$  y extraeremos la información que se encuentra

en la primera entrada  $\mathbb{E}[\tau_t|y_{1:n}]$  para estimar  $\tau_t$ .

Recordemos que en este modelo la matriz de covarianza del sistema la definimos como  $W_t = \text{diag}(1, 0, \dots, 0)$ , que es la propuesta por la función `d1mModSeas()` en caso de no ser especificada. Esto significa que a pesar de que el periodo es  $s = 12$ , casi seguramente  $\tau_{t+12} \neq \tau_t$  para toda  $t \geq 1$ , por lo que, a diferencia del modelo pasado, cada 12 tiempos el valor estimado de la función cíclica de periodo  $s$  de la cual ( $Y_t - m_t$  es una medición imprecisa) será distinto.

Por esta razón se puede pensar en los procesos generados por cada uno de los 12 factores estacionales  $\{\tau^{(i)}_n\}_{n \geq 1}$  tales que  $\tau^{(i)}_n = \tau_{i+(n-1)*12}$  para toda  $i = 1, 2, \dots, 12$ , y analizar el comportamiento de cada proceso. Esto es, estudiar el comportamiento en el tiempo de cada uno de los factores estacionales que definen al modelo.

A continuación se presenta el código en R que estima el vector de estados para cada tiempo y por lo tanto también estima el comportamiento de cada uno de los factores estacionales. Esto es, calcula el valor esperado de  $\tau^{(i)}_n|y_{1:t} = \tau_{i+(n-1)*12}|y_{1:t}$  para toda  $i = 1, 2, \dots, 12$  para los tiempos  $i + (n - 1) * 12$  que sean menores a  $t$  y donde  $y_t$  se refiere a la última observación disponible de la serie.

---

### Código en R

---

```
#Cálculo del algoritmo recursivo de filtración
mod2Smooth <- d1mSmooth(mod2Filt)

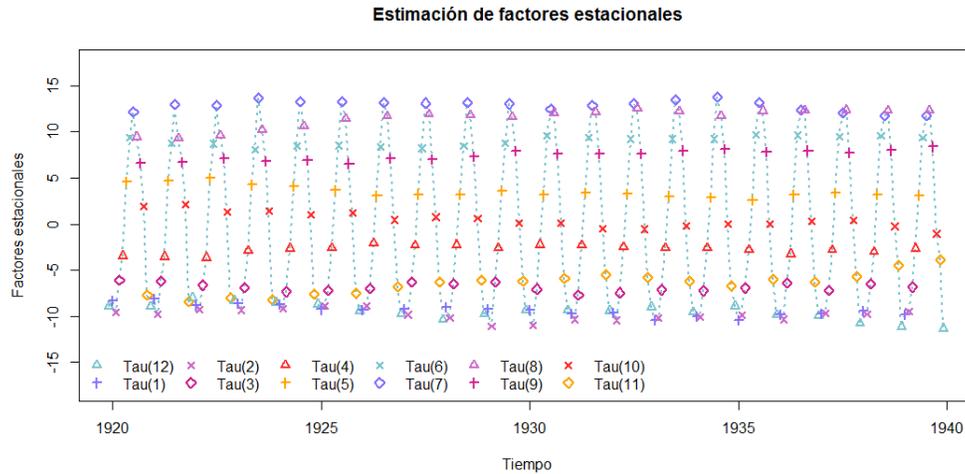
#Grafica de la estimación de los factores estacionales
plot(nottem - mod2Smooth$s[1,12], type = 'o', lty = 1, pch = 19,
     lwd = 1, col = "white", xlab = "Tiempo", ylab = "Factores estacionales",
     main = "Estimación de factores estacionales")
lines(mod2Smooth$s[,1], type = 'o',
      pch = c(2,3,4,5), lty = 3, lwd = 2, col = c("cadetblue3","slateblue1",
      "orchid3","violetred","firebrick1","orange"))
leg <- c("Tau(12)", "Tau(1)", "Tau(2)", "Tau(3)", "Tau(4)", "Tau(5)",
      "Tau(6)", "Tau(7)", "Tau(8)", "Tau(9)", "Tau(10)", "Tau(11)")
legend("bottomleft", legend = leg,
      col = c("cadetblue3","slateblue1","orchid3",
      "violetred","firebrick1","orange","cadetblue3","slateblue1","orchid3",
      "violetred","firebrick1","orange"),
      pch = c(2,3,4,5,2,3,4,5,2,3,4,5), ncol = 6, lwd = c(rep(2,12)),
      lty = NA, bty = "n")
```

---

De esta manera se genera la siguiente gráfica, donde se muestra el ciclo estimado definido por los factores estacionales. A su vez se usaron elementos distintos para representar a

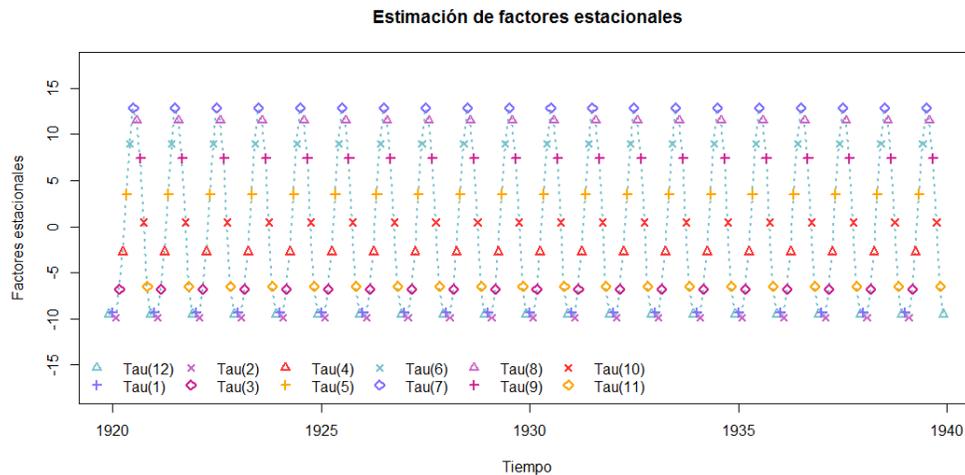
cada uno de los factores estacionales y se pueda apreciar su comportamiento en el tiempo. Cabe aclarar que  $\text{Tau}(j)$  en la gráfica se refiere a  $\{\tau^{(j)}_n\}_{n \geq 1}$ .

Figura 3.1.3.4: Estimación de factores estacionales del Modelo 2 y la serie `nottem`



En la gráfica es muy claro como la estimación de los factores estacionales cambia en el tiempo y por lo tanto también cambia la función cíclica (como función del tiempo) definida por estos. Con el fin de hacer una pequeña comparación, se ha corrido el último código presentado pero fijando  $W_t = \text{diag}(0, \dots, 0)$  para el componente cíclico para ver que sucede con este modelo cuando no se permite variación aleatoria al vector de estados. La gráfica generada es la que sigue

Figura 3.1.3.5: Estimación de factores estacionales del Modelo 2 (con matriz de covarianza observacional igual a 0) y la serie `nottem`



Notemos que aquí la estimación de cada uno de los factores estacionales es la misma en el tiempo, esto es porque como se fijó  $W_t = \text{diag}(0, \dots, 0)$  para el componente ciclico, entonces  $\tau_t = \tau_{t+12}$  para toda  $t$  y por lo tanto el estimador de  $\tau_t$ ,  $\mathbb{E}[\tau_t|y_{1:t}] = \mathbb{E}[\tau_{t+12}|y_{1:t}]$ , es precisamente el estimador de  $\tau_{t+12}$ .

### 3.1.4. Modelos Auto-regresivos y de Promedios Móviles (ARMA) como Modelos Dinámicos Lineales

Entre los modelos más ampliamente usados para el análisis de series de tiempo encontramos a la clase de los modelos Autoregresivos y de Promedios Móviles (ARMA). Estos están definidos por la relación

$$Y_t = \mu + \sum_{j=1}^p \phi_j(Y_{t-j} - \mu) + \sum_{j=1}^q \psi_j \varepsilon_{t-j} + \varepsilon_t$$

donde  $p$  y  $q$  son enteros no negativos,  $\{\varepsilon_t\}_{t \geq 1}$  es un ruido blanco<sup>18</sup> con varianza  $\sigma_\varepsilon^2$  y los parámetros  $\phi_1, \phi_2, \dots, \phi_p, \psi_1, \psi_2, \dots, \psi_q$  satisfacen condiciones necesarias para que  $\{Y_t\}$  sea un proceso (débilmente) estacionario<sup>19</sup>.

Se dice que este es un modelo de ARMA( $p, q$ ). En el caso en que  $q = 0$  el modelo simplemente se llama Modelo AR( $p$ ) (Autoregresivo de orden  $p$ ) y queda como

$$Y_t = \mu + \sum_{j=1}^p \phi_j(Y_{t-j} - \mu) + \varepsilon_t.$$

Por el otro lado si  $p = 0$  el modelo se llama, como era de esperarse, Modelo MA( $q$ ) (de Promedios Móviles de orden  $q$ ) y está descrito por la siguiente ecuación:

$$Y_t = \sum_{j=1}^q \psi_j \varepsilon_{t-j} + \varepsilon_t.$$

Para que sea posible ajustar alguno de estos modelos a una serie de tiempo es necesario que esta sea estacionaria. Para series no estacionarias, o al menos para una gran parte de ellas, se puede ajustar un modelo llamado Modelo Autoregresivo Integrado de Promedio Móviles (ARIMA) el cual se caracteriza por que la serie formada por las  $d$ -ésimas diferencias de la serie original  $\Delta^d Y_t = Y_t - Y_{t-d}$  se comporta como un modelo ARMA( $p, q$ ) y en dado caso se escribe que es un modelo ARIMA( $p, d, q$ ).

La breve discusión anterior puede referirse a series de tiempo univariadas o multivariadas, para lo que sigue vamos a considerar únicamente el caso en que  $Y_t$  es un escalar para

<sup>18</sup>Un ruido blanco es un proceso estocástico  $\{X_t\}_{t \geq 0}$  tal que para toda  $t$   $\mathbb{E}[X_t] = 0$  y  $\text{Var}(X_t) = \sigma^2$  y además  $X_r$  es independiente de  $X_s$  para toda  $r \neq s$ .

<sup>19</sup>Decimos que un proceso estocástico  $\{X_t\}$  es (débilmente) estacionario si el vector aleatorio  $(X_1, \dots, X_n)^T$  tiene el mismo vector de medias y misma matriz de covarianza que el vector  $(X_{1+h}, \dots, X_{n+h})^T$  para todo entero  $h$  y entero positivo  $n$ .

toda  $t$ .

Cada uno de estos modelos se puede representar como MDL si hacemos la suposición adicional que la distribución de los errores  $\{\varepsilon_t\}$  es Gaussiana. Es decir, dado un proceso ARIMA(p,d,q) con errores gaussianos se puede encontrar un MDL cuyo proceso observable sea igual en distribución que el proceso mencionado. Para fines de este documento solamente discutiremos como representar a los modelos ARMA, ya que los MDLs nos permiten describir distintos rasgos de la serie mediante distintos componentes, por lo que si existe alguna tendencia o ciclicidad que hicieran la serie no estacionaria, se pueden modelar aparte y sumarlas a algún componente ARMA(p,q).

En la práctica el enfoque más utilizado es precisamente el descrito en el párrafo anterior, sin embargo no está de más conocer la representación de un modelo ARIMA a través de un MDL, si se desea hacer esto, se invita al lector a revisar el libro [1] de la bibliografía.

Consideremos entonces el caso estacionario ARMA(p,q). Supongamos por simplicidad que  $\mu = 0$ , es claro que, bajo el contexto de los MDLs, siempre se puede hacer esta suposición ya que una media distinta de cero o que cambia en el tiempo se puede modelar mediante un componente distinto al componente autoregresivo de promedio móviles.

En primer lugar consideremos un Modelo ARMA(p,0) ó AR(p). La ecuación mediante la cual está descrita este modelo, bajo la suposición  $\mu = 0$  y suponiendo que los errores tienen distribución gaussiana, es:

$$Y_t = \sum_{i=1}^p \phi_i Y_{t-i} + \varepsilon_t \quad \varepsilon_t \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2).$$

Sean

$$\theta_t = \begin{pmatrix} \theta_{1t} \\ \theta_{2t} \\ \vdots \\ \theta_{pt} \end{pmatrix}, \quad G = \begin{pmatrix} \phi_1 & 1 & 0 & \cdots & 0 \\ \phi_2 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \\ \phi_{p-1} & 0 & \cdots & 0 & 1 \\ \phi_p & 0 & \cdots & 0 & 0 \end{pmatrix},$$

$$W = \text{diag}(\sigma_\varepsilon^2, 0, \dots, 0) \quad F = (1, 0, \dots, 0), \quad y \quad V = 0.$$

De tal manera que

$$\theta_{t+1} = \begin{pmatrix} \theta_{1t} \\ \theta_{2t} \\ \vdots \\ \theta_{pt} \end{pmatrix} \begin{pmatrix} \phi_1 & 1 & 0 & \cdots & 0 \\ \phi_2 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \\ \phi_p & 0 & \cdots & 0 & 1 \\ \phi_{p-1} & 0 & \cdots & 0 & 0 \end{pmatrix} + \begin{pmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

ó análogamente

$$\begin{aligned}
\theta_{1t} &= \phi_1\theta_{1t-1} + \theta_{2t-1} + \varepsilon_t \\
\theta_{2t} &= \phi_2\theta_{1t-1} + \theta_{3t-1} \\
&\vdots \\
\theta_{p-1t} &= \phi_{p-1}\theta_{1t-1} + \theta_{pt-1} \\
\theta_{pt} &= \phi_p\theta_{1t-1},
\end{aligned}$$

sustituyendo  $\theta_{2t-1} = \phi_2\theta_{1t-2} + \theta_{3t-2}$  en la primera ecuación se tiene que

$$\theta_{1t} = \phi_1\theta_{1t-1} + \phi_2\theta_{1t-2} + \theta_{3t-2} + \varepsilon_t,$$

a su vez notemos que  $\theta_{3t-2} = \phi_3\theta_{1t-3} + \theta_{4t-3}$  por lo que reescribimos la ecuación anterior como

$$\theta_{1t} = \phi_1\theta_{1t-1} + \phi_2\theta_{1t-2} + \phi_3\theta_{1t-3} + \theta_{4t-3} + \varepsilon_t,$$

y continuando con la sustitución en la ecuación  $\theta_{j_{t-(j-1)}}$  por  $\phi_j\theta_{1t-j} + \theta_{j+1t-j}$  para  $j = 4, \dots, p$  (donde claramente  $\theta_{p+1t-p} = 0$ ) se obtiene

$$\theta_{1t} = \sum_{j=1}^p \phi_j\theta_{1t-j} + \varepsilon_t.$$

Ahora, por otro lado tenemos que

$$\begin{aligned}
Y_t &= F\theta_t = \theta_{1t} \\
&= \sum_{j=1}^p \phi_j\theta_{1t-j} + \varepsilon_t \\
&= \sum_{j=1}^p \phi_j Y_{t-j} + \varepsilon_t.
\end{aligned}$$

Con esto se verifica que el proceso observable  $\{Y_t\}$  del MDL previamente definido tiene la misma distribución que un proceso AR(p).

Notemos que el vector de estados conjuntamente con la matriz  $G$  se encargan de ir guardando la información de los  $p$  tiempos pasados de forma que al tiempo actual se pueda incorporar esta información y usarla a través del primer componente del vector de estados.

Para el caso general de un ARMA(p,q) hay que hacer un cambio en el MDL representante de un AR(p), de tal forma que este nuevo modelo pueda ir guardando los errores de

tiempos pasados e incorporarlos a la observación actual de la manera que lo hace un Modelo de Promedio Móviles. Para esto, lo primero que hacemos es definir  $r = \max\{p, q + 1\}$  y rescribir al modelo ARMA como

$$Y_t = \sum_{j=1}^r \phi_j(Y_{t-j}) + \sum_{j=1}^{r-1} \psi_j \varepsilon_{t-j} + \varepsilon_t,$$

donde evidentemente  $\phi_j = 0$  si  $j > p$  y  $\psi_j = 0$  si  $j > q$ . Esto se hace para encontrar una dimensión adecuada de las matrices  $G$  y  $R$  definidas a continuación. Definamos, entonces, a las matrices y vectores involucradas en el modelo

$$\theta_t = \begin{pmatrix} \theta_{1t} \\ \theta_{2t} \\ \vdots \\ \theta_{rt} \end{pmatrix}, \quad G = \begin{pmatrix} \phi_1 & 1 & 0 & \cdots & 0 \\ \phi_2 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \\ \phi_{r-1} & 0 & \cdots & 0 & 1 \\ \phi_r & 0 & \cdots & 0 & 0 \end{pmatrix},$$

$$R = \begin{pmatrix} 1 \\ \psi_1 \\ \vdots \\ \psi_{r-1} \end{pmatrix}, \quad W = RR^T \sigma_\varepsilon^2, \quad F = (1, 0, \dots, 0), \quad V = 0.$$

De esta manera el MDL queda descrito por

$$Y_t = F_t \theta_t$$

$$\theta_t = G \theta_{t-1} + R \varepsilon_t.$$

Realizando los productos de  $G \theta_{t-1}$  y  $R \varepsilon_t$  la ecuación del sistema queda de la siguiente forma

$$\begin{aligned} \theta_{1t} &= \phi_1 \theta_{1t-1} + \theta_{2t-1} + \varepsilon_t \\ \theta_{2t} &= \phi_2 \theta_{1t-1} + \theta_{3t-1} + \psi_1 \varepsilon_t \\ &\vdots \\ \theta_{r-1t} &= \phi_{r-1} \theta_{1t-1} + \theta_{rt-1} + \psi_{r-2} \varepsilon_t \\ \theta_{rt} &= \phi_r \theta_{1t-1} + \psi_{r-1} \varepsilon_t. \end{aligned}$$

En la primera ecuación podemos ir sustituyendo consecutivamente  $\theta_{j_{t-(j-1)}}$ , para  $j = 2, 3, \dots, r$ , de la misma forma que como se hizo para el modelo AR(p), de donde se obtiene

$$\theta_{1t} = \sum_{j=1}^r \phi_j \theta_{1t-j} + \sum_{j=1}^{r-1} \psi_j \varepsilon_{t-j} + \varepsilon_t,$$

y nuevamente como  $Y_t = F\theta_t = \theta_{1t}$  es evidente que  $\{Y_t\}$  es igual en distribución que un proceso ARMA(p,q).

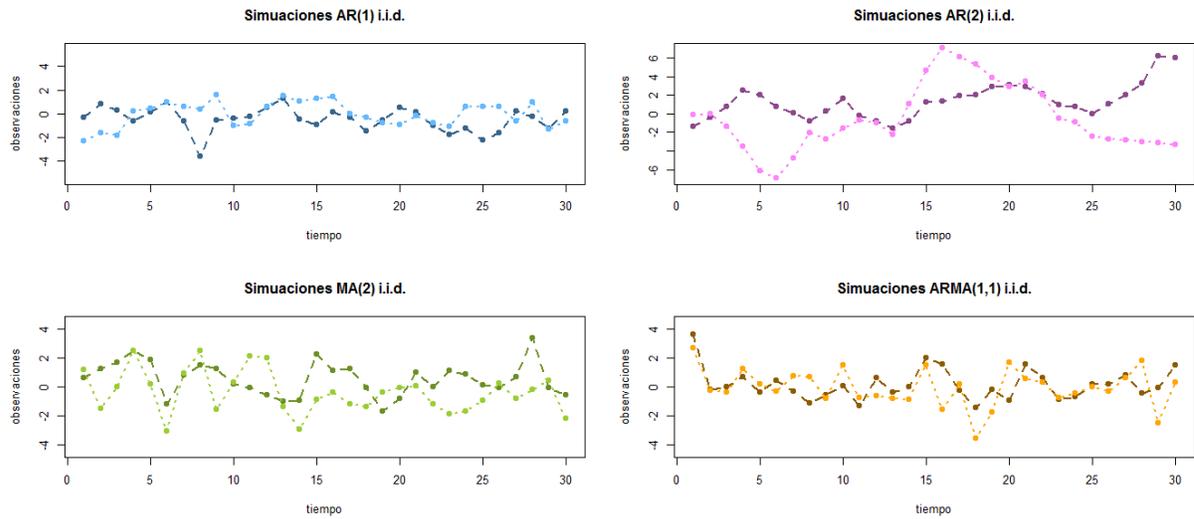
Para el caso particular de un MA(q) lo único que se hace es sustituir el todo el procedimiento anterior  $p$  por 0 lo que obviamente implica que  $\phi_j = 0$  para toda  $j$ .

Una vez más, la librería `d1m` de R nos proporciona una función a través de la cual es posible especificar cualquier Modelo ARMA en su representación como MDL. Se presenta una breve descripción de la misma a continuación.

- `d1mModARMA()`: Esta función retorna un objeto de clase `d1m` tras recibir como parámetros
  - (i) `ar`: Un vector que contiene los coeficientes de autoregresión. En caso de no ser especificado se considera un vector nulo.
  - (ii) `ma`: Se refiere a un vector que contiene los coeficientes de promedios móviles. Al igual que `ar`, en caso de no ser especificado se considera un vector nulo.
  - (iii) `sigma2`: La varianza de los errores  $\varepsilon_t$  del modelo ARMA.
  - (iv) `dV`: La varianza de la ecuación de observaciones  $V_t$ , de no ser especificada se considera igual a 0.
  - (v) `m0`: El vector de medias de la distribución inicial del vector de estados. Si no se especifica la función considera  $\mu_0 = (0, \dots, 0)$ .
  - (vi) `C0`: La matriz de covarianza inicial del vector de estados. En caso de no especificarse se toma  $C_0 = I$ .

Meramente con fines ilustrativos, a continuación se presenta una gráfica en la que se dibujan dos muestras de algunos de los modelos ARMA más utilizados en la práctica, estos son AR(1), AR(2), MA(2) y ARMA(1,1).

Figura 3.1.4.1: Simulaciones de Modelos Autoregresivos de Promedios Móviles



**Nota:** Para el modelo AR(1) se ha tomado  $\phi_1 = 0.5$ . Para el modelo AR(2) se consideraron  $\phi_1 = 1.5$  y  $\phi_2 = -0.56$ . Para el tercer modelo MA(2) se tomó  $\psi_1 = 0.8$  y  $\psi_2 = -0.5$ . Y para el cuarto y último modelo propuesto, ARMA(1,1), se consideró  $\phi_1 = 0.4$  y  $\psi_1 = -0.8$ . El resto de los parámetros se consideraron los que sugiere la librería a través de la función MDL.

### 3.1.5. Modelos de Regresión

Es de interés en muchas aplicaciones incluir el efecto de variables explicativas en una serie de tiempo  $\{Y_t\}$ . Por ejemplo, analizar alguna repuesta clínica  $Y_t$  a diferentes dosis de drogas  $x_t$  en el tiempo, ó el tamaño de la población de algún tipo de bacterias  $Y_t$  al ser cometida a distintas temperaturas en un lugar fijo.

El modelo de regresión lineal estándar sobre  $x$  es definido como

$$Y_t = \alpha_0 + \alpha_1 x_t + \varepsilon_t, \quad \varepsilon_t \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2).$$

Comúnmente la suposición de que los errores  $\varepsilon_t$  son independientes e idénticamente distribuidos no es realista. Una posible solución es suponer algún tipo de dependencia entre los errores como por ejemplo que  $\{\varepsilon_t\}$  es un modelo Autoregresivo. Otra solución es pensar que la relación entre  $x$  y  $Y$  cambia en el tiempo, es decir plantear un modelo de la forma

$$Y_t = \alpha_{0t} + \alpha_{1t} x_t + \varepsilon_t, \quad \varepsilon_t \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2),$$

modelando la evolución temporal como  $\alpha_{it} = \alpha_{i,t-1} + w_{it}$  para  $i = 0, 1$ , donde  $w_{0t}$  y  $w_{1t}$  son v.a. Gaussianas independientes.

Es relativamente sencillo e intuitivo plantear a este modelo como MDL. Basta definir

$$\theta_t = \begin{pmatrix} \alpha_{0t} \\ \alpha_{1t} \end{pmatrix}, \quad F = (1, x_t), \quad V_t = \sigma_\varepsilon^2$$

de manera que

$$\begin{aligned} Y_t &= F\theta_t + v_t, \quad v_t \sim \text{Norm}(0, \sigma_\varepsilon^2) \\ &= \alpha_{0t} + \alpha_{1t}x_t + \varepsilon_t \end{aligned}$$

Y elegir  $G = I$ , y  $W$  como una matriz diagonal de  $2 \times 2$  tal que los elementos de sus diagonal son las varianzas de  $w_{0t}$  y  $w_{1t}$ . Esto para que la evolución del modelo sea la deseada.

Puede darse el caso en que la serie de tiempo  $Y_t$  no este afectada por una, sino por  $p > 1$  variables aleatorias explicatorias  $x_{1t}, \dots, x_{pt}$ . En esta situación el modelo queda como

$$Y_t = \alpha_t x_t + \varepsilon_t, \quad \varepsilon_t \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2),$$

donde  $\alpha_t = (\alpha_{0t}, \alpha_{1t}, \dots, \alpha_{pt})$  y  $x_t = (1, x_{1t}, \dots, x_{pt})^T$ . La evolución temporal se plantea como  $\alpha_{jt} = \alpha_{j,t-1} + w_{jt}$  (para toda  $j = 0, 1, \dots, p$ ) con  $w_{jt} \sim \mathcal{N}(0, W_{jt})$  y  $w_{jt}$  independiente de  $w_{it}$  siempre que  $i \neq j$ .

La generalización de la representación MDL del modelo con una variable explicativa a  $p$  variables explicativas resulta natural e intuitiva. Basta definir

$$\theta_t = \begin{pmatrix} \alpha_{0t} \\ \alpha_{1t} \\ \vdots \\ \alpha_{pt} \end{pmatrix}, \quad F = (1, x_{1t}, \dots, x_{pt}), \quad V_t = \sigma_\varepsilon^2, \quad G = I \quad y \quad W_t = \text{diag}(W_{0t}, W_{1t}, \dots, W_{pt}).$$

Para los modelos en los que la relación entre  $Y_t$  y la(s) variable(s) explicativa(s) no varía en el tiempo basta considerar  $W_t = 0$ .

En la librería `d1m` de R, también existe una función que representa a los modelos de regresión. Esta función se llama

■ `d1mModReg()`: Esta función recibe como parámetros

- (i) **X**: El valor de la variable(s) explicativas a cada tiempo  $t$ .
- (ii) **addInt**: Es un parámetro lógico, Puede ser **TRUE** o **FALSE** y se refiere a si se debe de agregar una intersección. Esto es, usando la notación anterior, en caso de ser **FALSE** significa que  $\alpha_{0t} = 0$  para toda  $t$ . La función lo considera verdadero de no ser especificado lo contrario.
- (iii) **dV**: Es la varianza observacional. El default es 1.

- (iv)  $dW$ : Se refiere los elementos de la diagonal del ruido del sistema. De no ser especificada se considera la matriz nula, es decir  $W_t = 0$ .
- (v)  $m_0$ : El vector de medias inicial. Si el usuario no lo define, el sistema lo considerará igual a  $(0, \dots, 0)$ .
- (vi)  $C_0$ : La matriz de covarianza de la distribución inicial del vector de estados. La función la definirá por default igual a  $diag(10^7, \dots, 10^7)$

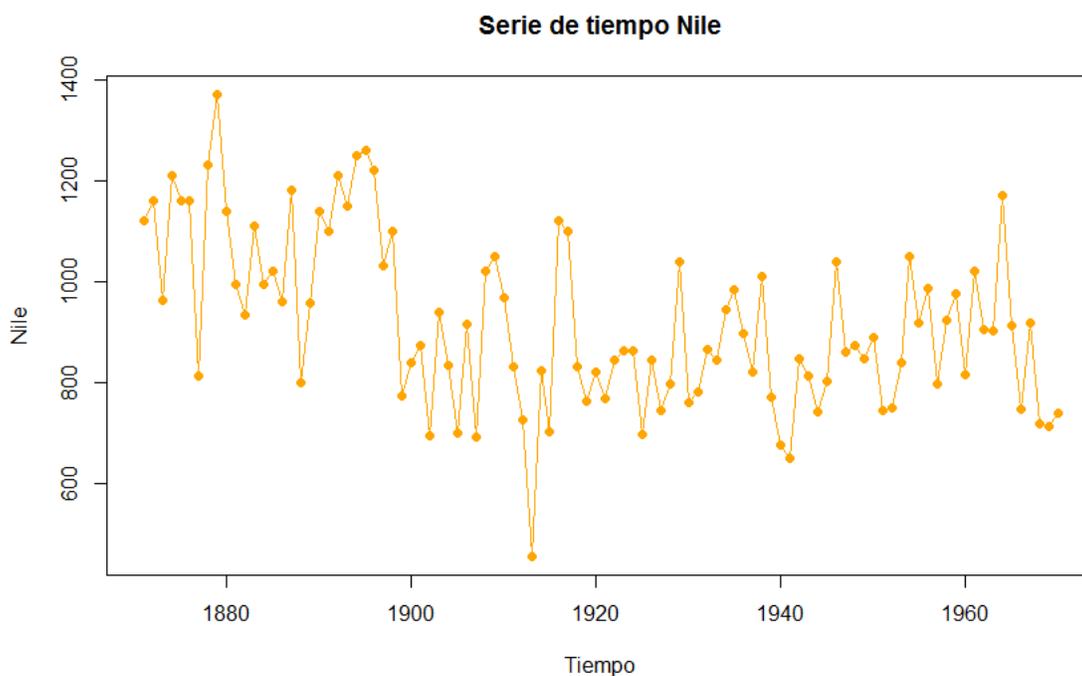
Con esto concluimos la sección de la especificación del modelo.

## 4. Aplicación de los Modelos Dinámicos Lineales a un caso práctico.

En los capítulos pasados se discutieron algunos aspectos teóricos de los Modelos Dinámicos Lineales y a través de ejemplos se ilustraron algunas características de los mismos. En este último capítulo aplicaremos estos conceptos a un caso práctico con la ayuda del paquete `dlm` que se encuentra en el software estadístico R.

La serie de tiempo que analizaremos, ilustrada a continuación, representa el volumen del Río Nilo en Aswan en los años 1870-1970, esta serie se encuentra guardada en R como `Nile`.

Figura 4.0.5.1: Observaciones de la serie `Nile`



Como se mencionó en capítulos pasados, dado un MDL y una serie de tiempo resultan de interés dos problemas principalmente, el primero es estimar a los estados y el segundo hacer predicción sobre futuras observaciones. Estos problemas se resuelven calculando  $\pi(\theta_s|y_{1:t})$  para alguna  $s \geq 0$  y  $\pi(Y_{t+k}|y_{1:t})$  para alguna  $k \geq 1$  respectivamente. Para el caso particular de un MDL, se sabe que todas estas densidades son Gaussianas, por lo que basta con calcular las medias y las varianzas de cada una para que estén completamente especificadas.

Para poder calcular densidades de suavización ( $\pi(\theta_s|y_{1:t})$  con  $s < t$ ) o de predicción

ya sea de estados o observaciones ( $\pi(\theta_s|y_{1:t})$  o  $\pi(Y_s|y_{1:t})$  con  $s > t$ ) necesitamos primero conocer la densidad de filtración ( $\pi(\theta_t|y_{1:t})$ ). Y esta última se puede calcular mediante el algoritmo recursivo de filtración que calcula  $\pi(\theta_t|y_{1:t-1})$ ,  $\pi(Y_t|y_{1:t-1})$  y  $\pi(\theta_t|y_{1:t})$  para  $t = 1, 2, \dots$  de manera recursiva.

Además de esto también nos interesa mucho conocer que tan adecuado es el modelo que se está proponiendo. Se ha planteado revisar el modelo a través del análisis de los residuales estandarizados los cuales son  $\tilde{e}_t = \frac{e_t}{\sqrt{Q_t}} = \frac{Y_t - f_t}{\sqrt{Q_t}}$  donde  $f_t = \mathbb{E}[Y_t|y_{1:t-1}]$  y  $Q_t = \text{Var}(Y_t|y_{1:t-1})$ . Por lo que para hacer este análisis basta calcular el algoritmo recursivo de filtración.

De esta manera se ha decidido analizar a la serie Nile en los siguientes pasos:

- (i) Proponer un modelo y especificarlo a través de la librería `dlm`
- (ii) Realizar el chequeo del modelo que se eligió (con las funciones `residuals()`, `tsdiag()` y `qqnorm()`).
- (iii) Calcular los algoritmos recursivos de filtración (con la función `dlmFilter()`); y estimar a los estados para los años 1870 – 1980 mediante la funciones `dlmSmooth()` y `dmlSmooth()`
- (iv) Predecir observaciones futuras con la función `dmlForecast()`.
- (v) Proponer modelos alternativos y realizar un análisis de sensibilidad del modelo.

## 4.1. Especificación del modelo

Lo primero que hay que hacer es proponer un modelo viable para ajustar a la serie de tiempo. A simple vista podemos ver que la serie aparenta no tener componente cíclico ni tendencia, sin embargo su media es claramente distinta de cero, por lo que podríamos proponer ajustar un modelo polinómico de orden 1, o lo que es lo mismo, un Modelo de Nivel Local. Recordemos que este modelo es de la forma:

$$\begin{aligned} Y_t &= F_t \theta_t + v_t & v_t &\sim \mathcal{N}(0, V_t), \\ \theta_t &= G_t \theta_{t-1} + w_t & w_t &\sim \mathcal{N}(0, W_t). \end{aligned}$$

con  $F_t = 1 = G_t$  para toda  $t$  y además

$$\theta_0 \sim \mathcal{N}(\mu_0, C_0).$$

A nuestra propuesta asignaremos  $\mu_0 = 1,100$  y  $C_0 = 10^7$ . Se decidió tomar la media inicial igual a 1,100 guiándonos en la información que nos brinda la gráfica de la serie

a simple vista, y como la información que se tiene acerca del valor de  $\mu_0$  es poca, se optado por elegir un valor muy grande para la varianza inicial.

En cuanto a  $V_t$  y  $W_t$  se ha decidido asignar el estimador máximo verosímil a cada una de estas varianzas. En este caso, es claro que no resulta trivial calcular estos estimadores. En la literatura se ha analizado y se han propuesto soluciones a este problema (revisar, por ejemplo, el libro [4] de la bibliografía). Además, por suerte para nosotros, existe una solución práctica y sencilla en R que nos propone la librería `d1m` a través de su función `d1mMLE`.

A grandes rasgos lo que hace esta función es, dado un vector de parámetros desconocidos, `parm`, (por supuesto parámetros desconocidos del MDL a especificar), dada una función `build` de los parámetros desconocidos que regresa un MDL y dada una serie de tiempo, `y`, calcula los estimadores máximo verosímiles de cada uno de los parámetros desconocidos. Esta función utiliza a su vez la función `d1mLL`, la cual calcula la log-verosimilitud<sup>20</sup> de un MEE, y la función `optim` la cual ofrece un algoritmo de optimización muy general que se puede particularizar a través de los parámetros que reciba.

En nuestro caso, el código para calcular los estimadores máximo verosímiles es el siguiente:

---

### Código en R

---

```
library(d1m)

buildFun <- function(x){
d1mModPoly(1, dV = exp(x[1]), dW = exp(x[2]), m0 = 1100)
}

fit <- d1mMLE(Nile, parm = c(0,0), build = buildFun)

d1mNile <- buildFun(fit$par)

print(d1mNile$W)
print(d1mNile$V)
```

---

De esta manera, la varianza de la ecuación de estados queda como  $W = 1468$  mientras que para la varianza observacional asignamos  $V = 15099$ .

En este punto, nuestro modelo ya ha quedado completamente especificado (considerando que ya habíamos asignado  $\mu_0 = 1,100$  y  $C_0 = 10^7$ ). Llamémosle Modelo 1 a este modelo, mismo que en R se especifica como sigue

---

<sup>20</sup>logaritmo natural de la verosimilitud

```
#Especificación del Modelo 1
mod1 <- dlmModPoly(order = 1, dV = 15099, dW = 1468, m0 = 1100)
```

---

## 4.2. Evaluación del modelo

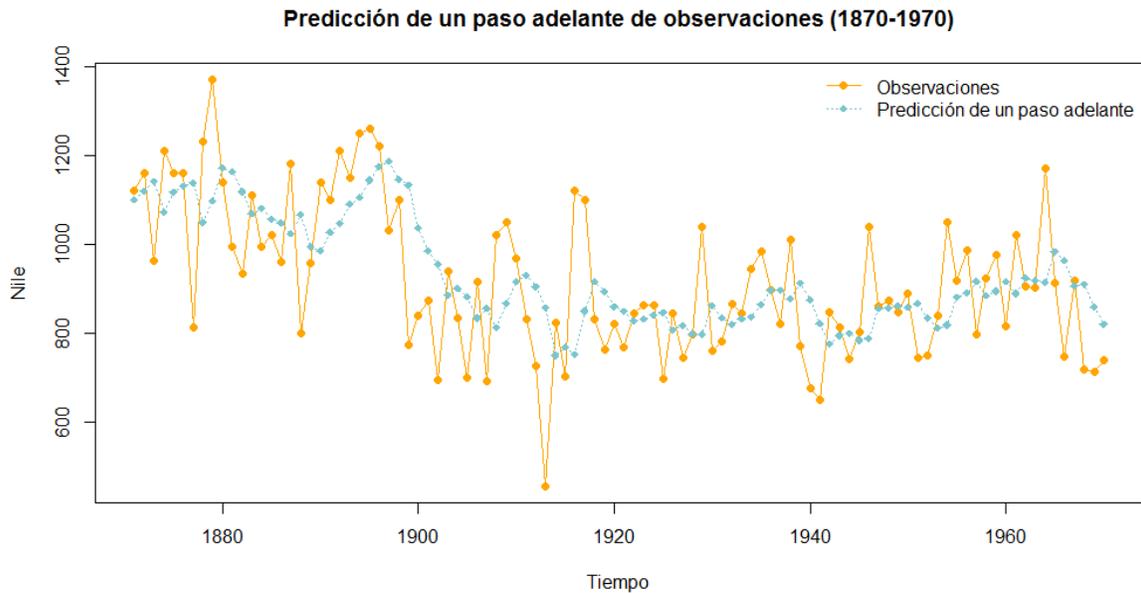
Como ya se ha visto lo primero que debemos hacer para realizar estimación de estado y/o predicción de observaciones futuras es calcular el filtro para ambos modelos. Este filtro, entre otras cosas, calcula  $f_t = \mathbb{E}[Y_t | y_{1:t-1}]$  que es precisamente la predicción que se hace de la observación al tiempo  $t$  dada la información disponible hasta el tiempo anterior  $t - 1$ .

A continuación se presenta el código en R que calcula el filtro y la gráfica que ilustra las observaciones y las predicciones de ellas (dada la información disponible hasta el momento anterior).

```
mod1Filt <- dlmFilter(Nile,mod1)
plot(Nile, type = 'o', pch = 19, col = "orange", xlab = "Tiempo",
main = "Predicción de un paso adelante de observaciones (1870-1970)")
lines(mod1Filt$f, type = 'o', lty = 3, pch = 18, col = "cadetblue3")
leg <- c("Observaciones", "Predicción de un paso adelante",)
legend("topright", legend = leg,
col = c("orange", "cadetblue3"),
lty = c(1, 3), pch = c(19, 18), bty = "n")
```

---

Figura 4.2.0.2: Predicción de un paso adelante de las observaciones según Modelo1



En la gráfica anterior podemos apreciar las predicciones de un paso adelante de las observaciones así como las observaciones mismas. La diferencia entre estas son los errores de predicción, a través de los cuales podemos realizar la evaluación del modelo dividiéndolos entre su desviación estándar (esto porque las observaciones y por lo tanto los errores de predicción son univariadas). Al dividir los errores de predicción entre su desviación estándar obtenemos los residuales estandarizados del Modelo 1. Ya hemos visto que si estos se comportan como una muestra i.i.d. de una distribución Normal estándar diremos que el modelo es adecuado.

El siguiente código extrae los residuales estandarizados del modelo filtrado y genera una serie de gráficas que nos hablan del comportamiento de estos.

---

### Código en R

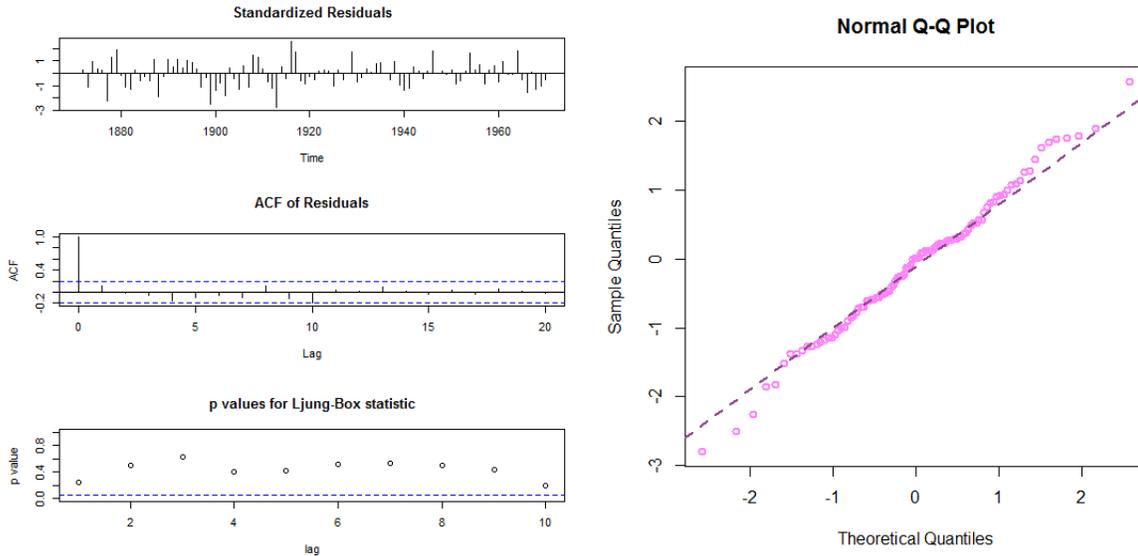
---

```
mod1res <- residuals(mod1Filt)
tsdiag(mod1Filt)
qqnorm(mod1res$res, pch = 1, lwd = 2, col = "orchid1")
qqline(mod1res$res, lwd = 2, lty = 2, col = "orchid4")
```

---

Las gráficas generadas son las siguientes

Figura 4.2.0.3: Verificación del Modelo



En la gráfica de los residuales podemos ver que la media de ellos aparenta ser cero, además encontramos que a simple vista sí aparentan comportarse como una muestra i.i.d de una distribución  $\mathcal{N}(0, 1)$ , ya que no hay evidencia de lo contrario.

La gráfica de la función de autocorrelación nos dice que los residuales si aparentan ser variables independientes. Mientras que los p-valores para la prueba de Ljung-Box nos indican que la autocorrelaciones conjuntamente hasta un lag son cero.

Por último la gráfica donde se comparan los cuantiles teóricos de una distribución Normal con los cuantiles de los residuales nos hablan de que estos son similares, por lo que podemos concluir que el proceso de los residuales estandarizados sí se comporta como una sucesión de variables aleatorias i.i.d. con distribución  $\mathcal{N}(0, 1)$

### 4.3. Estimación de estados

Para estimar los estados para los años 1870 – 1970 basta calcular  $\pi(\theta_s|y_{1:t})$  para  $s = 1, 2, \dots, 100$  donde  $y_{1:t}$  son las observaciones disponibles del nivel del Río Nilo.

Para  $s = t$  la densidad requerida ya la calculamos con el algoritmo de filtración. Cuando  $s < t$  estas densidades se calculan con un algoritmo recursivo hacia atrás de suavización, donde a partir de  $\pi(\theta_t|y_{1:t})$  se calcula  $\pi(\theta_{t-1}|y_{1:t})$  y basándonos en esta última se determina  $\pi(\theta_{t-2}|y_{1:t})$  y así consecutivamente.

Para el caso en que  $s > t$  también nos basamos en la densidad de filtración al tiempo  $t$  para calcular  $\pi(\theta_{t+1}|y_{1:t})$  y a partir de esta especificar  $\pi(\theta_{t+2}|y_{1:t})$  y así hasta que encontrar  $\pi(\theta_{t+10}|y_{1:t})$ .

Ya hemos calculado el filtro para cada modelo, utilizando las funciones `d1mSmooth()` y `d1mForecast()` podemos calcular los algoritmos recursivos de suavización y de predicción como sigue

---

### Código en R

---

```
mod1Smooth <- d1mSmooth(mod1Filt)
mod1Fore <- d1mForecast(mod1Filt)
```

---

El estimador de los estados a cada tiempo  $r$  será  $\hat{\theta}_r = \mathbb{E}[\theta_r|y_{1:t}]$ . Como ya se revisó, este estimador es el más adecuado pues todas las distribuciones de suavización y de filtración son Gaussianas.

Como la función `d1mSmooth()` regresa las medias y las varianzas (o la descomposición SVD de las varianzas) de estas densidades, y en adición se sabe que estas densidades son Normales, podemos conocer de manera explícita  $\pi(\theta_r|y_{1:t})$ . El hecho de conocer explícitamente dichas densidades nos permite construir intervalos de probabilidad para los estados,  $\theta_r$ , de manera muy sencilla. Por ejemplo, para construir un intervalo del 95 % de probabilidad para  $\theta_r$  basta calcular los cuantiles<sup>21</sup> 0.975 y 0.025 de una distribución Normal con media  $\mathbb{E}[\theta_r|y_{1:t}]$  y varianza  $Var(\theta_r|y_{1:t})$ .

A continuación se construyen en R intervalos de confianza para el proceso de estados y se genera una gráfica donde se presenta la estimación de este proceso junto con los intervalos del 95 % de probabilidad para el Modelo 1.

---

### Código en R

---

```
#Construcción de intervalos del 95% de probabilidad para estimación de edos.
#Modelo 1
#suavización
S1 <- unlist(d1mSvd2var(mod1Smooth$U.S,mod1Smooth$D.S))
q0.025smooth1 <- qnorm(0.025, mean = mod1Smooth$s, sd = sqrt(S1), lower = FALSE)
q0.975smooth1 <- qnorm(0.971, mean = mod1Smooth$s, sd = sqrt(S1), lower = FALSE)
#predicción
R1 <- unlist(mod1Fore$R)
q0.025fore1 <- qnorm(0.025, mean = mod1Fore$a, sd = sqrt(R1), lower = FALSE)
q0.975fore1 <- qnorm(0.971, mean = mod1Fore$a, sd = sqrt(R1), lower = FALSE)
```

---

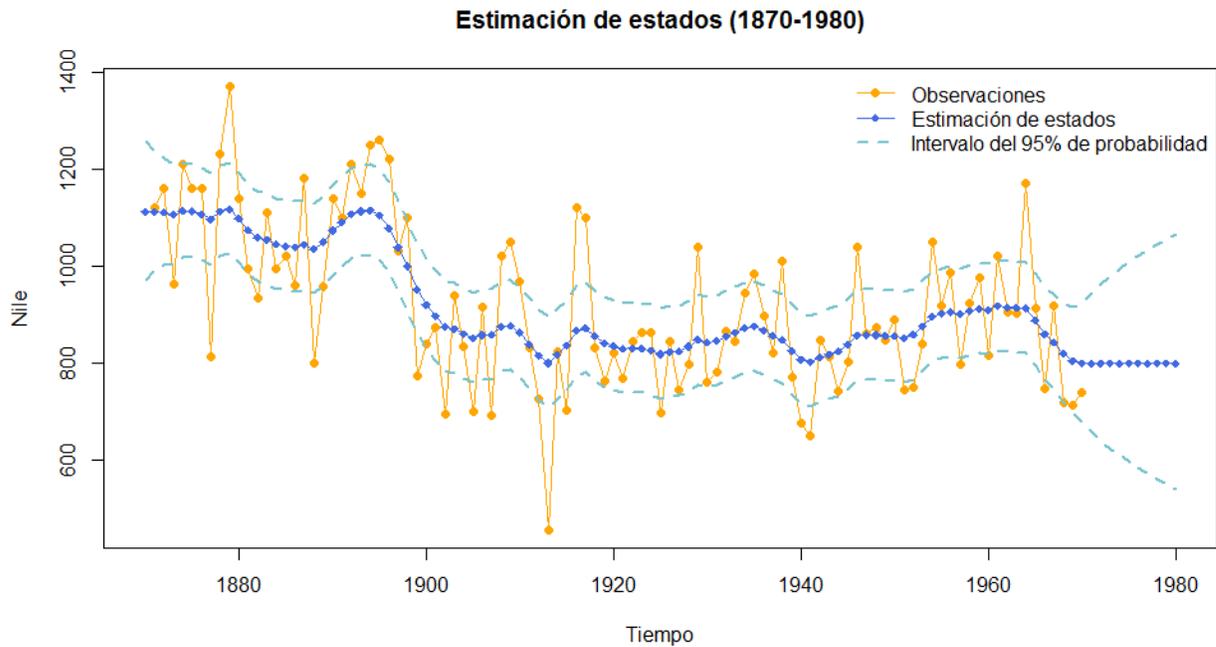
<sup>21</sup>El cuantil  $\alpha$  de la distribución de una v.a.  $X$  se define como el número  $q_\alpha$  tal que  $\mathbb{P}[X \leq q_\alpha] = \alpha$

```

#Gráfica de estimación de edos. e intervalos de probabilidad según Modelo 1
plot(Nile, type = 'o', lwd = 1, pch = 19, col = "orange", xlim = c(1870,1980),
xlab = "Tiempo", main = "Estimación de estados (1870-1980)")
lines(ts(c(mod1Smooth$s,mod1Fore$a), start = 1870),
type = 'o', lwd = 1, pch = 18, col = "royalblue3")
lines(ts(c(q0.025smooth1,q0.025fore1), start = 1870),
lty = 2, lwd = 2, col = "cadetblue3")
lines(ts(c(q0.975smooth1,q0.975fore1), start = 1870),
lty = 2, lwd = 2, col = "cadetblue3")
leg <- c("Observaciones", "Estimación de estados",
"Intervalo del 95% de probabilidad")
legend("topright", legend = leg,
col = c("orange", "royalblue", "cadetblue3"),
lty = c(1, 1, 2), pch = c(19, 18, NA), lwd = c(1,1,2), bty = "n")

```

Figura 4.3.0.4: Estimación de estados según el Modelo 1 dada la serie de tiempo Nile



En la gráfica los puntos color azul oscuro representan al estimador del estado a cada tiempo  $r$ ,  $\mathbb{E}[\theta_r|y_{1:t}]$ . Es decir, el valor que, dada la información que proporcionan las observaciones acerca de él, se espera que tenga el estado a cada tiempo.

Las líneas rayadas de color azul claro representan un intervalo del 95 % de probabilidad para el proceso de estados dadas las observaciones del nivel del Río Nilo. Es decir, si  $\alpha_r$  es el límite inferior del intervalo al tiempo  $r$  (valor que toma la línea de abajo en la gráfica) y  $\beta_r$  es el límite superior del intervalo al tiempo  $r$  (valor que toma la línea de arriba), entonces  $\mathbb{P}[\alpha_r \leq \theta_r < \beta_r | y_{1:t}] = 0.95$ .

Notese conforme más grande se  $k$   $\alpha_{t+k}$  y  $\beta_{t+k}$  (con  $t = 100$ ) estarán más alejados, esto era de esperarse ya que entre más pasos adelante queramos predecir más incierta será la predicción, por lo que el intervalo del probabilidad 95 % es mayor.

De esta manera ya hemos encontrado un estimador para los estados y además un intervalo tal que el proceso de estados pertenece a él con una probabilidad de 0.95. Veamos ahora que información podemos obtener del proceso de observaciones para tiempos futuros.

#### 4.4. Predicción de observaciones futuras

En ocasiones se desea hacer predicción acerca de observaciones futuras, en nuestro caso podría, por ejemplo interesarnos encontrar el valor esperado de la observación correspondiente al nivel del Río Nilo para el año 1971, 1972 y algunos otros posteriores a estos.

Como se ha venido haciendo, la predicción de las observaciones del año  $1970 + k$  será  $f_t(k) = \mathbb{E}[Y_{t+k} | y_{1:t}]$  donde  $y_{1:t}$  se refiere a las observaciones disponible del Río Nilo. Esta media la podemos calcular usando el algoritmo recursivo de predicción. El cual ya hemos calculado para poder hacer predicción sobre el proceso de estados. Recordemos que el código que usamos fue `mod1Fore <- dlmForecast(mod1Filt, nAhead = 10, sampleNew = 10)` . Con este código, además de calcular las medias y varianzas de las densidades de predicción, también hemos generado 10 muestras de las siguientes 10 observaciones.

Al igual que como se hizo con la estimación de estados, para las observaciones que no conocemos se puede construir un intervalo del 95 % de probabilidad dada la información disponible, es decir, se pueden encontrar, para  $t = 100$  y  $k \geq 1$ ,  $\alpha_t(k)$  y  $\beta_t(k)$  tales que  $\mathbb{P}[\alpha_t(k) \leq Y_{t+k} < \beta_t(k) | y_{1:t}] = 0.95$  . Nuevamente como estamos tratando con distribuciones Normales basta calcular los cuantiles 0.025 y 0.975 de una Normal con media  $\mathbb{E}[Y_{t+k} | y_{1:t}]$  y varianza  $Var(Y_{t+k} | y_{1:t})$  para encontrar un intervalo  $[\alpha_t(k), \beta_t(k)]$  que cumpla lo que deseamos, es claro que este intervalo no es único.

Se ha calculado en R dicho intervalo y se ha generado una gráfica donde se muestran las 10 muestras obtenidas, el valor esperado de las observaciones futuras y un intervalo del 95 % de confianza para estas.

---

#### Código en R

---

```
#Construcción de intervalos del 95% de probabilidad
```

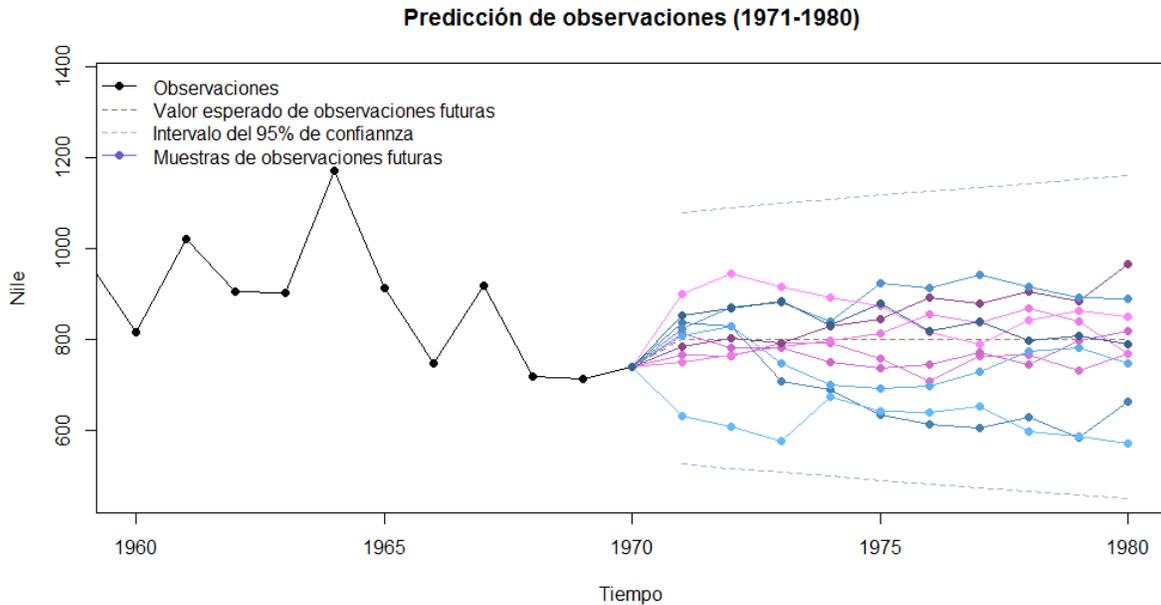
```

#para predicción de observaciones
Q1 <- unlist(mod1Fore$Q)
q0.025foreobs1 <- qnorm(0.025, mean = mod1Fore$f, sd = sqrt(Q1),
                        lower = FALSE)
q0.975foreobs1 <- qnorm(0.971, mean = mod1Fore$f, sd = sqrt(Q1),
                        lower = FALSE)

#Gráfica muestras de observaciones futuras
f1 <- ts(c(Nile[100],unlist(mod1Fore$newStates[1])), start = 1970)
f2 <- ts(c(Nile[100],unlist(mod1Fore$newStates[2])), start = 1970)
f3 <- ts(c(Nile[100],unlist(mod1Fore$newStates[3])), start = 1970)
f4 <- ts(c(Nile[100],unlist(mod1Fore$newStates[4])), start = 1970)
f5 <- ts(c(Nile[100],unlist(mod1Fore$newStates[5])), start = 1970)
f6 <- ts(c(Nile[100],unlist(mod1Fore$newStates[6])), start = 1970)
f7 <- ts(c(Nile[100],unlist(mod1Fore$newStates[7])), start = 1970)
f8 <- ts(c(Nile[100],unlist(mod1Fore$newStates[8])), start = 1970)
f9 <- ts(c(Nile[100],unlist(mod1Fore$newStates[9])), start = 1970)
f10 <- ts(c(Nile[100],unlist(mod1Fore$newStates[10])), start = 1970)
plot(Nile, type = 'o', lwd = 1, pch = 19, col = "black", xlim = c(1960,1980),
     xlab = "Tiempo", main = "Predicción de observaciones (1971-1980)")
lines(ts(q0.025foreobs1, start = 1971), lty = 2, lwd = 1, col = "slategray3")
lines(ts(q0.975foreobs1, start = 1971), lty = 2, lwd = 1, col = "slategray3")
lines(ts(c(Nile[100],mod1Fore$f), start = 1970), lty = 2, lwd = 1, col = "wheat4")
lines(f1, type = 'o', pch = 19, lty = 1, col = "orchid")
lines(f2, type = 'o', pch = 19, lty = 1, col = "orchid1")
lines(f3, type = 'o', pch = 19, lty = 1, col = "orchid2")
lines(f4, type = 'o', pch = 19, lty = 1, col = "orchid3")
lines(f5, type = 'o', pch = 19, lty = 1, col = "orchid4")
lines(f6, type = 'o', pch = 19, lty = 1, col = "steelblue")
lines(f7, type = 'o', pch = 19, lty = 1, col = "steelblue1")
lines(f8, type = 'o', pch = 19, lty = 1, col = "steelblue2")
lines(f9, type = 'o', pch = 19, lty = 1, col = "steelblue3")
lines(f10, type = 'o', pch = 19, lty = 1, col = "steelblue4")
leg <- c("Observaciones", "Valor esperado de observaciones futuras",
        "Intervalo del 95% de confiannza", "Muestras de observaciones futuras")
legend("topleft", legend = leg,
     col = c("black", "wheat4", "slategray3", "slateblue"),
     lty = c(1, 2, 2, 1), pch = c(19, NA, NA, 19), bty = "n")

```

Figura 4.4.0.5: Muestra de tamaño 10 de las 10 siguientes observaciones de la serie Nile



En la gráfica 4.4.0.5 se ilustran los intervalos de confianza, al igual que como sucede con los estados, conforme más pasos adelante queramos predecir, el intervalo se agrandará esto es  $|\beta_t(k_1) - \alpha_t(k_1)| < |\beta_t(k_2) - \alpha_t(k_2)|$  siempre que  $k_1 < k_2$ . Es visible también en la gráfica que las muestras de las observaciones futuras se encuentran todas dentro del intervalo de confianza construido, lo cual esperabamos ya que la probabilidad que las observaciones pertenezcan al intervalo es grande.

## 4.5. Propuesta de un Modelo Dinámico

En esta sección vamos a proponer un segundo modelo para la serie de tiempo Nile. Este segundo modelo tiene la característica de que no es constante en el tiempo.

La serie de tiempo aparenta tener un cambio drástico de nivel alrededor de los años 1898 y 1899 (que corresponden a las observaciones número 28 y 29 respectivamente). Esto se debe a que durante el año 1898 se contruyó una presa cerca del Río Nilo, la presa Ashwan, la cual se espera que cause un cambio más grande en el nivel local del Río. Fácilmente puede ser incorporada esta información al modelo, aprovechando que los MDLs no son necesariamente constantes en el tiempo. Hagamos entonces una segunda propuesta de modelo basándonos en el primer modelo y asignando un valor más grande a  $W_t$  para las observaciones número 28 y 29.

Nombremos Modelo 2 a la variante dinámica del Modelo 1, es decir al modelo para el

que  $W_{28}$  y  $W_{29}$  son mayores.

Ya conocemos la manera de especificar el modelo invariante en el tiempo a través del paquete `d1m`, veamos la manera de especificar el modelo dinámico en este mismo. Para este efecto haremos uso de algunos componentes de los objetos de clase `d1m` que no se han mencionado anteriormente. Estos componentes son `JFF`, `JGG`, `JW`, `JV` y `X`, cabe mencionar que en caso de no ser especificado lo contrario (cuando el modelo es constante en el tiempo) todos estos componentes se consideran nulos.

Para el caso particular del Modelo 2 utilizaremos únicamente `X` y `JW`, esto por que la varianza del sistema es la única matriz que no es constante en el tiempo. `JW` representa una matriz del mismo tamaño que el componente `W` del objeto tal que si `JW[i,j]` es igual a algún entero positivo `k` entonces `W[i,j]` toma el valor de `X[s,k]` al tiempo `s`. Si, por otro lado, `JW[i,j]` es igual a cero, esto significa que `W[i,j]` es constante en el tiempo. De manera análoga funcionan `JFF`, `JGG`, `JV` y `JV`.

---

### Código en R

---

```
#Especificación del Modelo 2
mod2 <- d1mModPoly(order = 1, dV = 15100, dW = 1468, m0 = 1100)
X <- matrix(mod2$W, nc = 1, nr = length(Nile))
X[28,] <- 12*mod2$W
X[29,] <- 12*mod2$W
mod2$X <- X
mod2$JW <- matrix(1,1,1)
```

---

Notemos que en este caso `X` es una matriz con una única columna debido a que solamente la varianza del sistema cambia en el tiempo y esta es univariada.

Esta manera de almacenar la información nos permite ahorrar espacio ya que no tenemos que guardar  $F_t$ ,  $G_t$ ,  $V_t$  y  $W_t$  para cada tiempo, basta con almacenar únicamente la información necesaria en `X` e indicar a través de `JFF`, `JGG`, `JV` y `JW` a que se refiere cada dato de `X`.

## 4.6. Comparación Modelo Estático vs. Modelo Dinámico

Para poder comparar ambos modelos lo primero que tenemos que hacer es calcular las densidades de filtración y de un paso adelante para el Modelo 2 mediante el comando `mod2Filt <- d1mFilter(Nile, mod2)` (las densidades de filtración del Modelo 1 ya la hemos calculado).

Una vez calculadas las densidades de un paso adelante de las observaciones para ambos modelos, generamos una gráfica a través del siguiente código que compara las predicción de un paso adelante del Modelo 1 y del Modelo 2 así como las observaciones de la serie Nile.

---

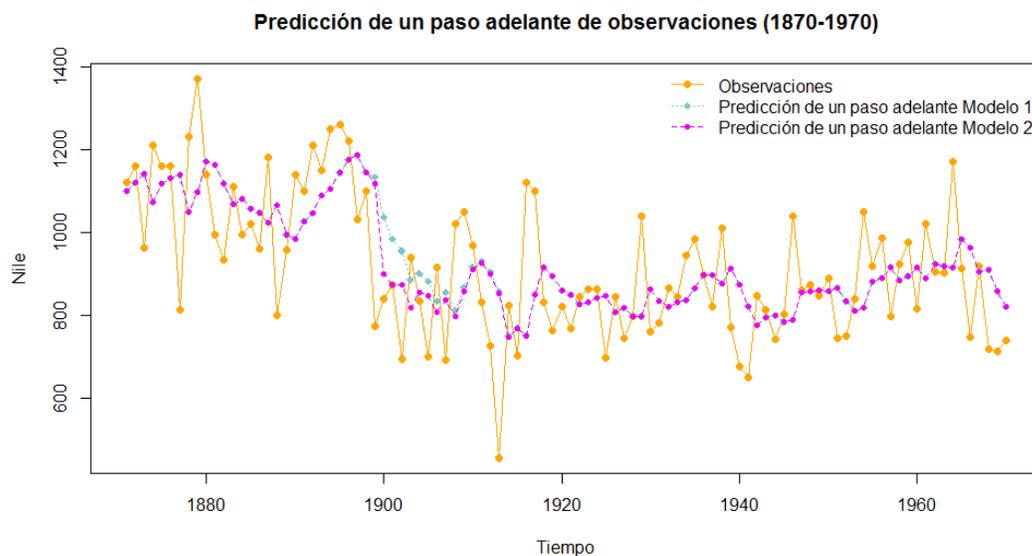
### Código en R

---

```
#Comparacion de modelos (Predicción de un paso adelante para obs.)
plot(Nile, type = 'o', pch = 19, col = "orange", xlab = "Tiempo",
main = "Predicción de un paso adelante de observaciones (1870-1970)")
lines(mod1Filt$f, type = 'o', lty = 3, pch = 18, col = "cadetblue3")
lines(mod2Filt$f, type = 'o', lty = 2, pch = 20, col = "magenta2")
legend("topright", legend = c("Observaciones", "Predicción de un paso adelante
modelo dinámico", "Predicción de un paso adelante modelo estático"),
col = c("orange", "cadetblue3", "magenta2"),
lty = c(1, 3, 2), pch = c(19, 18, 20), bty = "n")
```

---

Figura 4.6.0.6: Comparación de predicciones de un paso adelante para la observaciones según cada modelo



En la gráfica anterior podemos ver que las predicciones de un paso adelante de las observaciones según los dos modelos son similares. De hecho para los años anteriores a 1898 ambas predicciones son idénticas. Para los años 1898, 1899 y unos cuantos más adelante las predicciones difieren y poco a poco se va acortando esta diferencia hasta que alrededor

del año 1910 ambas se vuelven muy similares o al menos no existe diferencia significativa.

De manera similar podemos comparar la estimación de estados así como los intervalos del 95 % de probabilidad según ambos modelos para los años (1870-1970).

---

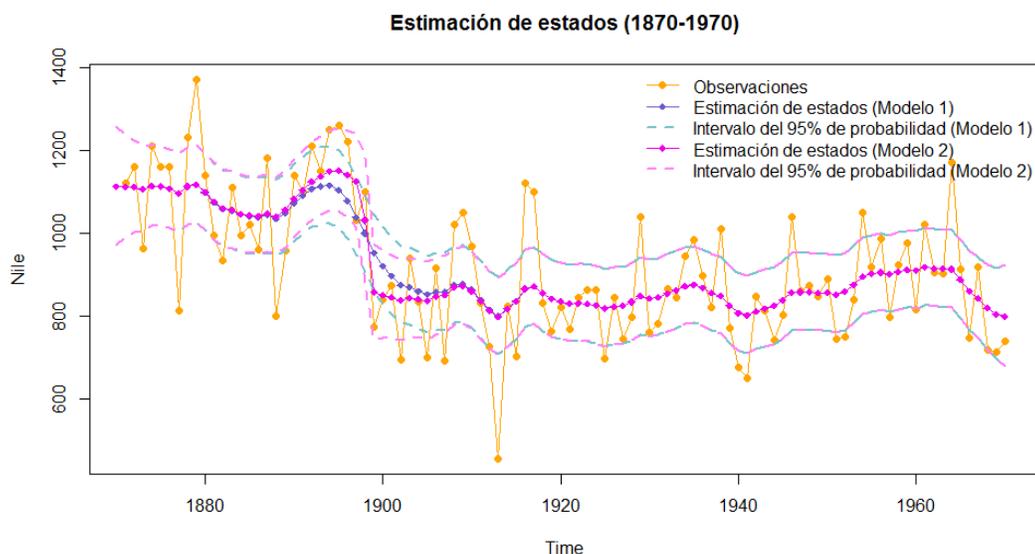
### Código en R

---

```
#Comparación (estimación de estados)
plot(Nile, type = 'o', lwd = 1, pch = 19, col = "orange",
     main = "Estimación de estados (1870-1970)")
lines(mod1Smooth$s, type = 'o', lwd = 1, pch = 18, col = "slateblue3")
lines(q0.025smooth1, lty = 2, lwd = 2, col = "cadetblue3")
lines(q0.975smooth1, lty = 2, lwd = 2, col = "cadetblue3")
lines(mod2Smooth$s, type = 'o', lwd = 1, pch = 18, col = "magenta2")
lines(q0.025smooth2, lty = 2, lwd = 2, col = "orchid1")
lines(q0.975smooth2, lty = 2, lwd = 2, col = "orchid1")
legend("topright", legend = c("Observaciones", "Estimación de estados (Modelo 1)",
                              "Intervalo del 95% de probabilidad (Modelo 1)", "Estimación de estados (Modelo 2)",
                              "Intervalo del 95% de probabilidad (Modelo 2)"),
      col = c("orange", "slateblue3", "cadetblue3", "magenta2", "orchid1"),
      lty = c(1, 1, 2, 1, 2), pch = c(19, 18, NA, 18, NA), lwd = c(1,1,2,1,2),
      bty = "n")
```

---

Figura 4.6.0.7: Comparación de estimación de estados según cada modelo



Notemos que en la gráfica del Modelo 2 el intervalo de probabilidad se vuelve más grande para los tiempos  $r$  en los que  $WR$  es mayor. También, alrededor de los tiempos  $r = 28, r = 29$  (cuando  $WR$  es mayor) la estimación del proceso de estados según el Modelo 2 se ve más influenciada por las observaciones que la estimación de los estados según el Modelo 1. Al igual que como sucede con la predicción de un paso adelante de las observaciones, tanto los intervalos de probabilidad como la estimación de estados son iguales (o al menos sin diferencia significativa) según ambos Modelos para los años cercanos a 1970.

El comparativo de la predicción de estados y de observaciones para los siguientes 10 años no se va a realizar debido a que la función `d1mForecast()` del paquete `d1m` no funciona con modelos que no son constantes en el tiempo. Sin embargo, en este caso esto no será un problema debido a que las densidades de filtración, suavización y predicción del Modelo 1 y del Modelo 2 son muy similares a partir del año 1910 y cada vez más parecidas. Alrededor del año 1970 la diferencia entre ambas no resulta significativa, por lo que podemos intuir que los intervalos de probabilidad y las predicciones tanto de estados como de observaciones de ambos modelos serán prácticamente iguales para ambos modelos. Por lo que no hay necesidad de calcularlas para el Modelo Dinámico.

Antes de dar por concluida esta sección conviene presentar la evaluación del modelo para nuestro Modelo Dinámico y usando los criterios MAP, MSE y MAPE comparar ambos modelos para ver cual es más adecuado.

---

### Código en R

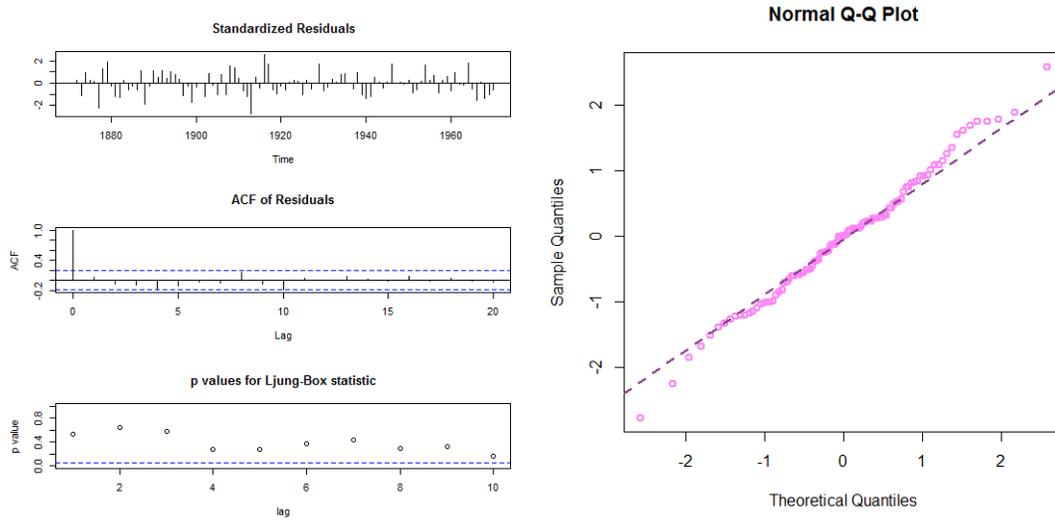
---

```
mod2res <- residuals(mod2Filt)
tsdiag(mod2Filt)
qqnorm(mod2res$res, pch = 1, lwd = 2, col = "orchid1")
qqline(mod2res$res, lwd = 2, lty = 2, col = "orchid4")
```

---

Las gráficas generadas son las siguientes:

Figura 4.6.0.8: Verificación del Modelo



Por lo que, bajo los mismos argumentos que para el Modelo Estático, podemos concluir que este modelo también resulta adecuado.

Claramente entre menores sean los errores de predicción, más adecuado será el modelo. Así que para decidir cual es el mejor de los dos modelos que propusimos. Se han calculado los criterios  $MAD = \sum_{i=1}^n |e_t|$ ,  $MSE = \sum_{i=1}^n (e_t)^2$  y  $MAPE = \sum_{i=1}^n \frac{|e_t|}{y_t}$  para cada modelo (donde  $n$  se refiere al número de observaciones que en este caso es 100):

---

**Código en R**

---

```
print(mean(abs(mod1Filt$f-Nile)))
print(mean((mod1Filt$f-Nile)^2))
print(mean(abs(mod1Filt$f-Nile)/Nile))
print(mean(abs(mod2Filt$f-Nile)))
print(mean((mod2Filt$f-Nile)^2))
print(mean(abs(mod2Filt$f-Nile)/Nile))
```

---

Resultando la información presentada en la tabla de abajo

	MAD	MSE	MAPE
mod1	112.6843	20485.81	0.12983
mod2	109.3761	19574.5	0.12538

Los tres criterios indican que entre el modelos estático y el modelo que cambia en el tiempo, el que hace mejores predicciones es el modelo no estático.

Con esto damos por concluido este pequeño análisis de la serie de tiempo Nile utilizando un Modelo Dinámico Lineal.

## Conclusiones

Los modelos dinámicos lineales son una muy bonita aplicación de la Estadística Bayesiana. Son una herramienta alternativa muy útil para analizar las series de tiempo ya que tiene las siguientes importantes ventajas:

- Son modelos muy flexibles gracias al proceso auxiliar que se supone para determinar la dependencia entre las observaciones.
- El hecho de suponer una dependencia markoviana en el proceso auxiliar y una dependencia condicionalmente independiente entre las observaciones dado el proceso auxiliar, en los Modelos Espaciales de Estados, nos permite calcular de manera recursiva y secuencial en el tiempo las densidades de interés.
- En los Modelos Dinámicos Lineales, gracias a que todas las distribuciones de interés son gaussianas, el cálculo de las densidades de interés se torna considerablemente más sencillo.
- Al realizar el análisis de una serie de tiempo a través de un Modelo Dinámico Lineal, resulta muy conveniente que estos puedan variar en el tiempo.
- No hay necesidad de realizar alguna transformación preliminar a los datos que hagan a la serie de tiempo estacionaria para entonces poder ajustar un modelo, como es el caso del análisis de series de tiempo hecho de manera convencional.
- Se puede incorporar de manera muy sencilla y natural alguna información empírica que se tenga acerca de la serie de tiempo.
- Se puede incorporar la dependencia de alguna otra variable a través del ajuste de un modelo de regresión, además los modelos convencionales (autoregresivos y/o de promedios móviles) tiene una representación como Modelos Dinámicos Lineales.
- Gracias a la propiedad de aditividad (el hecho de que la suma de dos Modelos Dinámicos Lineales resulte en otro Modelo Dinámico Lineal) podemos describir distintos rasgos de la serie de tiempo a través de distintos componentes de un mismo modelo.
- A través de los errores de predicción y el proceso de innovaciones podemos evaluar que tan adecuado es un modelo para una serie de tiempo.

Estás son las ventajas más significativas de los Modelos Dinámicos Lineales que encontré a lo largo del desarrollo de este proyecto.

En lo personal, creo que siempre es bueno tener herramientas diversas al momento de enfrentarnos a problemas, ya que nos dan más posibilidades de resolverlo de una manera adecuada. Creo que no deberían de estar peleadas las distintas formas de resolver un problema, al contrario, es bueno tomar lo mejor de que cada una. Creo que la herramientas

que se propone en este documento vale la pena analizarla debido a las muchas ventajas que nos proporciona en cuanto al análisis de series de tiempo.

Para concluir este trabajo me gustaría hablar un poco de lo que este trabajo hizo por mí. Creo que gracias a este ejercicio tuve la oportunidad de asentar los conocimientos adquiridos. Me ayudo a realizar, por primera vez en mi vida una investigación más completa y profunda acerca de un tema. Me gusta mucho que el tema que reúna una parte teórica de las matemáticas y que las aplique a algo tan práctico como el análisis de las series de tiempo, sobre todo porque creo que este tipo de análisis resulta muy útil y en ocasiones necesario para la solución de muchos problemas de diversas disciplinas. Me gusta que sea un tema relativamente nuevo (comparado con la antigüedad que pueden llegar a tener algunos temas en el área de las matemáticas), ya que creo que la manera de contribuir a nuestro mundo, a nuestra sociedad y a nuestro país es innovando. Estoy consciente que este tema no es una innovación mía, y no intento quitar el mérito a aquella persona que lo tiene, sin embargo creo que el conocer y entender los temas de vanguardia, las nuevas cosas que se van desarrollando, sin olvidarse de las bases que las sustentan, es la manera de educar a la mente para que eventualmente pueda obtener la suficiente madurez para innovar.

Espero que el lector haya encontrado este trabajo de utilidad y/o de interés. Me gustaría que más allá de ser un trabajo para lograr una titulación, ayudara a alguien a conocer un nuevo tema, o al menos lograr un mejor entendimiento de él. Sin más por el momento agradezco la atención y espero que lo hayan disfrutado.

## A. Anexos

### A.1. Gráficas

**Definición A.1.0.1.** Una gráfica  $G$  es un trío ordenado  $(V(G), E(G), \Psi_G)$  que consiste en un conjunto de vértices  $V(G) \neq \emptyset$ , un conjunto de aristas  $E(G)$  ajeno a  $V(G)$  (es decir,  $V(G) \cap E(G) = \emptyset$ ) y un función de incidencia  $\Psi_G$  que asigna a cada una de las aristas de  $G$  un par de vértices (no ordenados y no necesariamente distintos) de  $(G)$ .

Ejemplo:

Sea  $G = (V(G), E(G), \Psi_G)$  tal que:

$$V(G) = \{a, b, c, d, e\}$$

$$E(G) = \{f, g, h, i, j\}$$

tales que:

$$\Psi_G(f) = ad = da$$

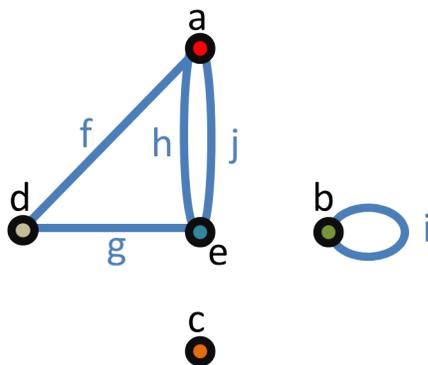
$$\Psi_G(g) = ed = de$$

$$\Psi_G(h) = ea = ae$$

$$\Psi_G(i) = bb$$

$$\Psi_G(j) = ea = ae.$$

Hay muchas formas de dibujar  $G$ , siempre y cuando se cumpla su descripción (es decir el conjunto de vértices, el conjunto de aristas y la función de incidencia). A continuación se presenta un dibujo de  $G$ :



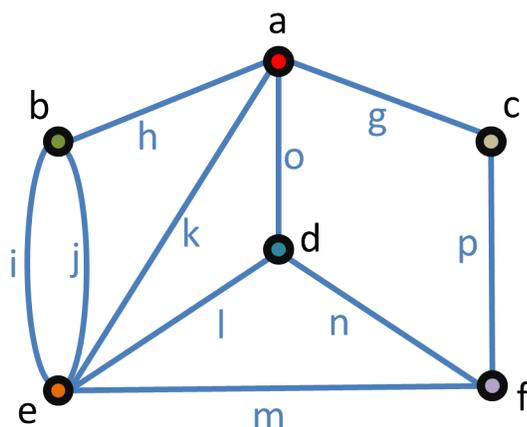
Las gráficas obtienen su nombre porque, como ya vimos, se pueden representar gráficamente, y esta representación gráfica nos ayuda a entender muchas de sus propiedades.

**Definición A.1.0.2.** Sea  $G$  una gráfica. Un camino en  $G$  es una secuencia finita y no nula  $W = v_0e_1v_1\dots v_{k-1}e_kv_k$  de vértices y aristas alternados (esto es, entre cada dos vértices hay una arista) tal que  $\forall 1 \leq i \leq k$ ,  $v_{i-1}$  y  $v_i$  son los extremos de  $e_i$  (es decir,  $\Psi_G(e_i) = v_{i-1}v_i$ ). El primer y el último elemento de un camino siempre serán vértices, al primer vértice se le conoce como origen y al último se le llama término.

Un recorrido en  $G$  es un camino  $W = v_0e_1v_1\dots v_{k-1}e_kv_k$  en la cual no se repiten aristas, es decir  $\forall 1 \leq i \leq k$  y  $\forall 1 \leq j \leq k$ ,  $i \neq j \Rightarrow e_i \neq e_j$ .

Por último, una trayectoria en  $G$  es un recorrido  $W = v_0e_1v_1\dots v_{k-1}e_kv_k$  en el cual no se repiten vértices, esto es  $\forall 0 \leq i \leq k$  y  $\forall 0 \leq j \leq k$ ,  $i \neq j \Rightarrow v_i \neq v_j$ . Notemos que una trayectoria es un camino en el cual todas sus aristas y todos sus vértices son distintos.

Ejemplo:



En la gráfica anterior  $W_1 = ahbieibjemfndoaod$  es un camino,  $W_2 = biejbhagcpfmekaod$  es un recorrido y  $W_3 = bhaodlemfpc$  es una trayectoria.

A un camino que tiene como origen a un vértice  $u$  y como término a un vértice  $v$  le llamamos  $(u, v)$ -camino. Si este camino no repite aristas se dice que es un  $(u, v)$ -recorrido, si en adición no repite vértices decimos que es una  $(u, v)$ -trayectoria.

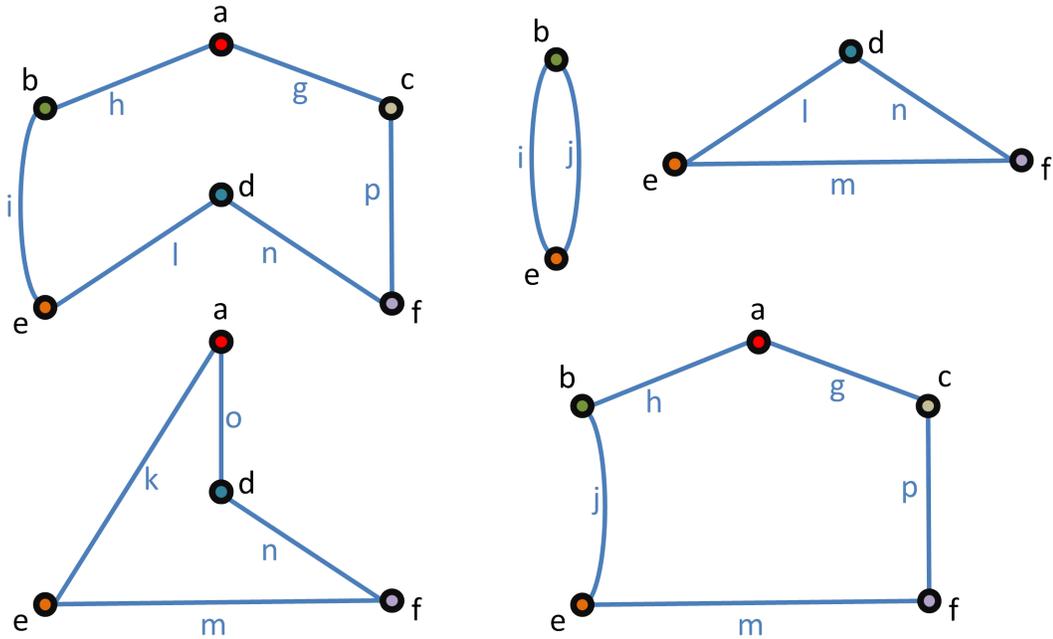
**Definición A.1.0.3.** Sean  $a$  y  $b$  dos vértices de una gráfica. Se dice que estos vértices están conectados si existe una  $(a, b)$ -trayectoria en la gráfica.

**Definición A.1.0.4.** Un camino cerrado es un camino  $W = v_0e_1v_1\dots v_{k-1}e_kv_k$  tal que  $v_0 = v_k$ , es decir, su origen y su término son el mismo vértice.

Un recorrido cerrado que cumple que su origen y sus vértices internos son distintos se llama ciclo. En otras palabras, un ciclo es un recorrido  $W = v_0e_1v_1\dots v_{k-1}e_kv_k$  tal que  $v_0 = v_k$  y  $v_i \neq v_j \forall 1 \leq i \leq k-1$  y  $\forall 1 \leq j \leq k-1$  con  $i \neq j$ .

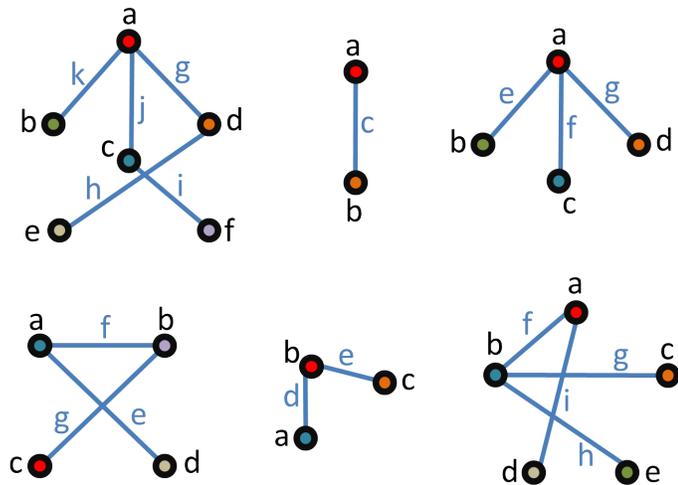
Ejemplo:

Considerando la gráfica del ejemplo anterior. Podemos ver que existen varios ciclos en ella, entre ellos  $agcpfndleibha$ ,  $eibje$ ,  $dnfmeld$ ,  $akemfndoa$ ,  $cgahbjemfpc$ :



**Definición A.1.0.5.** Decimos que una gráfica es acíclica si esta no contiene ningún ciclo.

A continuación se presentan algunos ejemplos de gráficas acíclicas:



**Definición A.1.0.6.** Una digráfica  $G$  es un trío ordenado  $(V(G), A(G), \Psi_G)$  que consiste en un conjunto de vértices  $V(G) \neq \emptyset$ , un conjunto de arcos  $A(G)$  ajeno a  $V(G)$  (es decir,

$V(G) \cap A(G) = \emptyset$ ) y un función de incidencia  $\Psi_G$  que asigna a cada uno de los arcos de  $G$  un par de vértices ordenados (no necesariamente distintos) de  $(G)$ .

En una digráfica los arcos tienen un sentido, es decir, se originan en algun vértice y tienen su destino en otro vértice (que puede ser el mismo vértice de origen). Recordemos que en las gráficas que hemos venido discutiendo las aristas no tenían un origen ni destino, tenía simplemente dos extremos.

Ejemplo:

Sea  $G = (V(G), A(G), \Psi_G)$  tal que:

$$V(G) = \{a, b, c, d, e, f\}$$

$$A(G) = \{g, h, i, j, k\}$$

tales que:

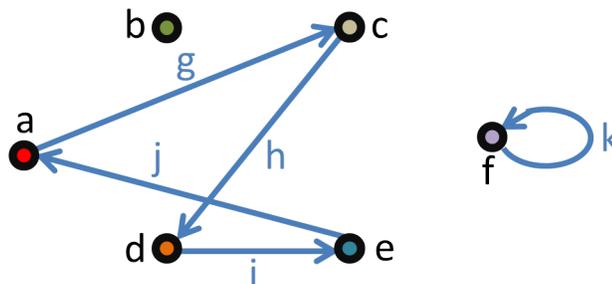
$$\Psi_G(g) = ac$$

$$\Psi_G(h) = cd$$

$$\Psi_G(i) = de$$

$$\Psi_G(j) = ea$$

$$\Psi_G(k) = ff.$$



## A.2. Matrices

**Definición A.2.0.7.** Una matriz  $A$  con entradas de un campo escalar  $F$  es una arreglo rectangular de la forma.

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,m} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,m} \end{pmatrix},$$

donde cada entrada  $a_{i,j} \in F$  para toda  $1 \leq i \leq n$ ,  $1 \leq j \leq m$ .

En este caso decimos que  $A$  es una matriz de  $n \times m$  ya que tiene  $n$  filas y  $m$  columnas.

En este documento denotemos  $A = (a_{i,j})_{i=1:n,j=1:m}$  para referirnos a la matriz anteriormente presentada, con el fin de utilizar una notación más corta. Además el campo escalar sobre el que están definidas todas las matrices (a lo largo de este documento) es  $\mathbb{R}$ , es decir  $F = \mathbb{R}$ .

### A.2.1. Suma de matrices y multiplicación por un escalar

**Definición A.2.1.1.** Definamos la suma de matrices. Sean  $A = (a_{i,j})_{i=1:n,j=1:m}$  y  $B = (b_{i,j})_{i=1:n,j=1:m}$  dos matrices de  $n \times m$ . Entonces  $A + B = C = (c_{i,j})_{i=1:n,j=1:m}$  donde  $c_{i,j} = a_{i,j} + b_{i,j}$ . Evidentemente para que sea posible sumar dos matrices, estas tienen que tener las mismas dimensiones.

**Definición A.2.1.2.** Dada una matriz de  $A = (a_{i,j})_{i=1:n,j=1:m}$ . Se define a la multiplicación de esta por un escalar  $\alpha$  a la matriz  $\alpha A = C = (c_{i,j})_{i=1:n,j=1:m}$  tal que  $c_{i,j} = \alpha a_{i,j}$ .

**Definición A.2.1.3.** La matriz nula de  $n \times m$ , a la cual denotaremos  $0_{n \times m}$  es la matriz tal que todos sus elementos son iguales a cero.

Esta matriz tiene la propiedad de que para cualquier matriz  $A$  de  $n \times m$ , resulta que  $A + 0_{n \times m} = 0_{n \times m} + A = A$ .

**Proposición A.2.1.1.** Sean  $A$ ,  $B$  y  $C$  matrices con las mismas dimensiones, y sean  $\alpha$ ,  $\beta$  escalares, entonces las siguientes afirmaciones son ciertas:

- (a)  $A + B = B + A$ .
- (b)  $A + (B + C) = (A + B) + C$ .
- (c)  $\alpha(A + B) = \alpha A + \alpha B$ .
- (d)  $\alpha(\beta A) = (\alpha\beta)A$ .
- (e)  $A + 0 = A$ .
- (f)  $A + (-A) = 0$ .

donde  $0$  es la matriz nula de las mismas dimensiones de  $A$ .

### A.2.2. Multiplicación de matrices

**Definición A.2.2.1.** Se define a la multiplicación de matrices de la siguiente forma. Sean  $A = (a_{i,j})_{i=1:n,j=1:m}$  y  $B = (b_{i,j})_{i=1:m,j=1:p}$  dos matrices de  $n \times m$  y  $m \times p$  respectivamente. Entonces  $AB = C = (c_{i,j})_{i=1:n,j=1:p}$  donde  $c_{i,j} = \sum_{k=1}^m a_{i,k} b_{k,p}$ . Notemos que para que se pueda realizar esta operación entre dos matrices, el número de columnas de la matriz de la izquierda debe ser igual al número de filas de la matriz de la derecha.

**Definición A.2.2.2.** Se define como matriz identidad de  $n \times n$  a la matriz  $I_n = (\delta_{i,j})_{i=1:n,j=1:n}$  tal que si  $i \neq j$  entonces  $\delta_{i,j} = 0$  y por el contrario, si  $i = j$  entonces  $\delta_{i,j} = 1$ . A  $\delta$  definida de esta forma se le llama delta de Kronecker.

La matriz identidad es la única matriz que cumple que si  $C$  es una matriz de  $m \times n$  y  $B$  una matriz de  $n \times p$  entonces  $CI_n = C$  y  $I_n B = B$ . Evidentemente si  $A$  es una matriz cuadrada de  $n \times n$  entonces  $AI_n = I_n A = A$ .

Mostrar que esto es cierto es muy sencillo. Ya que si  $A = (a_{i,j})_{i=1:n,j=1:n}$ , entonces para toda  $1 \leq i \leq n$  y para toda  $1 \leq j \leq n$  se tiene que

$$(I_n A)_{i,j} = \sum_{k=1}^n \delta_{i,k} a_{k,j} = a_{i,j} = \sum_{k=1}^n a_{i,k} \delta_{k,j} = (AI_n)_{i,j}$$

donde  $(I_n A)_{i,j}$  denota al elemento que está en la  $i$ -ésima fila y en la  $j$ -ésima columna de  $I_n A$ , y análogamente  $(AI_n)_{i,j}$  denota al elemento que está en la  $i$ -ésima fila y en la  $j$ -ésima columna de  $AI_n$ .

**Proposición A.2.2.1.** Sean  $A$  una matriz de  $n \times m$   $B$  y  $C$  matrices de  $m \times p$  y  $D$  una matriz de  $p \times q$ , sea  $\alpha$  cualquier escalar, entonces:

- (a)  $A(B + C) = AB + AC$ .
- (b)  $(B + C)D = BD + CD$ .
- (c)  $\alpha(AB) = (\alpha A)B = A(\alpha B)$ .
- (d)  $A0 = 0A = 0$ , donde  $0$  es una matriz nula con  $m$  filas.

**Definición A.2.2.3.** Decimos que una matriz es cuadrada si su número de filas es igual a su número de columnas

**Definición A.2.2.4.** Decimos que una matriz  $D = (d_{i,j})_{i=1:n,j=1:n}$  cuadrada de  $n \times n$  es diagonal si  $d_{i,j} = 0$  para toda  $i \neq j$  y  $d_{i,i} = d_i$  para toda  $1 \leq i \leq n$ . Es decir  $D$  es de la forma

$$D = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \end{pmatrix}.$$

Nótese que dadas dos matrices  $A$  y  $B$  diagonales de  $n \times n$ , se tiene que

$$AB = \begin{pmatrix} a_1 & 0 & \cdots & 0 \\ 0 & a_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_n \end{pmatrix} \begin{pmatrix} b_1 & 0 & \cdots & 0 \\ 0 & b_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & b_n \end{pmatrix} = \begin{pmatrix} a_1 b_1 & 0 & \cdots & 0 \\ 0 & a_2 b_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_n b_n \end{pmatrix}.$$

Claramente la multiplicación de dos matrices diagonales es una matriz diagonal, y para este caso la igualdad  $AB = BA$  sí se cumple.

### A.2.3. Matrices transpuestas y matrices simétricas

**Definición A.2.3.1.** Sea  $A = (a_{i,j})_{i=1:n,j=1:m}$ , se dice que la transpuesta de  $A$  es  $A^T = (a_{j,i})_{j=1:m,i=1:n}$ . Es decir, la transpuesta de  $A$  es la matriz resultante de cambiar sus filas por columnas.

**Proposición A.2.3.1.** Dadas cualesquiera matrices  $A$  y  $B$  y para cualquier escalar  $\alpha$ , las siguientes afirmaciones acerca de su transpuesta se cumplen:

- (a)  $(A^T)^T = A$ .
- (b)  $(A + B)^T = A^T + B^T$ .
- (c)  $(AB)^T = B^T A^T$ .
- (d)  $(\alpha A)^T = \alpha A^T$ .

**Definición A.2.3.2.** Decimos que una matriz  $A$  cuadrada de  $n \times n$  es simétrica si  $a_{i,j} = a_{j,i}$  para toda  $1 \leq i \leq n$ ,  $1 \leq j \leq n$ .

**Nota:** Es claro que si  $A$  es una matriz simétrica entonces  $A^T = A$ .

### A.2.4. Matrices inversas

**Definición A.2.4.1.** Sea  $A$  una matriz cuadrada, decimos que  $A$  es invertible si existe una matriz  $B$  (de  $n \times n$ ) tal que  $AB = BA = I_n$ . En dicho caso llamamos a  $B$  la matriz inversa de  $A$  y la denotamos por  $A^{-1}$ .

**Proposición A.2.4.1.** Dadas dos matrices invertibles  $A$  y  $B$  de  $n \times n$ , sus inversas tienen las siguientes propiedades

- (a)  $A^{-1}$  es única.
- (b)  $(A^{-1})^{-1} = A$ .
- (c)  $(AB)^{-1} = B^{-1}A^{-1}$ .
- (d)  $(\alpha A)^{-1} = \frac{1}{\alpha}A^{-1}$ .
- (e)  $(A^T)^{-1} = (A^{-1})^T$ .

**Definición A.2.4.2.** Sea  $A$  una matriz de  $n \times m$ , decimos que  $A$  tiene inversa por la izquierda si existe una matriz  $B$  de  $m \times n$  tal que  $BA = I_m$ , y decimos que  $A$  tiene inversa por la derecha si existe una matriz  $C$  de  $n \times m$  tal que  $CA = I_n$ .

### A.2.5. Determinantes

**Definición A.2.5.1.** Sea  $A = (a_{i,j})_{i=1:2,j=1:2}$  una matriz de  $2 \times 2$ , definimos al determinante de  $A$  denotado por  $|A|$  ó  $\det(A)$  como el escalar  $a_{1,1}a_{2,2} - a_{1,2}a_{2,1}$ .

**Definición A.2.5.2.** Sea  $A = (a_{i,j})_{i=1:n,j=1:n}$  una matriz de  $n \times n$  con  $n \geq 2$ , definimos al determinante de  $A$  como el escalar

$$|A| = \sum_{j=1}^n (-1)^{i+j} a_{i,j} |\widetilde{A}_{i,j}|$$

para cualquier fila  $i$  fija, donde  $\widetilde{A}_{i,j}$  es la matriz de  $(n-1) \times (n-1)$  resultante de quitar la  $i$ -ésima fila y la  $j$ -ésima columna de  $A$ .

**Nota:** Para una escalar  $a$  su determinante es el mismo, es decir,  $|a| = a$ .

**Definición A.2.5.3.** Sea  $A$  una matriz de  $n \times n$ , definimos a la matriz adjunta de  $A$  denotada por  $\text{Adj}(A)$  a la matriz tal que para toda  $i, j \in \{1, 2, \dots, n\}$

$$\text{Adj}(A)_{i,j} = |\widetilde{A}_{i,j}|$$

donde  $\text{Adj}(A)_{i,j}$  representa al elemento de  $\text{adj}(A)$  que se encuentra en la  $i$ -ésima fila y en la  $j$ -ésima columna,  $\widetilde{A}_{i,j}$ , al igual que en la definición anterior, es la matriz de  $(n-1) \times (n-1)$  resultante de quitar la  $i$ -ésima fila y la  $j$ -ésima columna de  $A$ .

**Proposición A.2.5.1.** Sean  $A$  y  $B$  dos matrices de  $n \times n$  y  $\alpha$  un escalar, las siguientes afirmaciones acerca de sus determinantes se cumplen:

- (a)  $\det(A^T) = \det(A)$ .
- (b)  $\det(AB) = \det(A)\det(B) = \det(BA)$ .
- (c)  $\det(\alpha A) = \alpha^n \det(A)$ .
- (d)  $A$  es invertible si y sólo si  $\det(A) \neq 0$ .
- (e) Si  $A$  es invertible entonces  $\det(A^{-1}) = \frac{1}{\det(A)}$ .
- (d) Si  $A$  es invertible  $A^{-1} = \frac{1}{\det(A)} \text{Adj}(A)$ .

### A.2.6. Matrices ortogonales y matrices definidas positivas

**Definición A.2.6.1.** Decimos que dos vectores  $v = (v_1, v_2, \dots, v_n)^T$  y  $u = (u_1, u_2, \dots, u_n)^T$  son ortogonales si

$$v^T u = \sum_{i=1}^n v_i u_i = u^T v = 0.$$

**Definición A.2.6.2.** Definimos a la norma de un vector  $v = (v_1, v_2, \dots, v_n)^T$  como

$$\|v\| = \sqrt{v^T v} = \sqrt{\sum_{i=1}^n (v_i)^2}.$$

**Definición A.2.6.3.** Decimos que dos vectores  $v$  y  $u$  son ortonormales si

(i)  $v$  y  $u$  son ortogonales, y

(ii)  $\|v\| = \|u\| = 1$ .

**Definición A.2.6.4.** Decimos que una matriz cuadrada  $A$  es ortogonal si  $A^{-1} = A^T$ . Esto es  $A^T A = A A^T = I$ .

Notemos que el hecho de que una matriz sea ortogonal significa que sus filas y sus columnas son ortonormales. Para ver esto consideremos a una matriz  $A$  de  $n \times n$  ortogonal, denotamos a la  $i$ -ésima fila de  $A$  como  $A_{(i)}$  y a la  $j$ -ésima columna de  $A$  como  $A^{(j)}$ . Es claro que  $(A^T)_{(i)} = (A^{(i)})^T$  así como  $(A^T)^{(i)} = (A_{(i)})^T$ . Como  $A^T A = I$  Esto significa que si  $i = j$  entonces  $1 = (A^T)_{(i)} A^{(j)} = (A^{(i)})^T A^{(j)}$  y si  $i \neq j$  entonces  $0 = (A^T)_{(i)} A^{(j)} = (A^{(i)})^T A^{(j)}$ . De manera análoga, como consecuencia de que  $A A^T = I$  se puede ver que si  $i = j$  esto implica que  $A_{(i)} (A_{(j)})^T = 1$  y por el contrario, si  $i \neq j$  entonces  $A_{(i)} (A_{(j)})^T = 0$ . Lo que significa que todas las columnas de  $A$  tienen norma igual a 1 y que tanto sus filas como sus columnas son ortogonales entre ellas.

Notemos también, que dado un vector  $x = (x_1, x_2, \dots, x_n)^T$  y una matriz  $A$  ortogonal, si definimos  $y = Ax$ . Entonces es fácil ver que  $y$  tiene la misma norma que  $x$  ya que

$$y^T y = (Ax)^T (Ax) = x^T A^T A x = x^T x.$$

**Definición A.2.6.5.** Dada una matriz cuadrada  $A$  de  $n \times n$  decimos que  $\lambda$  es un valor propio de  $A$  si

$$\det(A - \lambda I) = 0.$$

**Proposición A.2.6.1.** Dada una matriz simétrica  $A$  existe una matriz  $C$  ortogonal de las mismas dimensiones de  $A$  tal que

$$C^T A C = D$$

donde

$$D = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}$$

es una matriz diagonal tal que los elementos de su diagonal  $\lambda_1, \lambda_2, \dots, \lambda_n$  son los valores propios de  $A$ .

La demostración de esta proposición se encuentra en muchos libros de Álgebra Lineal, como por ejemplo en el libro [9] de la bibliografía.

Por la proposición anterior se tiene que

$$\begin{aligned} \det(A) &= \frac{1}{\det(C)} \det(A) \det(C) = \det(C^T A C) \\ &= \det(C^T) \det(A) \det(C) = \det(D) = \prod_{i=1}^n \lambda_i \end{aligned}$$

**Definición A.2.6.6.** Definimos a la forma cuadrática basada en la matriz simétrica  $A$  de  $n \times n$  como

$$Q_A = Q_A(x) = x^T A x = \sum_{i=1}^n \sum_{j=1}^n a_{i,j} x_i x_j$$

donde  $x = (x_1, x_2, \dots, x_n)^T \in \mathbb{R}^n$ .

**Definición A.2.6.7.** Decimos que una matriz simétrica  $A = (a_{i,j})_{i=1:n,j=1:n}$  de  $n \times n$  es definida positiva (semidefinida positiva) si se cumplen las siguientes condiciones equivalentes

- $Q_A(x) > 0$  ( $\geq 0$ ) para todo vector no nulo  $x = (x_1, x_2, \dots, x_n)^T \in \mathbb{R}^n$ .
- Todos los valores propios de  $A$  son mayores que cero (mayores o iguales a cero).
- $|A_k| > 0$  ( $\geq 0$ ) para toda  $1 \leq k \leq n$ , donde  $A_k = (a_{i,j})_{i=1:k,j=1:k}$  es la submatriz de  $A$  de tamaño  $k \times k$  que resulta de quitar las últimas  $n - k$  filas y las últimas  $n - k$  columnas de  $A$ .

En algunos libros se utiliza la notación  $A > 0$  ( $\geq 0$ ) para decir que  $A$  es una matriz definida (semidefinida positiva).

Ahora consideremos a una matriz  $A$  de  $n \times n$  simétrica y semidefinida positiva, por la proposición A.2.6.1 existe una matriz ortogonal  $C$  (de  $n \times n$ ) tal que  $C^T A C = D$  donde  $D = (d_{i,j})_{i=1:n,j=1:n}$  es la matriz diagonal tal que  $d_{i,i} = \lambda_i$  donde  $\lambda_1, \lambda_2, \dots, \lambda_n$  son los valores propios de  $A$ , Ahora, el hecho de que  $C$  es ortogonal significa que  $C^T = C^{-1}$  por lo que  $A = C D C^T$ . Sea  $\hat{D} = (\hat{d}_{i,j})_{i=1:n,j=1:n}$  la matriz diagonal tal que  $\hat{d}_{i,i} = +\sqrt{d_{i,i}} = +\sqrt{\lambda_i}$ , es claro que  $\hat{D} \hat{D} = D$ . Por último definamos a la matriz  $B = C \hat{D} C^T$ , se tiene entonces que

$$B^2 = B B = (C \hat{D} C^T)(C \hat{D} C^T) = C \hat{D} \hat{D} C^T = C D C^T = A.$$

Si  $A$  es una matriz definida positiva sea  $\tilde{D} = \hat{D}$ . Por otro lado si  $A$  tiene  $1 < m < n$  valores propios iguales a 0, podemos suponer sin pérdida de generalidad que  $\lambda_i > 0$  para toda  $i = 1 : n - m$  y  $\lambda_i = 0$  para toda  $i = n - m + 1 : n$  de tal forma que  $D = (d_{i,j})_{i=1:n,j=1:n}$

es una matriz diagonal tal que  $d_{i,i} = \lambda_i > 0$  para  $i = 1 : n - m$  y  $d_{i,i} = \lambda_i = 0$  para  $i = n - m + 1 : n$ . En este caso definamos a la matriz de  $n \times (n - m)$

$$\tilde{D} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_{n-m} \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}.$$

Es fácil ver que, en cualquiera de los dos casos,  $\tilde{D}\tilde{D}^T = D$ . Además, en el caso que  $A$  es definida positiva  $\tilde{D}$  es una matriz de  $n \times n$  invertible. Por otro lado, cuando  $A$  es semidefinida positiva  $\tilde{D}$  tiene inversa por la izquierda así como  $\tilde{D}^T$  tiene inversa por la derecha. Si hacemos  $G = C\tilde{D}$ , se tiene entonces que

$$GG^T = (C\tilde{D})(C\tilde{D})^T = C\tilde{D}\tilde{D}^T C^T = C\tilde{D}\tilde{D}C^T = CDC^T = A$$

para ambos casos.

### A.3. Propiedades importantes de la distribución Normal

Decimos que una variable aleatoria  $X$  cuya función de densidad es de la forma

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$

sigue una distribución Normal con media  $\mu$  y varianza  $\sigma^2$ , y lo denotaremos de la siguiente manera:  $X \sim \mathcal{N}(\mu, \sigma^2)$ .

Es claro que la función de distribución de  $X$  será:

$$F_X(x) = \int_{-\infty}^x f_X(t)dt = \int_{-\infty}^x \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right\} dt$$

Más adelante veremos que efectivamente  $\mathbb{E}[X] = \mu$  y  $Var(X) = \sigma^2$ .

**Definición:** Se define a la **función generadora de momentos** de una variable aleatoria  $X$  como:

$$M_X(t) = \mathbb{E}[e^{tX}] \quad \forall t \in \mathbb{R}.$$

Se le llama función generadora de momentos ya que

$$\mathbb{E}[X^n] = M_X^{(n)}(0) = \frac{d^n M_X(t)}{dt^n} \Big|_{t=0}$$

**Demostración:**(de la afirmación anterior)

Sea  $X$  una variable aleatoria continua con función de densidad  $f_X(x)$ , sea  $t \in \mathbb{R}$  entonces

$$\begin{aligned}M_X(t) &= \mathbb{E}[e^{tX}] \\&= \int_{-\infty}^{\infty} e^{tx} f_X(x) dx \\&= \int_{-\infty}^{\infty} \left(1 + tx + \frac{t^2 x^2}{2!} + \dots + \frac{t^n x^n}{n!} + \dots\right) f_X(x) dx \\&= 1 + t\mathbb{E}[X] + \frac{t^2 \mathbb{E}[X^2]}{2!} + \dots + \frac{t^n \mathbb{E}[X^n]}{n!} + \dots\end{aligned}$$

Al derivar  $M_X(t)$  con respecto a  $t$  una vez obtenemos

$$M'_X(t) = \mathbb{E}[X] + t\mathbb{E}[X^2] + \dots + \frac{t^{n-1} \mathbb{E}[X^n]}{(n-1)!} + \dots$$

Si derivamos  $M_X(t)$  con respecto a  $t$  dos veces se tiene que

$$M''_X(t) = \mathbb{E}[X^2] + \dots + \frac{t^{n-2} \mathbb{E}[X^n]}{(n-2)!} + \dots$$

De la misma forma, al realizar la  $n$ -ésima derivada de  $M_X(t)$  con respecto a  $t$  se obtendrá

$$M_X^{(n)}(t) = \mathbb{E}[X^n] + t\mathbb{E}[X^{n+1}] + \frac{t^2 \mathbb{E}[X^{n+2}]}{2!} + \dots$$

Notemos que para cualquier  $n \in \mathbb{N}$  se cumple la ecuación anterior. Por lo tanto, al evaluar  $M_X^{(n)}(t)$  en  $t = 0$  llegaremos a lo que queríamos demostrar:

$$M_X^{(n)}(0) = \mathbb{E}[X^n].$$

**Proposición:** Sea  $X \sim \mathcal{N}(\mu, \sigma)$ . Su función generadora de momentos es  $M_X(t) = \exp\left\{\mu t + \frac{t^2 \sigma^2}{2}\right\}$ .

**Demostración:**

$$\begin{aligned}
M_X(t) &= \int_{-\infty}^{\infty} \exp\{tx\} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{\frac{-x^2 + 2x\mu - \mu^2 + 2xt\sigma^2}{2\sigma^2}\right\} dx \\
&= \exp\left\{-\frac{\mu^2}{2\sigma^2}\right\} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{\frac{-x^2 + 2x(\mu + t\sigma^2)}{2\sigma^2}\right\} dx \\
&= \exp\left\{-\frac{\mu^2}{2\sigma^2} + \frac{(\mu + t\sigma^2)^2}{2\sigma^2}\right\} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{\frac{-x^2 + 2x(\mu + t\sigma^2) - (\mu + t\sigma^2)^2}{2\sigma^2}\right\} dx \\
&= \exp\left\{\frac{2\mu t\sigma^2 + t^2\sigma^4}{2\sigma^2}\right\} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x - (\mu + t\sigma^2))^2}{2\sigma^2}\right\} dx \\
&= \exp\left\{\mu t + \frac{t^2\sigma^2}{2}\right\}.
\end{aligned}$$

**Proposición:** Sea  $X \sim \mathcal{N}(\mu, \sigma^2)$ . Entonces  $\mathbb{E}[X] = \mu$  y  $\text{Var}(X) = \sigma^2$ .

**Demostración:**

Hasta ahora ya sabemos que  $\mathbb{E}[X^n] = M_X^{(n)}(0)$  para cualquier  $n \in \mathbb{N}$  y que  $M_X(t) = \exp\left\{\mu t + \frac{t^2\sigma^2}{2}\right\}$ .

$$\begin{aligned}
&\Rightarrow M_X'(t) = (\mu + t\sigma^2) \exp\left\{\mu t + \frac{t^2\sigma^2}{2}\right\} \\
&\Rightarrow M_X''(t) = (\sigma^2) \exp\left\{\mu t + \frac{t^2\sigma^2}{2}\right\} + (\mu + t\sigma^2)^2 \exp\left\{\mu t + \frac{t^2\sigma^2}{2}\right\}.
\end{aligned}$$

Evaluando cada una de las dos funciones anteriores en  $t = 0$  se obtiene

$$\begin{aligned}
\mathbb{E}[X] &= M_X'(0) = \mu, \\
\mathbb{E}[X^2] &= M_X''(0) = \sigma^2 + \mu.
\end{aligned}$$

Por lo que se concluye que

$$\mathbb{E}[X] = \mu$$

y

$$\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \sigma^2.$$

**Proposición:** Sean  $X$  y  $Y$  variables aleatorias. Si  $X$  y  $Y$  son independientes, entonces  $M_{X+Y}(t) = M_X(t)M_Y(t)$ .

**Demostración:**

$$\begin{aligned}
 M_{X+Y}(t) &= \mathbb{E}[e^{(X+Y)t}] \\
 &= \mathbb{E}[e^{Xt}e^{Yt}] \\
 &= \mathbb{E}[e^{Xt}]\mathbb{E}[e^{Yt}] \\
 &= M_X(t)M_Y(t).
 \end{aligned}$$

**Proposición:** Sea  $X \sim \mathcal{N}(\mu, \sigma^2)$ , sean  $a, b \in \mathbb{R}$  y  $Y = aX + b$ . Entonces  $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$ .

**Demostración:**

$$\begin{aligned}
 M_Y(t) &= \mathbb{E}[e^{(aX+b)t}] \\
 &= \mathbb{E}[e^{X(at)+bt}] \\
 &= e^{bt} \left( \exp \left\{ \mu(at) + \frac{\sigma^2(at)^2}{2} \right\} \right) \\
 &= \exp \left\{ bt + a\mu(t) + \frac{a^2\sigma^2(t^2)}{2} \right\} \\
 &= \exp \left\{ (a\mu + b)t + \frac{a^2\sigma^2(t^2)}{2} \right\}
 \end{aligned}$$

$$\therefore Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2). \quad (\text{A.3.0.1})$$

**Proposición:** Sean  $X$  y  $Y$  variables aleatorias independientes tales que  $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$  y  $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ . Sea  $Z = X + Y$ , entonces  $Z \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$ .

**Demostración:**

$$\begin{aligned}
 M_Z(t) &= M_X(t)M_Y(t) \\
 &= \exp \left\{ \mu_X t + \frac{t^2\sigma_X^2}{2} \right\} \exp \left\{ \mu_Y t + \frac{t^2\sigma_Y^2}{2} \right\} \\
 &= \exp \left\{ t(\mu_X + \mu_Y) + \frac{t^2(\sigma_X^2 + \sigma_Y^2)}{2} \right\}.
 \end{aligned}$$

### A.3.1. Normal multivariada

**Definición A.3.1.1.** Un vector aleatorio es un vector formado por una o más variables aleatorias escalares.

**Definición A.3.1.2.** Sea  $X = (X_1, X_2, \dots, X_n)^T$  un vector aleatorio, definimos al vector de medias de  $X$  como  $\mu = \mathbb{E}[X]$  donde  $\mu$  es tal que cada uno de sus componentes  $\mu_i = \mathbb{E}[X_i]$  para toda  $i = 1, 2, \dots, n$ .

La matriz de covarianza de  $X$  se define como  $\Sigma = Cov(X) = \mathbb{E}[(X - \mu)(X - \mu)^T]$  donde sus componentes son tales que  $\sigma_{i,j} = Cov(X_i, X_j)$  para toda  $i, j = 1, 2, \dots, n$ . Cuando  $i = j$  entonces  $\sigma_{i,i} = Cov(X_i, X_i) = Var(X_i)$ .

Notemos que  $\Sigma$  es una matriz simétrica (es decir,  $\Sigma = \Sigma^T$ ) ya que para toda  $i, j = 1, 2, \dots, n$   $\sigma_{i,j} = Cov(X_i, X_j) = Cov(X_j, X_i) = \sigma_{j,i}$ .

Algunas veces también se escribe  $Var(X)$ , donde  $X$  es un vector aleatorio, para referirse a la matriz de covarianza de  $X$ .

**Proposición A.3.1.1.** Sea  $X = (X_1, X_2, \dots, X_n)^T$  un vector aleatorio con vector de medias  $\mu = (\mu_1, \mu_2, \dots, \mu_n)^T$  y matriz de covarianza

$$\Sigma = \begin{pmatrix} \sigma_{1,1} & \sigma_{1,2} & \cdots & \sigma_{1,n} \\ \sigma_{2,1} & \sigma_{2,2} & \cdots & \sigma_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n,1} & \sigma_{n,2} & \cdots & \sigma_{n,n} \end{pmatrix}.$$

Se tiene entonces que  $\Sigma$  es una matriz semidefinida positiva.

**Demostración:**

Se sabe que la expresión

$$\left( \sum_{k=1}^n a_k (X_k - \mu_k) \right)^2 \geq 0$$

para todo  $a = (a_1, a_2, \dots, a_n)^T \in \mathbb{R}^n$ , por lo que

$$\begin{aligned} 0 &\leq \mathbb{E} \left[ \left( \sum_{k=1}^n a_k (X_k - \mu_k) \right)^2 \right] \\ &= \mathbb{E} \left[ \sum_{k=1}^n \sum_{j=1}^n a_k a_j (X_k - \mu_k)(X_j - \mu_j) \right] \\ &= \sum_{k=1}^n \sum_{j=1}^n a_k a_j \mathbb{E}[(X_k - \mu_k)(X_j - \mu_j)] \\ &= \sum_{k=1}^n \sum_{j=1}^n a_k a_j \sigma_{i,j} \\ &= Q_{\Sigma}(a). \end{aligned}$$

donde  $Q_{\Sigma}$  es la forma cuadrática asociada a  $\Sigma$ .

□

**Teorema A.3.1.1.** Sea  $X = (X_1, X_2, \dots, X_n)^T$  un vector aleatorio con vector de medias  $\mu$  y matriz de covarianzas  $\Sigma$ . Sea  $B = (b_{i,j})_{i=1:m, j=1:n}$  una matriz constante de tamaño  $m \times n$  y  $b = (b_1, b_2, \dots, b_m)^T$  un vector constante de dimensiones  $m \times 1$ . Definamos  $Y = BX + b$ . Entonces:

$$\mathbb{E}[Y] = B\mathbb{E}[X] + b \quad \text{y} \quad \text{Cov}(Y) = B\Sigma B^T.$$

**Demostración:**

Se tiene que  $Y = (Y_1, Y_2, \dots, Y_m)^T$  donde

$$\begin{aligned} Y_i &= \sum_{j=1}^n b_{i,j} X_j + b_j \\ \Rightarrow \\ \mathbb{E}(Y_i) &= \sum_{j=1}^n b_{i,j} \mathbb{E}[X_j] + b_j. \end{aligned}$$

Por lo tanto:

$$\mathbb{E}[Y] = B\mathbb{E}[X] + b = B\mu + b.$$

De manera análoga se puede demostrar que  $\mathbb{E}[X^T B^T] = \mathbb{E}[X^T] B^T$ .

Por otra parte, en cuanto a la covarianza de  $Y$  se tiene que

$$\begin{aligned} \text{Cov}(Y) &= \mathbb{E}[(Y - \mathbb{E}[Y])(Y - \mathbb{E}[Y])^T] \\ &= \mathbb{E}[(BX + b - B\mathbb{E}[X] - b)(BX + b - B\mathbb{E}[X] - b)^T] \\ &= \mathbb{E}[(BX - B\mathbb{E}[X])(BX - B\mathbb{E}[X])^T] \\ &= \mathbb{E}[B(X - \mathbb{E}[X])(B(X - \mathbb{E}[X]))^T] \\ &= \mathbb{E}[B(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T B^T] \\ &= B\mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T] B^T \\ &= B\text{Cov}(X)B^T = B\Sigma B^T. \end{aligned}$$

□

**Definición A.3.1.3.** Sean  $X = (X_1, X_2, \dots, X_n)^T$  y  $Y = (Y_1, Y_2, \dots, Y_m)^T$  dos vector aleatorios con vector de medias  $\mu_X$  y  $\mu_Y$  respectivamente, se define a la covarianza entre  $X$  y  $Y$  como  $\text{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)^T]$ .

Notemos que  $\text{Cov}(X, Y)$  es una matriz de  $n \times m$  tal que su componente  $\sigma_{i,j} = \text{Cov}(X_i, Y_j)$  con  $i = 1, 2, \dots, n$  y  $j = 1, 2, \dots, m$ .

**Definición A.3.1.4.** Decimos que un vector aleatorio  $X = (X_1, X_2, \dots, X_n)^T$  es un vector Gaussiano o tiene una distribución Normal multivariada ( $X \sim \mathcal{N}_n(\mu, \Sigma)$ , donde  $\mu = \mathbb{E}[X]$  y  $\Sigma = \text{Cov}(X)$ ) si cualquier combinación lineal de sus componentes resulta en una variable aleatoria con distribución Normal. Es decir, para toda  $A = (a_1, a_2, \dots, a_n) \in \mathbb{R}^n$  se tiene que  $AX$  distribuye Normal.

Cabe aclarar que debido a que a lo largo de este documento solo trabajaremos con el caso no degenerado de la distribución Normal multivariada, por esta razón, en la definición anterior así como en lo que resta de esta sección,  $\Sigma$  es una matriz definida positiva, y por ende  $|\Sigma| > 0$ .

**Proposición A.3.1.2.** Como consecuencia de esta definición es posible demostrar que si  $X = (X_1, X_2, \dots, X_n)^T$  es un vector gaussiano, entonces

- (a) Cada uno de sus componentes es una variable aleatoria Normal.
- (b) Cualquier subvector de  $X$  es un vector Gaussiano.
- (c) Si  $Y = (Y_1, Y_2, \dots, Y_n)^T$  y  $X = (X_1, X_2, \dots, X_n)^T$  son vectores Gaussianos independientes, entonces  $X + Y$  es también un vector Gaussiano.

**Demostración:**

Inciso(a):

Sea  $X_i$  la  $i$ -ésima entrada de  $X$ , definamos a la matriz  $A_i = ((a_1, a_2, \dots, a_n))$  tal que  $a_i = 1$  y  $a_j = 0$  para toda  $j \neq i$ . Como  $X$  es un vector Gaussiano, entonces,  $A_i X = X_i$  es un variable aleatoria con distribución Normal, por lo que se obtiene lo que se deseaba demostrar.

Inciso(b):

Sea  $i_1, i_2, \dots, i_m \subseteq 1, 2, \dots, n$  de manera muy sencilla se puede probar que el subvector de  $X$ ,  $\bar{X} = (X_{i_1}, X_{i_2}, \dots, X_{i_m})^T$ , es un vector Gaussiano. Para demostrar esto consideremos cualquier vector  $A \in \mathbb{R}^m$ . Definamos al vector  $C = (c_1, c_2, \dots, c_n)$  de la siguiente manera:

- Si  $j = i_k$  para alguna  $k \in \{1, 2, \dots, m\}$  entonces  $c_j = a_k$ .
- Si  $j \neq i_k$  para toda  $k \in \{1, 2, \dots, m\}$  entonces  $c_j = 0$ .

Es claro que  $A\bar{X} = \sum_{k=1}^m a_k X_{i_k} = \sum_{j=1}^n c_j X_j = CX$  como  $X$  es un vector Gaussiano esto significa que  $CX = A\bar{X}$  es una v.a con distribución normal por lo que se concluye que en efecto  $\bar{X}$  es un vector Gaussiano.

Inciso(c):

Sea  $A \in \mathbb{R}^n$ , se tiene que  $A(X + Y) = AX + AY$  como  $X$  y  $Y$  son vectores Gaussianos independientes entonces  $AX$  y  $AY$  son v.a. Normales independientes, esto implica  $A(X + Y) = AX + AY$  es una v.a Normal lo que significa  $X + Y$  es un vector Gaussiano. □

**Proposición A.3.1.3.** Sea  $X \sim \mathcal{N}_n(\mu, \Sigma)$ . Sea  $Y = BX + b$  donde  $B$  es una matriz constante de  $m \times n$  y  $b$  un vector constante de tamaño  $m$ . Entonces  $Y \sim \mathcal{N}_m(B\mu + b, B\Sigma B^T)$ .

**Demostración:**

Sea  $A = (a_1, a_2, \dots, a_m) \in \mathbb{R}^m$ .

$$AY = A(BX + b) = A(BX) + Ab = CX + c,$$

donde  $C = AB$  es un vector constante de tamaño  $1 \times n$  y  $c = Ab$  es una constante. Como  $X$  es un vector Gaussiano es claro que  $CX + c$  es una variable aleatoria con distribución Normal. Lo que significa que  $Y$  también es un vector Gaussiano.

Lo único que queda por hacer es calcular el vector de medias de  $Y$  así como su matriz de covarianzas. Sin embargo, esta tarea ya se realizó en el teorema anterior donde se encontró que efectivamente  $\mathbb{E}[Y] = BX + b$  y  $Cov(Y) = B\Sigma B^T$ . □

**Definición A.3.1.5.** Sea  $X = (X_1, X_2, \dots, X_n)^T$  un vector aleatorio. Definimos a su función generadora de momentos como:

$$M_X(t) = \mathbb{E}[e^{tX}],$$

donde  $t = (t_1, t_2, \dots, t_n) \in \mathbb{R}^n$ .

**Proposición A.3.1.4.**  $X \sim \mathcal{N}_n(\mu, \Sigma)$  si y sólo si función generadora de momentos es:

$$M_X(t) = \exp \left\{ t\mu + \frac{1}{2}t\Sigma t^T \right\}.$$

**Demostración :**

( $\implies$ ) Sabemos que  $X$  es un vector Gaussiano, esto significa que  $Z = tX$  es un variable aleatoria Normal con media  $t\mu$  y varianza  $t\Sigma t^T$ , esto significa que podemos escribir la función generadora de momentos de  $X$  en términos de  $Z$  ya que:

$$M_X(t) = \mathbb{E}[e^Z] = M_Z(1) = \exp \left\{ t\mu + \frac{1}{2}t\Sigma t^T \right\}.$$

( $\impliedby$ ) Se sabe que  $X$  es tal que  $M_X(t) = \exp \left\{ t\mu + \frac{1}{2}t\Sigma t^T \right\}$ . Sea  $A = (a_1, a_2, \dots, a_n) \in \mathbb{R}^n$ . Calculemos entonces la función generadora de momentos de la variable aleatoria  $AX$ . Consideremos  $t \in \mathbb{R}$ , se tiene entonces que

$$\begin{aligned}
M_{AX}(t) &= \mathbb{E}[e^{tAX}] = M_X(tA) \\
&= \left\{ tA\mu + \frac{1}{2}tA\Sigma(tA)^T \right\} \\
&= \left\{ tA\mu + \frac{1}{2}tA\Sigma A^T t^T \right\} \\
&= \left\{ tA\mu + \frac{1}{2}t^2 A\Sigma A^T \right\}.
\end{aligned}$$

Aquí podemos reconocer que  $M_{AX}(t)$  es la función generadora de momentos de una variable aleatoria con distribución  $\mathcal{N}(A\mu, A\Sigma A^T)$ . Por esto concluimos que  $AX$  distribuye Normal y por lo tanto  $X$  es un vector Gaussiano en el sentido de la definición A.3.1.4.  $\square$

Esta proposición nos dice que existe una definición equivalente de vector Gaussiano a la definición A.3.1.4, la cual se enunciará a continuación.

**Definición A.3.1.6.** *Se dice que  $X = (X_1, X_2, \dots, X_n)^T$  es un vector Gaussiano ó  $X \sim \mathcal{N}_n(\mu, \Sigma)$  si y sólo si la función generadora de momento de  $X$  está dada por:*

$$M_X(t) = \exp \left\{ t\mu + \frac{1}{2}t\Sigma t^T \right\},$$

donde  $t = (t_1, t_2, \dots, t_n) \in \mathbb{R}^n$ .

A continuación se presentara una tercera definición de vector Gaussiano junto con la prueba de que esta nueva definición es equivalente a las definiciones anteriores.

**Definición A.3.1.7.** *Se dice que  $X = (X_1, X_2, \dots, X_n)^T$  es un vector Gaussiano o  $X \sim \mathcal{N}_n(\mu, \Sigma)$  ( $\Sigma \geq 0$ ) si y sólo si  $X \stackrel{d}{=} \Lambda W + \mu$  donde  $W = (W_1, W_2, \dots, W_m)^T$  es un vector aleatorio tal que  $W_i \sim \mathcal{N}(0, 1)$  para toda  $i = 1, 2, \dots, m$  y todos sus componentes son independientes,  $\Lambda$  es una matriz constante de  $n \times m$  que cumple  $\Sigma = \Lambda\Lambda^T$  y  $\mu$  es un vector constante de  $n \times 1$ .*

**Proposición A.3.1.5.** *Si  $X$  es un vector Gaussiano en el sentido de la definición A.3.1.7, entonces  $X$  es un vector Gaussiano en el sentido de la definición A.3.1.4.*

**Demostración:**

Sea  $A = (a_1, a_2, \dots, a_n) \in \mathbb{R}^n$ , entonces

$$AX = A(\Lambda W + \mu) = (A\Lambda)W + A\mu = BW + b,$$

donde  $B = A\Lambda$  es un vector constante de  $1 \times m$  y  $b = A\mu$  es una constante.

Como las  $W_i$ 's son variables aleatorias Normales independientes, esto significa que  $BW + b$  es también una variable aleatoria Normal y por lo tanto  $AX$  resulta en una

variable aleatoria Normal, esto significa que  $X$  es un vector Gaussiano en el sentido de la definición A.3.1.4. □

**Proposición A.3.1.6.** *Sea  $X$  un vector Gaussiano en el sentido de la definición A.3.1.6, entonces  $X$  es un vector Gaussiano en el sentido de la definición A.3.1.7.*

**Demostración:**

Se sabe que  $M_X(t) = \exp\left\{t\mu + \frac{1}{2}t\Sigma t^T\right\}$ . Como  $\Sigma$  es una matriz semidefinida positiva, se sabe que existe una matriz  $\Lambda$  de  $n \times m$  (posiblemente  $m = n$ ) tal que  $\Lambda\Lambda^T = \Sigma$ , definamos entonces  $W = (W_1, W_2, \dots, W_m)^T$  donde las  $W_i$ 's son variables aleatorias independientes con distribución Normal estándar, es decir  $W_i \sim \mathcal{N}(0, 1)$  para toda  $i$ . A continuación calcularemos la función generadora de momentos de  $Z = \Lambda W + \mu$ .

Para esto primero notemos que  $\mathbb{E}[W] = (\mathbb{E}[W_1], \mathbb{E}[W_2], \dots, \mathbb{E}[W_m])$  donde  $\mathbb{E}[W_i] = 0$  para toda  $i$ , en adición, como las  $W_i$ 's son independientes entonces  $Cov(W_i, W_j) = 0$  cuando  $i \neq j$  y  $Cov(W_i, W_i) = Var(W_i) = 1$  por lo que  $Cov(W) = I_n$ <sup>22</sup>.

Antes de calcular la función generadora de momentos de  $Z$  calculemos la de  $W$ . Sea  $u = (u_1, u_2, \dots, u_m) \in \mathbb{R}^m$  entonces:

$$\begin{aligned} M_W(u) &= \mathbb{E}[e^{uW}] = \mathbb{E}\left[\exp\left\{\sum_{i=1}^m u_i W_i\right\}\right] \\ &= \mathbb{E}\left[\prod_{i=1}^m \exp\{u_i W_i\}\right] \\ &= \prod_{i=1}^m \mathbb{E}[\exp\{u_i W_i\}] \\ &= \prod_{i=1}^m M_{W_i}(u_i) \\ &= \prod_{i=1}^m \exp\left\{\frac{1}{2}u_i^2\right\} \\ &= \exp\left\{\frac{1}{2}\sum_{i=1}^m u_i^2\right\} \\ &= \exp\left\{\frac{1}{2}uu^T\right\}. \end{aligned}$$

Sea  $t = (t_1, t_2, \dots, t_n) \in \mathbb{R}^n$ , calculemos ahora  $M_Z(t)$ .

---

<sup>22</sup> $I_n$  se refiere a la matriz identidad de tamaño  $n \times n$

$$\begin{aligned}
M_Z(t) &= \mathbb{E}[e^{tZ}] = \mathbb{E}[e^{t(\Lambda W + \mu)}] \\
&= \mathbb{E}[e^{t\Lambda W} e^{t\mu}] \\
&= e^{t\mu} \mathbb{E}[e^{t\Lambda W}] \\
&= e^{t\mu} M_W(t\Lambda) \\
&= e^{t\mu} \exp \left\{ \frac{1}{2} t\Lambda(t\Lambda)^T \right\} \\
&= \exp \left\{ t\mu + \frac{1}{2} t\Lambda\Lambda^T t^T \right\} \\
&= \exp \left\{ t\mu + \frac{1}{2} t\Sigma t^T \right\}.
\end{aligned}$$

Como  $M_X(t) = M_Z(t)$  esto significa que  $X$  y  $Z$  tienen la misma distribución, lo cual significa que  $X$  se puede expresar como  $BW + b$  donde  $B = \Lambda$ ,  $b = \mu$  y  $W$  es un vector aleatorio formado por v.a.i.i.d con distribución normal estándar.

□.

Antes de concluir con esta parte del documento se presentara una cuarta y última definición de vector aleatorio Gaussiano, esta definición habla de la función de densidad conjunta de este mismo.

**Definición A.3.1.8.** Sea  $X = (X_1, X_2, \dots, X_n)^T$  un vector aleatorio, entonces  $X \sim \mathcal{N}_n(\mu, \Sigma)$  (con  $\Sigma > 0$ ) si y sólo si su función de densidad conjunta es:

$$f_X(x) = \left( \frac{1}{\sqrt{2\pi}|\Sigma|} \right)^n \exp \left\{ -\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu) \right\}.$$

Es claro que si  $\Sigma$  es una matriz definida positiva entonces  $|\Sigma| > 0$ .

Notemos que esta última definición solamente aplica para el caso en el que la matriz de covarianza es una matriz definida positiva.

**Proposición A.3.1.7.** Si  $X$  es un vector Gaussiano en el sentido de la definición A.3.1.7, entonces  $X$  es un vector Gaussiano en el sentido de la definición A.3.1.8 siempre y cuando  $\Sigma > 0$ .

**Demostración:**

Se sabe que si  $X \sim \mathcal{N}_n(\mu, \Sigma)$  en el sentido de la definición A.3.1.7 entonces  $X \stackrel{d}{=} \Lambda W + \mu$  donde  $W = (W_1, W_2, \dots, W_m)^T$  es un vector de v.a.i.i.d. tal que para toda  $i$   $W_i \sim \mathcal{N}(0, 1)$ . Esto significa que podemos escribir  $W = \Lambda^{-1}(X - \mu)$ <sup>23</sup>.

---

<sup>23</sup>La prueba de que  $\Lambda$  tiene inversa por la izquierda se puede revisar en la sección de anexos en la parte de matrices

Primero calculemos la función de densidad conjunta de  $W$ , lo cual es muy sencillo ya que los componentes de  $W$  son independientes.

$$\begin{aligned} f_W(w) &= \prod_{i=1}^m f_{W_i}(w_i) \\ &= \prod_{i=1}^m \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{w_i^2}{2}\right\} \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^m \exp\left\{-\frac{w^T w}{2}\right\}. \end{aligned}$$

Además, como estamos en el caso en que  $\Sigma$  es positiva definida entonces  $\Sigma = \Lambda\Lambda^T$  donde  $\Lambda$  es una matriz de  $n \times n$  por lo que se sabe que  $m = n$ , de forma que

$$f_W(w) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left\{-\frac{w^T w}{2}\right\}.$$

Consideremos a  $W$  como función de  $X$  de tal manera que para toda  $x \in \mathbb{R}^n$ ,  $w(x) = \Lambda^{-1}(x - \mu) = (w_1, w_2, \dots, w_n)$ , donde  $\Lambda^{-1} = (\lambda_{ij}^{-1})_{i=1:n, j=1:n}$ . Claramente  $w_i = \sum_{j=1}^n \lambda_{ij}^{-1}(x_j - \mu_j)$ .

Como una consecuencia del teorema de cambio de variable se sabe que:

$$\begin{aligned} f_X(x) &= f_W(w(x)) |J_w(x)| \\ &= f_W(\Lambda^{-1}(x - \mu)) \left| \left( \frac{\delta w_i}{\delta x_j} \right)_{i=1:n, j=1:n} \right| \\ &= f_W(\Lambda^{-1}(x - \mu)) \left| (\lambda_{ij}^{-1})_{i=1:n, j=1:n} \right| \\ &= f_W(\Lambda^{-1}(x - \mu)) |\Lambda^{-1}| \\ &= \left(\sqrt{2\pi}\right)^{-n/2} \exp\left\{-\frac{(\Lambda^{-1}(x - \mu))^T (\Lambda^{-1}(x - \mu))}{2}\right\} |\Lambda^{-1}| \\ &= \left(\sqrt{2\pi}\right)^{-n/2} \exp\left\{-\frac{(x - \mu)^T (\Lambda^{-1})^T \Lambda^{-1} (x - \mu)}{2}\right\} |\Lambda\Lambda^T|^{-1/2} \\ &= \left(\sqrt{2\pi}\right)^{-n/2} \exp\left\{-\frac{(x - \mu)^T (\Lambda\Lambda^T)^{-1} (x - \mu)}{2}\right\} |\Sigma|^{-1/2} \\ &= \left(\sqrt{2\pi}\right)^{-n/2} |\Sigma|^{-1/2} \exp\left\{-\frac{(x - \mu)^T (\Sigma)^{-1} (x - \mu)}{2}\right\}, \end{aligned}$$

donde  $J_w(x)$  es el Jacobiano de  $W$  con respecto a  $x$ .

□

**Proposición A.3.1.8.** Si  $X = (X_1, X_2, \dots, X_n)^T$  es un vector Gaussiano en el sentido de la definición A.3.1.8, entonces su función generadora de momentos es

$$M_X(t) = \exp \left\{ t\mu + \frac{1}{2}t\Sigma t^T \right\},$$

donde  $t = (t_1, t_2, \dots, t_n) \in \mathbb{R}^n$  y  $\Sigma > 0$ .

**Demostración:**

$$\begin{aligned} M_X(t) &= \mathbb{E}[e^{tX}] \\ &= \int_{\mathbb{R}^n} \exp\{tx\} \left( \frac{1}{\sqrt{2\pi|\Sigma|}} \right)^n \exp \left\{ -\frac{(x-\mu)^T \Sigma^{-1} (x-\mu)}{2} \right\} dx_1 dx_2 \dots dx_n \\ &= \int_{\mathbb{R}^n} \left( \frac{1}{\sqrt{2\pi|\Sigma|}} \right)^n \exp \left\{ -\frac{x^T \Sigma^{-1} x - 2\mu^T \Sigma^{-1} x + \mu^T \Sigma^{-1} \mu + 2tx}{2} \right\} dx_1 dx_2 \dots dx_n \\ &= \exp \left\{ -\frac{\mu^T \Sigma^{-1} \mu}{2} \right\} \int_{\mathbb{R}^n} \left( \frac{1}{\sqrt{2\pi|\Sigma|}} \right)^n \exp \left\{ -\frac{x^T \Sigma^{-1} x - 2(\mu^T + t\Sigma)\Sigma^{-1} x}{2} \right\} dx_1 dx_2 \dots dx_n \\ &= \exp \left\{ -\frac{\mu^T \Sigma^{-1} \mu}{2} + \frac{(\mu + t\Sigma)\Sigma^{-1}(\mu + t\Sigma)^T}{2} \right\} \times \dots \\ &\quad \dots \times \int_{\mathbb{R}^n} \left( \frac{1}{\sqrt{2\pi|\Sigma|}} \right)^n \exp \left\{ -\frac{(x - (\mu + t\Sigma)^T)^T \Sigma^{-1} (x - (\mu + t\Sigma)^T)}{2} \right\} dx_1 dx_2 \dots dx_n \\ &= \exp \left\{ \frac{-\mu^T \Sigma^{-1} \mu + \mu^T \Sigma^{-1} \mu + t\mu + \mu^T t^T + t\Sigma t^T}{2} \right\} \\ &= \exp \left\{ \frac{2t\mu + t\Sigma t^T}{2} \right\} \\ &= \exp \left\{ t\mu + \frac{t\Sigma t^T}{2} \right\}. \end{aligned}$$

□

Con esto hemos demostrado que las distintas definiciones de vector Gaussiano que hemos propuesto son equivalentes. A continuación presentemos un par de propiedades más que tienen los vectores Gaussianos antes de dar por concluida esta sección.

**Proposición A.3.1.9.** Sean  $X = (X_1, X_2, \dots, X_n)^T$  y  $Y = (Y_1, Y_2, \dots, Y_n)^T$  dos vectores Gaussianos con vectores de medias  $\mu_X$  y  $\mu_Y$  y matrices de covarianza  $\Sigma_X$  y  $\Sigma_Y$  respectivamente. Entonces  $X$  y  $Y$  son independientes si y sólo si  $X$  y  $Y$  son no correlacionadas (es decir  $\Sigma_{XY} = \text{Cov}(X, Y) = 0 = \text{Cov}(Y, X)^T = \Sigma_{YX}^T$ ).

**Demostración:**

( $\implies$ )

Supongamos que  $X$  y  $Y$  son independientes. Esto significa que

$$\begin{aligned} \text{Cov}(X, Y) &= \mathbb{E}[(X - \mu_X)(Y - \mu_Y)^T] \\ &= \mathbb{E}[(X - \mu_X)]\mathbb{E}[(Y - \mu_Y)^T] = 0. \end{aligned}$$

( $\Leftarrow$ )

Supongamos que  $\Sigma_{XY} = \text{Cov}(X, Y) = 0 = \text{Cov}(Y, X)^T = \Sigma_{YX}^T$ . Sea

$$Z = \begin{pmatrix} X \\ Y \end{pmatrix}$$

Es claro entonces que la distribución conjunta de  $X$  y  $Y$  es gaussiana con vector de medias

$$\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}$$

y matriz de covarianza

$$\Sigma = \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{pmatrix} = \begin{pmatrix} \Sigma_X & 0 \\ 0 & \Sigma_Y \end{pmatrix}.$$

Con esto podemos escribir a  $(Z - \mu)^T \Sigma^{-1} (Z - \mu)$  como

$$\begin{aligned} &(X^T - \mu_X^T, Y^T - \mu_Y^T) \begin{pmatrix} \Sigma_X & 0 \\ 0 & \Sigma_Y \end{pmatrix}^{-1} \begin{pmatrix} X - \mu_X \\ Y - \mu_Y \end{pmatrix} \\ &= (X^T - \mu_X^T, Y^T - \mu_Y^T) \begin{pmatrix} \Sigma_X^{-1} & 0 \\ 0 & \Sigma_Y^{-1} \end{pmatrix} \begin{pmatrix} X - \mu_X \\ Y - \mu_Y \end{pmatrix}. \end{aligned}$$

Lo que implica que

$$\begin{aligned} &\exp\{(Z - \mu)^T \Sigma^{-1} (Z - \mu)\} \\ &= \exp\{(X - \mu_X)^T \Sigma_X^{-1} (X - \mu_X) + (Y - \mu_Y)^T \Sigma_Y^{-1} (Y - \mu_Y)\} \\ &= \exp\{(X - \mu_X)^T \Sigma_X^{-1} (X - \mu_X)\} \exp\{(Y - \mu_Y)^T \Sigma_Y^{-1} (Y - \mu_Y)\}. \end{aligned}$$

Con esto se concluye que la función de densidad conjunta se puede escribir como

$$f(Z) = g(X)h(Y),$$

lo que se traduce en que  $X$  y  $Y$  son independientes.

**Proposición A.3.1.10.** Sean  $X$  y  $Y$  dos vectores aleatorios:

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} \quad y \quad Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_q \end{pmatrix}$$

Entonces la distribución conjunta de  $Z = (X, Y)^T$  es Gaussiana si y sólo si

- (i)  $X$  tiene una distribución Gaussiana, y  
(ii) La distribución condicional de  $Y|X$  es Gaussiana.

**Demostración:**

Sea

$$Z = \begin{pmatrix} X \\ Y \end{pmatrix} \quad \text{donde} \quad X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} \quad \text{y} \quad Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_q \end{pmatrix}$$

con vector de medias

$$\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} = \begin{pmatrix} \mu_{X_1} \\ \vdots \\ \mu_{X_p} \\ \mu_{Y_1} \\ \vdots \\ \mu_{Y_q} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_{p+q} \end{pmatrix}$$

Y matriz de covarianza

$$\Sigma = \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{pmatrix},$$

donde

- $\mu_X = (\mu_{X_1}, \mu_{X_2}, \dots, \mu_{X_p})$  es el vector de medias de  $X$ ,
- $\mu_Y = (\mu_{Y_1}, \mu_{Y_2}, \dots, \mu_{Y_q})$  es el vector de medias de  $Y$ ,
- $\Sigma_X$  es la matriz de covarianza de  $X$ ,
- $\Sigma_Y$  es la matriz de covarianza de  $Y$  y
- $\Sigma_{XY} = Cov(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)^T] = Cov(Y, X)^T = \Sigma_{XY}^T$

Definamos al vector  $\hat{Y} = Y - \Sigma_{YX}\Sigma_X^{-1}X$ . Es fácil ver que  $X$  y  $\hat{Y}$  son independientes ya que

$$\begin{aligned} Cov(X, \hat{Y}) &= Cov(X, Y - \Sigma_{YX}\Sigma_X^{-1}X) \\ &= Cov(X, Y) - Cov(X, X)(\Sigma_{YX}\Sigma_X^{-1})^T \\ &= Cov(X, Y) - Cov(X, X)\Sigma_X^{-1}\Sigma_{XY} \\ &= Cov(X, Y) - \Sigma_{XY} \\ &= 0, \end{aligned}$$

además

$$\tilde{Z} = \begin{pmatrix} X \\ \hat{Y} \end{pmatrix} = \begin{pmatrix} I_p & 0_{pxq} \\ -\Sigma_{YX}\Sigma_X^{-1} & I_q \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} = AZ$$

donde  $A$  es una matriz invertible tal que su inversa es

$$A^{-1} = \begin{pmatrix} I_p & 0_{pxq} \\ \Sigma_{YX}\Sigma_X^{-1} & I_q \end{pmatrix},$$

por lo que

$$Z = \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} I_p & 0_{pxq} \\ \Sigma_{YX}\Sigma_X^{-1} & I_q \end{pmatrix} \begin{pmatrix} X \\ \hat{Y} \end{pmatrix} = A^{-1}\tilde{Z}.$$

Ahora, como  $A$  y  $A^{-1}$  son matrices constantes, por propiedades de vector Gaussiano se tiene que  $Z$  es vector Gaussiano  $\iff \tilde{Z}$  es vector Gaussiano.

Además  $Y|X$  es vector Gaussiano  $\iff Y|X + (\Sigma_{YX}\Sigma_X^{-1}X)|X$  es un vector Gaussiano<sup>24</sup>  $\iff \hat{Y}|X = \hat{Y}$  es un vector Gaussiano.

Por esto basta demostrar que  $\tilde{Z}$  es un vector gaussino  $\iff X$  y  $\hat{Y}$  son vectores Gaussianos.

( $\implies$ )

Como  $\tilde{Z}$  es un vector gaussino entonces todos sus subvectores tienen distribución Gaussiana, lo que significa que en particular  $X$  y  $\hat{Y} = \hat{Y}|X$  son vectores Gaussianos.

( $\impliedby$ )

Sea  $A = (A_1, A_2) \in \mathbb{R}^{p+q}$  con  $A_1 \in \mathbb{R}^p$  y  $A_2 \in \mathbb{R}^q$ . Entonces

$$A\tilde{Z} = \begin{pmatrix} A_1 & A_2 \end{pmatrix} \begin{pmatrix} X \\ \hat{Y} \end{pmatrix} = A_1X + A_2\hat{Y}.$$

Como  $A_1X$  y  $A_2\hat{Y}$  son variables aleatorias Normales independientes entonces  $A\tilde{Z}$  es también una variable aleatoria Normal por lo que  $\tilde{Z}$  es un vector Gaussiano. □

**Definición A.3.1.9.** Decimos que  $\{X_t\}_{t=1}^{\infty}$  es un proceso Gaussiano si todas sus distribuciones finito dimensionales son gaussianas. Esto es  $(X_{t_1}, X_{t_2}, \dots, X_{t_n})$  es un vector Gaussiano para cualesquiera  $t_1 < t_2 < \dots < t_n \in \mathbb{N}$ .

---

<sup>24</sup>esto porque que  $(\Sigma_{YX}\Sigma_X^{-1}X)|X$  es un vector constante de  $qx1$

## Referencias

- [1] Giovanni Petris, Sonia Petrone and. Patrizia Campagnoli (2009). *Dynamic Linear Models with R*. Springer.
- [2] James O. Berger (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer - Verlag.
- [3] Giovanni Petris (2009). *dln: an R package for Bayesian analysis of Dynamic Linear Models*.
- [4] James D. Hamilton (1994). *Handbook of Econometrics*. Volumen IV. Elsevier.
- [5] Allan Gut (2009). *An Intermediate Course in Probability*. Springer.
- [6] Peter J. Brockwell and. Richard A. Davis (1996). *Introduction to Time Series and Forecasting*. Springer - Verlag.
- [7] William W. S. Wei (2006). *Time Series Analysis*. Pearson - Addison Wesley.
- [8] J. A. Bondy and U. S. R. Murty (1976). *Graph Theory with Applications*. Elsevier.
- [9] Stephen H. Friedberg, Arnold J. Insel and Lawrence E. Spence (1979). *Linear Algebra*. Prentice Hall.