Universitat Autònoma de Barcelona

Avances en el análisis de series temporales categóricas con aplicaciones

SERGIO BAENA MIRABETE

Tesis Doctoral

Avances en el análisis de series temporales categóricas con aplicaciones

SERGIO BAENA MIRABETE

Supervisado por PERE PUIG CASADO

Departamento de Matemáticas

Doctorado en Matemáticas

2019-2020

Universitat Autònoma
de Barcelona

A few years ago the city council of Monza, Italy, barred pet owners from keeping goldfish in curved goldfish bowls. The measure's sponsor explained the measure in part by saying that it is cruel to keep a fish in a bowl with curved sides because, gazing out, the fish would have a distorted view of reality. But how do we know we have the true, undistorted picture of reality? Might not we ourselves also be inside some big goldfish bowl and have our vision distorted by an enormous lens? The goldfish's picture of reality is different from ours, but can we be sure it is less real?.

Stephen Hawking & Leonard Mlodinow, *The Grand Design*.

Agradecimientos

En primer lugar, me gustaría empezar expresando mi agradecimiento a mi tutor, el Dr. Pere Puig, por hacer de esta travesía una experiencia tan agradable.

Agradecer también a Anabel Blasco por la revisión que ha hecho de cada uno de los artículos que componen esta tesis y sus útiles comentarios. Mil gracias.

También me gustaría agradecer a Amanda Fernández por ayudarme en todo el proceso final de depósito de la tesis. Me ha resultado de gran ayuda.

Finalmente, me gustaría dedicar esta tesis, con todo mi cariño, a mis hermanos Sandra y Dani, a mis sobrinos Aitor y Roger y, cómo no, a mis padres, Antonio e Isabel, por estar siempre ahi apoyándome. De todos ellos he aprendido muchísimas cosas, entre ellas, la constancia, el esfuerzo y el placer de aprender.

Sergio Baena Mirabete
Barcelona, Abril 2020

Índice general

Introducción

Una cadena de Markov[1] (o modelo de Markov) de orden k ($1 \leq k < \infty$), tomando valores en un conjunto finito de estados $A = \{0, 1, ..., m-1\}$, $m \geq 2$, es un modelo probabilístico muy utilizado para describir series de memoria corta. Sin embargo, los modelos de Markov en el contexto de series temporales categóricas pueden ser problemáticos debido a que el número de parámetros a estimar aumenta rápidamente con el orden k: una cadena con m estados tiene $m^k(m-1)$ parámetros a ser estimados. Numerosos estudios han centrado el interés en este hecho aportando algunas soluciones para reducir el número de parámetros. Raftery (1985) propuso el modelo MTD[2] para aproximar cadenas de Markov de mayor orden con un menor número de parámetros. Fokianos & Kedem (2003) mostraron la flexibilidad de los métodos de regresión para tratar con series temporales categóricas, permitiendo así reducir el número de parámetros. Para procesos más complicados, una combinación de más de una cadena de Markov puede ser una interesante alternativa. Modelos de mixturas de cadenas de Markov han sido propuestos en la literatura desde diferentes enfoques. En Park & Basawa (2002), los autores propusieron un modelo de mixtura de cadenas de Markov en el cual, condicionado a una variable no observable

[1] Recibe su nombre del célebre matemático ruso Andréi Márkov (1856-1922), que lo introdujo en 1913 en un estudio donde analiza la frecuencia con la que aparecen las vocales en un texto extraído de la novela de Aleksándr Pushkin titulada *Eugene Onegin*. Una traducción al inglés de dicho estudio puede consultarse en Markov (2006).

[2] Por sus siglas del inglés, *Mixture Transition Distribution model.*

(latente), las variables de respuesta seguían un proceso de Markov y el proceso latente se asumía independiente e idénticamente distribuido. Otro enfoque muy utilizado en la literatura son los llamados Modelos Latentes de Markov[3] que también asumen la existencia de un proceso no observable que afecta la distribución de las variables de respuesta. En este caso, el proceso latente sigue una cadena de Markov y, dado este proceso, las variables de respuesta se asumen que son condicionalmente independientes. Desde este último enfoque, un estudio interesante se describe en Albert (1991) donde se propuso una mixtura de cadenas de Markov para modelizar la serie temporal de recuentos de ataques epilépticos. Un estudio similar se describe en Leroux & Puterman (1992) cuyos autores aplicaron un modelo de Markov basado en una mixtura de distribuciones de Poisson para modelizar el número de movimientos fetales. Vermunt et al. (1999) propusieron una versión más flexible donde las probabilidades iniciales y las probabilidades de transición del proceso latente podían depender de covariables. En esta línea, Bartolucci et al. (2009) utilizaron un Modelo Latente de Markov con covariables para evaluar los diferentes niveles de salud, no observables directamente, de ancianos residentes en una clínica durante un cierto periodo de tiempo. Jackson et al. (2013) y Jackson et al. (2015), en un estudio longitudinal sobre conductores noveles, propusieron el uso de Modelos Latentes de Markov para relacionar la conducción temeraria con la ocurrencia de un accidente. En otro interesante estudio, Bartolucci et al. (2014a) propusieron un modelo autoregresivo basado también en procesos latentes para analizar aspectos relativos a la salud de personas jubiladas. Sin embargo, en algunas aplicaciones es poco realista asumir que el proceso bajo estudio es el mismo para todos los individuos. En este sentido, Van de Pol & Langeheine (1990) propusieron el llamado Modelo Mixto Latente de Markov[4] donde se estiman parámetros específicos para diferentes subpoblaciones (no observables directamente) de individuos. Un ejemplo práctico de cómo estimar un Modelo Mixto Latente de Markov usando el paquete LMest de R (R Core Team, 2017) puede consultarse en Bartolucci et al. (2017).

[3]En inglés, *Latent Markov Model*, que son modelos ocultos de Markov (del inglés, *Hidden Markov Models*) pero adaptados al contexto de datos longitudinales y que constituyen otro enfoque ampliamente utilizado en la literatura. Un Modelo Latente de Markov puede ser considerado como una generalización de un modelo de mixtura donde las variables latentes, que controlan la componente de la mixtura a seleccionar para cada observación, no son independientes sino que siguen un proceso de Markov.

[4]Del inglés, *Mixed Latent Markov Model*.

Una extensión de este modelo fue propuesta por Altman (2007), en la cual se incorporan covariables y efectos aleatorios para tener en cuenta diferencias entre individuos pero con un menor número de parámetros. En los últimos años, la investigación teórica y práctica sobre estos modelos ha ido apareciendo con mayor frecuencia en la literatura científica. Un resumen detallado sobre la misma aparece en la Sección 1.2 del libro de Bartolucci et al. (2013).

La primera y segunda parte de la presente tesis tiene como objetivo proponer una clase de modelos relacionados con la literatura previamente descrita. Concretamente, estos modelos se describen en dos artículos que conforman esta tesis: Baena-Mirabete & Puig (2018) y Baena-Mirabete et al. (2019).

En Baena-Mirabete & Puig (2018) proponemos una clase de modelos multiestados, para estados ordenados, basados en cadenas de Markov de alto orden. Estos modelos propuestos son especialmente apropiados para cadenas de alto orden donde las transiciones entre estados son poco frecuentes y cuando se producen, tienen lugar, generalmente, entre estados cercanos. En este tipo de procesos, las estimaciones de cada una de las celdas de la matriz de transición fuera de la diagonal son poco fiables debido al bajo número de observaciones en cada una de ellas. Se sugiere aquí algunas soluciones prácticas para reducir el número de parámetros y resolver los problemas de estimación en este tipo de procesos. Estos modelos son ilustrados con una aplicación en el contexto financiero, concretamente en la medición del riesgo de crédito. El principal objetivo de las agencias de calificación, tales como Moody's, Standard & Poor's (S&P) o Fitch, es evaluar la capacidad de un emisor de deuda, por ejemplo, una entidad financiera o un ente público, para cumplir con sus pagos. Para ello, las agencias de calificación proporcionan una escala de riesgo que clasifica la deuda en diferentes categorías (*rating*) según la solvencia del emisor. Estas categorías asignadas por estas agencias se expresan en una escala ordinal, donde la posición en la escala refleja la capacidad del emisor para hacer frente a sus obligaciones. Por ejemplo, Standard & Poor's puntúa en una escala que va de AAA a D, donde AAA se corresponde con la mayor puntuación y D con la menor. El estudio de las transiciones de *ratings* de una cierta cartera de inversión a lo largo del tiempo es crucial para la mayoría de las instituciones financieras. Consideremos por ejemplo, para una cierta entidad financiera, una cartera de inversión formada por títulos de bonos calificados cada uno de

ellos con un cierto *rating*. La evolución de dicha cartera, por ejemplo en el transcurso del año, hacia calificaciones más bajas no cabe duda que tendría un impacto negativo en el valor de dicha cartera. Esto es así porque el precio de los bonos no depende solo del tipo de interés sino también del riesgo de incumplimiento (Kliger & Sarig, 2000). En este caso, la entidad financiera deberá anticipar, generalmente a un año vista, la posible pérdida en dicha cartera y para ello es imprescindible disponer de estimaciones fiables de las probabilidades de transición entre *ratings*.

En Baena-Mirabete et al. (2019), proponemos diferentes modelos de mixturas de cadenas de Markov de alto orden para datos binarios. Estos modelos de mixturas han sido analizados desde diferentes enfoques. En primer lugar, se analizan Modelos Latentes de Markov donde las probabilidades de transición del proceso latente son modelizadas a partir de los retardos de las variables observadas, similar a como se describe en Bartolucci & Pennoni (2007). Segundo, se presenta una adaptación de los modelos descritos en Bartolucci & Farcomeni (2010) al caso de datos longitudinales y variable respuesta binaria. Como veremos, estos últimos modelos son una generalización del modelo MTD de Raftery (1985). Finalmente, en un tercer enfoque, se propone un modelo de mixtura de cadenas de Markov en el cual se tiene en cuenta una variabilidad adicional a nivel individuo. Este último modelo extiende el previamente citado trabajo de Park & Basawa (2002) al caso de datos longitudinales. Estos modelos derivan de un estudio observacional realizado a un grupo de estudiantes entre 18 y 20 años a quienes les propusimos replicar mentalmente las realizaciones de una moneda justa de la forma más fidedigna posible. Después de una sesión informativa, cada estudiante debía producir una serie de 50 resultados de cara y cruz pulsando, repetidamente, entre dos opciones posibles mostradas en la pantalla de su ordenador: 'CARA' y 'CRUZ'. En el transcurso del experimento, los estudiantes no podían ver sus resultados previos simulados. La generación de secuencias *aleatorias* mentalmente producidas por individuos es una tarea muy común en el campo de la psicología experimental. Tradicionalmente, este tipo de estudios se ha realizado sobre adultos sanos, generalmente estudiantes universitarios. Sin embargo, existen también algunos estudios al respecto realizados sobre pacientes con algún tipo de trastorno neurológico como por ejemplo, pacientes con síndrome disejecutivo (Spatt & Goldenberg, 1993), esquizofrenia (Salamé & Danion, 2007) o Alzheimer (Brugger et al., 1996). Inclu-

so en el caso del Alzheimer, las tareas de generación aleatoria han sido propuestas para la detección precoz de la enfermedad (García-Viedma et al., 2015). Todos estos estudios han mostrado que las secuencias producidas por pacientes con algún tipo de daño cerebral son más estereotipadas que aquellas generadas por el grupo de control. Por tanto, el análisis de este tipo de datos, más allá del interés que despierta en sí mismo, es de gran relevancia en el campo de la neuropsicología y otros campos relacionados.

La tercera parte de la tesis tiene como objetivo enunciar un resultado teórico basado en el cálculo de recurrencias para la obtención de las probabilidades de una gran clase de variables aleatorias tomando valores en $\mathbb{Z}$ (incluyendo tanto enteros positivos como negativos). Este resultado extiende el descrito en Sundt (1992) al incluir también distribuciones de recuento con soporte en los enteros negativos. Este enunciado y algunas aplicaciones relevantes se describen en Baena-Mirabete & Puig (2020), que constituye el tercer y último artículo que conforma la presente tesis.

Esta tesis está organizada en cinco capítulos. El Capítulo 1 constituye un resumen de los principales resultados metodológicos y aplicaciones que se presentan en cada uno de los artículos que forman parte de la tesis. El Capítulo 2 propone una conclusión definitiva basada en los alcances de esta tesis y algunas líneas de investigación futura. Finalmente, los Capítulos 3, 4 y 5 incluyen los manuscritos de los artículos mencionados previamente.

La presente tesis fue parcialmente financiada por el proyecto MTM2015-69493-R del Ministerio de Economía y Competitividad de España.

CAPÍTULO 1

Presentación global de los resultados y conclusiones

Este capítulo tiene como objetivo presentar y discutir los principales resultados de los artículos que conforman la presente tesis, previamente mencionados. Concretamente, el capítulo consiste de las siguientes cuatro secciones: (1.1) una sección donde se describen los modelos de Markov de alto orden para estados persistentes propuestos en Baena-Mirabete & Puig (2018) y aplicados en el contexto financiero de la medición del riesgo de crédito; (1.2) una sección donde se presentan diferentes modelos de mixturas de cadenas de Markov para datos binarios descritos en Baena-Mirabete et al. (2019), los cuales derivan de un estudio observacional en el que se analiza la aleatoriedad de secuencias generadas por un grupo de estudiantes que participaron en dicho estudio; (1.3) una sección donde se enuncia un resultado teórico basado en el cálculo de recurrencias para la obtención de las probabilidades de una extensa familia de distribuciones de recuento con soporte en $\mathbb{Z}$; y finalmente (1.4) una última sección donde se presentan los principales resultados y conclusiones de la tesis.

1.1

Modelos multiestados basados en cadenas de Markov de alto orden

En modelos basados en cadenas de Markov de orden k con m estados posibles, el número de parámetros a estimar aumenta rápidamente con el orden k de la cadena, concretamente, $m^k(m-1)$ parámetros. Este problema se acentúa cuando ciertas secuencias de estados son poco frecuentes, lo que conlleva a estimaciones poco fiables de las probabilidades de transición entre estos estados.

En el contexto del riesgo de crédito, por ejemplo, las transiciones entre *ratings* muy distantes entre ellos son muy poco comunes: imagínese una empresa puntuada en la mejor categoría crediticia (*rating* AAA); la probabilidad que en el transcurso de un año evolucione hacia la peor puntuación crediticia (*rating* C o incluso *rating default* D) es muy pequeña.[1] En este contexto, existen numerosos estudios sobre el análisis de las transiciones entre *ratings*, especialmente de bonos corporativos. Nickell et al. (2000) propuso una cadena de Markov de orden 1 para modelizar las transiciones entre *ratings*. Sin embargo, la investigación en este ámbito ha evolucionado hacia un consenso general: el conocimiento del *rating* previo es insuficiente para explicar la dinámica de las migraciones de *ratings*.[2] Esto nos lleva a la necesidad de explorar cadenas de mayor orden para analizar este tipo de procesos.

En esta sección, se presentan modelos multiestados basados en cadenas de Markov de alto orden. Estos modelos son aplicados a la medición del riesgo de crédito. Para un

[1] En riesgo de crédito, la probabilidad que un cierto activo permanezca en el mismo estado sin cambiar de *rating*, en un cierto horizonte de tiempo (p.ej. 1 año), es más grande que la probabilidad de moverse a cualquier otro estado. Esto da lugar a cadenas donde los diferentes estados tienden a repetirse, de forma consecutiva, a través del tiempo. Nos referiremos a este tipo de secuencias como cadenas de larga *persistencia* o cadenas *persistentes*, indiferentemente.

[2] Lando & Skϕdeberg (2002) constituye una interesante contribución a la literatura en este ámbito y, principalmente, dos conclusiones son extraídas de este estudio: la primera es conocida en la literatura como *rating momentum*, que significa que un previo empeoramiento del *rating* aumenta la probabilidad de empeorar nuevamente su *rating* y una mejora previa aumenta la probabilidad de mejorar su *rating*. La segunda conclusión tiene que ver con la relación entre el tiempo desde la última entrada en su actual *rating* y la probabilidad de transición. Los autores encuentran que cuanto más tiempo permanece sin cambiar de *rating* menos probable es un cambio del mismo en un futuro inmediato. Otros estudios, como en Fuertes & Kalotychou (2007), apoyan estos resultados encontrando evidencia de efectos *momentum* en la evolución de los *ratings* a lo largo del tiempo.

mayor detalle consúltese el Capítulo 3 (Baena-Mirabete & Puig (2018)).

1.1.1

Marco teórico

Considérese la variable aleatoria (v.a) discreta X_t tomando valores en un conjunto finito de estados $1, ..., m$. En adelante y sin pérdida de generalidad, el subíndice t se referirá a un momento del tiempo (p. ej. años), por tanto X_t denota el estado en el periodo actual t y X_{t-k} es el estado k periodos antes, siendo $k \leq t$. Una secuencia de v.a's sigue una cadena de Markov de orden k, denotada como MC$_k$, si se cumple,

$$P(X_t = i_t | X_0 = i_0, X_1 = i_1, ..., X_{t-1} = i_{t-1}) =$$
$$P(X_t = i_t | X_{t-k} = i_{t-k}, ..., X_{t-1} = i_{t-1})$$

donde $i_0, i_1, ..., i_t \in \{1, 2, ..., m\}$. Una cadena de Markov está completamente caracterizada por sus probabilidades de transición entre estados. Sea $n_{s;i_k,...,i_1}$ el número de transiciones observadas de la forma $X_{t-k} = i_k$, ..., $X_{t-1} = i_1$ y $X_t = s$ y $p^s_{i_k,...,i_1}$ su probabilidad de transición. Es inmediato ver que el estimador máximo verosímil (MLE) de la probabilidad de transición $p^s_{i_k,...,i_1}$ viene dado por

$$\hat{p}^s_{i_k,...,i_1} = \frac{n_{s;i_k,...,i_1}}{\sum_{j=1}^m n_{j;i_k,...,i_1}}.$$

Por tanto, el estimador MLE de $p^s_{i_k,...,i_1}$ puede fácilmente ser calculado por tabulación cruzada y su error estándar estimado usando la expresión común para la desviación estándar de una proporción,

$$\widehat{SE}(\hat{p}^s_{i_k,...,i_1}) = \sqrt{\frac{\hat{p}^s_{i_k,...,i_1}(1 - \hat{p}^s_{i_k,...,i_1})}{\sum_{j=1}^m n_{j;i_k,...,i_1}}}.$$

Cuando el orden k aumenta, muchas de las celdas de la tabulación cruzada pueden estar vacías ($n_{j;i_k,...,i_1} = 0$) y por tanto, no es posible hacer inferencia. A raíz de esto, si se desea explorar una dependencia de mayor orden, debe reducirse el número de parámetros. Una alternativa para reducir el número de parámetros en un MC$_k$, para estados ordenados, es considerar un modelo de regresión (logística) multinomial acumulativa como sigue,

$$P(X_t \leq s | X_{t-k} = i_{t-k}, ..., X_{t-1} = i_{t-1}) = \frac{\exp\left[\psi(s; i_{t-k}, ..., i_{t-1})\right]}{1 + \exp\left[\psi(s; i_{t-k}, ..., i_{t-1})\right]} \tag{1.1}$$

donde el predictor lineal tiene la forma,

$$\psi(s) := \psi(s; X_{t-k}, ..., X_{t-1}) = \beta_{s,0} + \sum_{i=1}^{k} \sum_{j=1}^{m-1} \beta_{s,j}^{(i)} I_{\{X_{t-i}=j\}}, \tag{1.2}$$

donde $\beta_{s,0}$ y $\beta_{s,j}^{(i)}$ son, respectivamente, los términos constantes y las pendientes y $I_{\{X_{t-i}=j\}}$ son funciones indicatrices que toman valor 1 si $X_{t-i} = j$ y 0 en caso contrario. Por tanto, se tienen $m-1$ ecuaciones logísticas para $s = 1, 2, ..., m-1$ donde los coeficientes son específicos para cada una de ellas. El sistema lineal anterior puede ser escrito más convenientemente en forma matricial como

$$\Psi = \eta_0 + B^{(1)} \cdot Y_1 + ... + B^{(k)} \cdot Y_k \tag{1.3}$$

donde $\Psi = (\psi(1), ..., \psi(m-1))^T$ es el vector de predictores lineales, $\eta_0 = (\beta_{1,0}, ..., \beta_{m-1,0})^T$ es el vector que contiene los términos constantes y $Y_i = (I_{\{X_{t-i}=1\}}, ..., I_{\{X_{t-i}=m-1\}})^T$ para cada retardo $i = 1, ..., k$. Finalmente, $B^{(i)}$ es la matriz cuadrada de parámetros:

$$B^{(i)} = \begin{bmatrix} \beta_{1,1}^{(i)} & \beta_{1,2}^{(i)} & \beta_{1,3}^{(i)} & ... & \beta_{1,m-1}^{(i)} \\ \beta_{2,1}^{(i)} & \beta_{2,2}^{(i)} & \beta_{2,3}^{(i)} & ... & \beta_{2,m-1}^{(i)} \\ & & ... & & \\ \beta_{m-1,1}^{(i)} & \beta_{m-1,2}^{(i)} & \beta_{m-1,3}^{(i)} & ... & \beta_{m-1,m-1}^{(i)} \end{bmatrix}$$

Por tanto, la probabilidad de transición $p_{i_k,...,i_1}^{s}$ se calcula, para $s > 1$, como

$$p_{i_k,...,i_1}^{s} = \frac{\exp\left[\psi(s; i_{t-k}, ..., i_{t-1})\right]}{1 + \exp\left[\psi(s; i_{t-k}, ..., i_{t-1})\right]} - \frac{\exp\left[\psi(s-1; i_{t-k}, ..., i_{t-1})\right]}{1 + \exp\left[\psi(s-1; i_{t-k}, ..., i_{t-1})\right]}$$

Nótese que las probabilidades de transición para $s = 1$ se calculan directamente de la ecuación (1.1). El modelo previo, conocido como el Modelo de Probabilidades No Proporcionales[3] (NPOM), tiene $k(m-1)^2 + (m-1)$ parámetros, muchos menos que el correspondiente MC_k que tiene $m^k(m-1)$ parámetros. Sin embargo, nótese que para $k = 1$ ambos modelos tienen el mismo número de parámetros a estimar. Un modelo más parsimonioso puede considerarse imponiendo el supuesto de probabilidades proporcionales[4], no obstante, este supuesto puede resultar muy restrictivo en determinados contextos.

[3]En inglés, *Non-proportional Odds Model. Véase Bender & Grouven (1998).*

[4]Del inglés, *Proportional Odds Model* (POM), que establece que los coeficientes para cada variable independiente son los mismos para todas las categorías ordenadas, y por tanto que $\beta_{1,j}^{(i)} = \beta_{2,j}^{(i)} = , ..., \beta_{m-1,j}^{(i)} = \beta_{j}^{(i)}$ para todo $j = 1, 2, ..., m-1$.

1.1.1.1

Modelos multiestados para cadenas de larga persistencia

Definimos una cadena persistente como aquella en que la matriz de transición a 1 paso es estrictamente diagonal dominante, es decir, $P(X_t = s|X_{t-1} = s) > 0{,}50$. Dada una cadena persistente, definimos

$$E_r^k(t) := E_r^k(t; X_{t-r}, ..., X_{t-k}) = \#\{X_{t-j} = X_{t-r}; j = r+1, ..., k\}$$

$$U_r^k(t) := U_r^k(t; X_{t-r}, ..., X_{t-k}) = \#\{X_{t-j} < X_{t-r}; j = r+1, ..., k\}$$

donde k y r son enteros, $k > r$. $E_r^k(t)$ representa el número de realizaciones entre los instantes $t-r-1$ y $t-k$ que son iguales al estado de la cadena en $t-r$. $U_r^k(t)$ cuenta el número de realizaciones por debajo (en el orden considerado) del estado de la cadena en $t-r$ entre los instantes $t-r-1$ and $t-k$. Nótese que el número de realizaciones en el intervalo $t-r-1$ and $t-k$ más grande que el estado de la cadena en $t-r$, denotado por $D_r^k(t)$, viene dado por $D_r^k(t) = k-r-E_r^k(t)-U_r^k(t)$. A modo ilustrativo, consideremos las siguientes cadenas: $(X_{t-3}, X_{t-2}, X_{t-1}, X_t) = (6, 5, 5, 4)$ y $(X_{t-3}, X_{t-2}, X_{t-1}, X_t) = (5, 5, 5, 4)$. Para la primera secuencia, $E_1^3(t) = 1, U_1^3(t) = 0$ y $D_1^3(t) = 1$. Para la segunda, $E_1^3(t) = 2$, $U_1^3(t) = 0$ y $D_1^3(t) = 0$. Nótese que, en este último caso, estas variables, conjuntamente con la información sobre el estado en $t-1$, provienen una información completa sobre la secuencia exacta de los estados de dicha cadena.

Definimos una cadena de Markov de larga persistencia de orden k y grado r, denotado[5] como LPM$_k(r)$, por medio de la relación:

$$P(X_t = i_t|X_0 = i_0, X_1 = i_1, ..., X_{t-1} = i_{t-1}) =$$
$$P(X_t = i_t|E_r^k(t) = n_e, U_r^k(t) = n_u, D_r^k(t) = n_d, X_{t-r} = i_{t-r}, ..., X_{t-1} = i_{t-1})$$

donde n_e,n_u,$n_d \in \{0, 1, ..., k-r\}$ tales que $n_e + n_u + n_d = k-r$. En la parte condicional, se tienen los términos referidos a la larga persistencia $E_r^k(t)$, $U_r^k(t)$, $D_r^k(t)$ y aquellos referidos a la corta persistencia representados por $X_{t-r} = i_{t-r}$,...,$X_{t-1} = i_{t-1}$.

Con el objeto de estimar las probabilidades de transición, generalizamos el modelo descrito en (1.3) a un LPM$_k(r)$. Por tanto, asumimos el siguiente modelo de regresión

[5] De sus siglas en inglés, *Long-Persistence Model*.

(logística) multinomial acumulativa:

$$P(X_t \leq s|E_r^k(t) = n_e, U_r^k(t) = n_u, X_{t-r} = i_{t-r},..., X_{t-1} = i_{t-1}) \\ = \frac{\exp\left[\psi(s; n_e, n_u, i_{t-r}, ..., i_{t-1})\right]}{1 + \exp\left[\psi(s; n_e, n_u, i_{t-r}, ..., i_{t-1})\right]}$$

donde el predictor lineal se define, de forma similar, como en la ecuación (1.2),

$$\psi(s) := \psi(s; E_r^k(t), U_r^k(t), X_{t-r},..., X_{t-1}) \\ = \beta_{s,0} + \sum_{i=1}^{r}\sum_{j=1}^{m-1} \beta_{s,j}^{(i)} I_{\{X_{t-i}=j\}} + \gamma_s^e E_r^k(t) + \gamma_s^u U_r^k(t)$$

donde γ_s^e y γ_s^u son los coeficientes para las variables referidas a la larga persistencia. Nuevamente, $m-1$ ecuaciones logísticas han sido definidas para $s = 1, 2, ..., m-1$. Por simplicidad, expresamos el sistema lineal anterior en forma matricial como en (1.3). Sin embargo, a fin de reducir el número de parámetros, sustituimos la matriz $B^{(i)}$ por la matriz diagonal-persistente $B^{*(i)}$ definida como sigue:

$$B^{*(i)} = \begin{bmatrix} \beta_{1,1}^{(i)} & \beta_2^{(i)} & \beta_3^{(i)} & ... & \beta_{m-1}^{(i)} \\ \beta_1^{(i)} & \beta_{2,2}^{(i)} & \beta_3^{(i)} & ... & \beta_{m-1}^{(i)} \\ & & ... & & \\ \beta_1^{(i)} & \beta_2^{(i)} & \beta_3^{(i)} & ... & \beta_{m-1,m-1}^{(i)} \end{bmatrix}$$

y tenemos,

$$\Psi = \eta_0 + B^{*(1)} \cdot Y_1 + ... + B^{*(r)} \cdot Y_r + \Gamma \cdot L_r^k$$

donde L_r^k es el vector de variables de larga persitencia $L_r^k = [E_r^k(t), U_r^k(t)]^T$ y Γ es la matriz con sus respectivos coeficientes, es decir $\Gamma = [\gamma^e, \gamma^u]$ donde $\gamma^e = (\gamma_1^e, ..., \gamma_{m-1}^e)^T$ y $\gamma^u = (\gamma_1^u, ..., \gamma_{m-1}^u)^T$. Es directo comprobar que el número de parámetros para el modelo LPM$_k(r)$, considerando m estados, es $(m-1)(3+2r)$, inferior al obtenido para un MC$_k$, que es $m^k(m-1)$. Nótese que el número de parámetros para el modelo LPM$_k(r)$ no depende de k, el orden del término referido a la larga persistencia. Por ejemplo, el número de parámetros para un LPM$_3(1)$ es igual a 25 en contraste con un MC$_3$, que es 1080.

1.1.2

Aplicación: medición del riesgo de crédito

La evaluación del riesgo de diferentes exposiciones crediticias es un objetivo importante tanto para los supervisores que diseñan los requisitos de capital regulatorio como para las

entidades bancarias implicadas en la asignación del capital económico. Una medida de riesgo para exposiciones de crédito ampliamente utilizada es el *Valor-en-Riesgo* (VaR[6]). Dada una cartera crediticia, el VaR, con un nivel de confianza α (%), es el cuantil estimado de la distribución de pérdidas correspondiente al valor que será excedido una fracción $1 - \alpha$ (%) de ocasiones por un inversor que invierta en dicha cartera. Por ejemplo, si una cartera de acciones tiene una VaR a un día de 40 millones de euros, con un nivel de confianza del $90\,\%$, esto significa que hay un $10\,\%$ de probabilidad que exista una pérdida en dicha cartera superior a 40 millones de euros, por lo que, si inviertes 100 veces en esta cartera, en 10 ocasiones obtendrás una pérdida superior a esa cifra. Dado que el *rating* de un producto financiero refleja su riesgo crediticio, un cambio en el *rating* tendrá una implicación directa en el precio de dicho producto. Si en lugar de disponer de un único producto financiero, se tiene un conjunto de productos que conforman una cierta cartera de inversión, los cambios de *rating* producidos en algunos de esos productos en un cierto horizonte de tiempo fijado (p.ej. 1 año), tendrán implicaciones en el valor de la cartera. En este caso, una estimación precisa de la matriz de transición entre *ratings* es crucial para una evaluación precisa del riesgo de crédito a partir del VaR. Como aplicación de los modelos propuestos, se analiza una muestra de 4747 emisores de bonos municipales en USA calificados por la agencia Standard & Poor's en el periodo comprendido entre los años 1990 y 2013. Se estiman diferentes modelos de cadenas de Markov, descritos anteriormente, con el objeto de obtener estimaciones de las probabilidades de transición entre *ratings* en el horizonte fijado de 1 año. Finalmente, se usan dichas estimaciones para obtener el VaR de una hipotética cartera de bonos con exposiciones simuladas. En este caso, se obtiene que el VaR estimado por un modelo de Markov de orden 1, muy utilizado en la literatura, es inferior al obtenido por modelos de mayor orden capaces de tener en cuenta efectos *momentum* comentados anteriormente. Para un mayor detalle de los resultados obtenidos así como interpretaciones relevantes en el contexto del riesgo de crédito, consúltese el Capítulo 3 (Baena-Mirabete & Puig (2018)).

[6]De sus siglas en inglés, *Value-at-Risk*.

1.2

Modelos de mixturas basados en cadenas de Markov con dos estados

Los modelos basados en cadenas de Markov para explicar o predecir la dinámica de un proceso con 2 posibles resultados son ampliamente utilizados en la literatura. Sea $\{X_t = i_t : t = 1, ..., T\}$ una secuencia de variables aleatorias con 2 posibles estados, es decir, $i_1, ..., i_T \in \{0, 1\}$. Una cadena de Markov de orden k (MC$_k$) puede ser estimada a partir de un modelo de regresión logística donde el predictor lineal viene dado por

$$\psi := \psi(X_{t-k}, ..., X_{t-1}; \boldsymbol{\beta}) = \beta_0 + \sum_{s=1}^{k} \beta_s X_{t-s} + \vartheta(k, \tilde{\boldsymbol{\beta}}),$$

donde $\boldsymbol{\beta}$ es el vector de parámetros y $\vartheta(k, \tilde{\boldsymbol{\beta}})$ es una función que representa las interacciones hasta orden k. Nótese que un MC$_k$ con 2 estados tiene 2^k parámetros a estimar.

Los modelos que aquí proponemos derivan de un estudio observacional realizado a un grupo de estudiantes a quienes les propusimos simular mentalmente las realizaciones de una moneda justa. Inicialmente, se estimaron varios modelos de cadenas de Markov para diferentes órdenes. Los resultados para estos modelos muestran que los individuos no recuerdan de forma exacta sus resultados previos pero sí, en cambio, un resumen de los mismos. Esto es debido a que la memoria corta tiene capacidad limitada (Miller, 1956). Esto nos ha llevado a plantear modelos de cadenas de Markov más parsimoniosos en los cuales no son necesarios detallar, de forma exacta, todas las combinaciones de estados posibles hasta un cierto orden k. Concretamente, se proponen modelos donde el predictor lineal tiene la forma

$$\psi = \beta_0 + \sum_{s=1}^{r} \beta_s X_{t-s} + \vartheta(r, \tilde{\boldsymbol{\beta}}) + \gamma \sum_{s=r+1}^{k} X_{t-s},$$

con $r < k$. En adelante, este modelo se denotará como SLMC$_k(r)$[7], con orden k y grado r. La memoria corta viene representada por los resultados de las r inmediatamente previas observaciones $X_{t-r} = i_{t-r}, ..., X_{t-1} = i_{t-1}$. La memoria larga viene representada por la influencia de los estados entre los instantes $t-r-1$ y $t-k$, resumidos por $\sum_{s=r+1}^{k} X_{t-s}$. En este caso, nótese que el número de parámetros a estimar es $2^r + 1$.

[7] De sus siglas en inglés, *k-order Short-Long-time memory Markov Chain*.

Sin embargo, el proceso mental de creación de secuencias binarias *aleatorias* por parte de individuos es más complicado y modelos basados en mixturas han obtenido mejores resultados en términos de ajuste. A continuación, se proponen diferentes modelos de mixturas basados en cadenas de Markov de alto orden. Estos modelos se describen con mayor detalle en el Capítulo 4 (Baena-Mirabete et al., 2019). Los datos y códigos usados en este apartado están disponibles en el siguiente link: *http://www.statmod.org/smij/archive.html*.

1.2.1 Modelo latente de Markov

Existe una extensa literatura sobre el análisis de datos longitudinales (p.ej. Fitzmaurice et al. (2008)). En esta literatura, los Modelos Latentes de Markov (Wiggins, 1973), en adelante referidos como LM, ocupan un papel relevante debido su aplicación en diferentes ámbitos. Sea $\{X_{jt} = i_t : t = 1, ..., T\}$, la secuencia observada de variables binarias en el tiempo para el j-ésimo individuo con $j = 1, ..., N$ y $i_1, ..., i_T \in \{0, 1\}$. El modelo LM considera que dado un valor concreto de una cierta variable latente (no observable) C_{jt} tomando un número finito de valores discretos (estados latentes), las variables observadas en el tiempo son independientes. Las probabilidades condicionadas $h_{1|c} := P(X_{jt} = 1|C_{jt} = c)$, con $c = 1, ..., z$ donde z toma finitos valores discretos, proponemos modelizarlas de acuerdo a una función logística como sigue,

$$h_{1|c} = \frac{\exp\{\alpha_c\}}{1 + \exp\{\alpha_c\}}.$$

En la expresión anterior, los parámetros α_c son los términos constantes correspondientes a cada uno de los estados latentes. En un modelo LM, se asume que $C_{j1}, ..., C_{jT}$ sigue una cadena de Markov de orden 1 con espacio de estados $\{1, ..., z\}$. Nosotros proponemos modelizar las probabilidades de transición de acuerdo a una parametrización logística multinomial,

$$\log \frac{P(C_{jt} = c|C_{j,t-1} = \ddot{c}, X_{j,t-k} = i_{t-k}, ..., X_{j,t-1} = i_{t-1})}{P(C_{jt} = \ddot{c}|C_{j,t-1} = \ddot{c}, X_{j,t-k} = i_{t-k}, ..., X_{j,t-1} = i_{t-1})} = \psi(i_{t-k}, ..., i_{t-1}; \boldsymbol{\beta}_{\ddot{c}\boldsymbol{c}}), \tag{1.4}$$

para $t = k + 2, ..., T$ y $\ddot{c}, c = 1, ..., z$ con $\ddot{c} \neq c$ y donde $\boldsymbol{\beta}_{\ddot{c}\boldsymbol{c}}$ es el vector de parámetros. Asumimos también que los procesos subyacentes pertenecen a la clase de predictores

lineales SLMC$_k(r)$, esto es,

$$\psi(X_{j,t-k},...,X_{j,t-1};\boldsymbol{\beta_{\ddot{c}c}}) = \beta_{0\ddot{c}c} + \sum_{s=1}^{r} \beta_{s\ddot{c}c} X_{j,t-s} + \vartheta_j(r,\tilde{\boldsymbol{\beta}}_{\boldsymbol{\ddot{c}c}}) + \gamma_{\ddot{c}c} \sum_{s=r+1}^{k} X_{j,t-s},$$

donde, de nuevo, $\vartheta_j(r,\tilde{\boldsymbol{\beta}}_{\boldsymbol{\ddot{c}c}})$ es la función que representa las interacciones hasta orden r y $\tilde{\boldsymbol{\beta}}_{\boldsymbol{\ddot{c}c}}$ el vector de parámetros correspondiente a estas interacciones. Nótese que $z(z-1)$ ecuaciones lineales se definen en (1.4) y, por consiguiente, hay $z(z-1)(2^r+1)$ parámetros contenidos en el conjunto de vectores $\boldsymbol{\beta_{\ddot{c}c}}$. Hemos propuesto modelizar las probabilidades iniciales del proceso latente, en el instante $k+1$, en función de las variables observadas retardadas adoptando, nuevamente, una parametrización multinomial logística como sigue,

$$\log \frac{P(C_{j,k+1}=c|X_{j,t-k}=i_{t-k},...,X_{j,t-1}=i_{t-1})}{P(C_{j,k+1}=1|X_{j,t-k}=i_{t-k},...,X_{j,t-1}=i_{t-1})} = \bar{\psi}(i_{t-k},...,i_{t-1};\boldsymbol{\beta_c}),$$

para c=2,...,z y de nuevo, asumimos que los predictores lineales $\bar{\psi}$ pertenecen a la clase SLMC$_k(r)$. Concretamente, proponemos una parsimoniosa especificación SLMC$_k(0)$, esto es, $\bar{\psi}(X_{j,t-k},...,X_{j,t-1};\boldsymbol{\beta_c}) = \beta_{0c} + \gamma_c \sum_{s=1}^{k} X_{j,t-s}$. Es directo ver que el número de parámetros a estimar es $3z+z(z-1)(2^r+1)-2$. El algoritmo EM[8] (Dempster et al., 1977) es muy utilizado para obtener estimaciones por máxima verosimilitud de los parámetros del modelo LM, especialmente cuando las variables del proceso latente tienen un soporte discreto. Un resumen detallado del algoritmo EM enfocado a la estimación de Modelos Latentes de Markov puede consultarse en Bartolucci et al. (2014b).

1.2.2

Una extensión del modelo MTD

El modelo MTD, como se ha mencionado anteriormente, fue propuesto por Raftery (1985) para aproximar cadenas de Markov de alto orden utilizando un menor número de parámetros respecto al correspondiente modelo de Markov completo[9]. Sean $\{C_{jt} : t = 1,...,T\}$ variables latentes independientes e idénticamente distribuidas (i.i.d.) con espacio de estados $\{1,...,k\}$. El modelo MTD de orden k asume que

$$P(X_{jt}=1|C_{jt}=c, X_{j,t-k}=i_{t-k},...,X_{j,t-1}=i_{t-1}) = P(X_{jt}=1|C_{jt}=c, X_{j,t-c}=i_{t-c}),$$

[8] Siglas del inglés, Expectation–Maximization (EM) algorithm.

[9] El término *completo* es usado aquí para referirnos a un modelo de Markov de un cierto orden k (MC$_k$) y 2 estados posibles, cuya matriz de transición está totalmente parametrizada (2^k parámetros).

para $t = k+1, ..., T$ y $c = 1, ..., k$. Por tanto, la variable respuesta X_{jt} depende únicamente de la variable retardada $X_{j,t-c}$, donde el retardo c es elegido por un mecanismo aleatorio no observable de forma directa.

Bartolucci & Farcomeni (2010) propusieron una generalización del modelo MTD basada en el supuesto que las variables latentes no eran i.i.d. sino que seguían un proceso de Markov de orden 1. Nosotros hemos adaptado este modelo al contexto de datos longitudinales y respuesta binaria. Concretamente, modelizamos

$$h_{j,1|c} := P(X_{jt} = 1|C_{jt} = c, X_{j,t-c} = i_{t-c}),$$

con $c = 1, ..., k$ y $h_{j,1|c}$ basada en una función logística como sigue,

$$h_{j,1|c} = \frac{\exp\{\psi(i_{t-c}; \boldsymbol{\beta_c})\}}{1 + \exp\{\psi(i_{t-c}; \boldsymbol{\beta_c})\}}. \tag{1.5}$$

El predictor lineal en la ecuación anterior viene dado por:

$$\psi(X_{j,t-c}; \boldsymbol{\beta_c}) = \beta_{0c} + \beta_{1c} X_{j,t-c},$$

donde $\boldsymbol{\beta_c} = (\beta_{0c}, \beta_{1c})$ es el vector de parámetros del predictor lineal para cada estado latente. Aquí, el proceso latente $\{C_{jt} : t = 1, ..., T\}$ sigue una cadena de Markov de orden 1, con probabilidades iniciales $w_c = P(C_{j,k+1} = c)$, $c = 1, ..., k$, y probabilidades de transición $\phi_{c_1,c_2} = P(C_{jt} = c_2|C_{j,t-1} = c_1)$, $c_1, c_2 = 1, ..., k$ para $t > k + 1$. Nótese que el número de parámetros en este modelo propuesto es $k^2 + 2k - 1$. Una reducción del número de parámetros puede obtenerse imponiendo la restricción $\boldsymbol{\beta_c} = \boldsymbol{\beta}$ en la ecuación (1.5) y, en este caso, el número de parámetros a estimar es $k^2 + 1$. Se han implementado varias rutinas en R (R Core Team, 2017) que estiman, por máxima verosimilitud, los modelos MTD descritos en este apartado usando el algoritmo EM (Dempster et al. (1977);Bartolucci & Farcomeni (2010)).

1.2.3 Mixtura de cadenas de Markov de alto orden

Suponemos, nuevamente, que tenemos una muestra de N individuos y para el j-ésimo individuo, observamos realizaciones de una secuencia de variables aleatorias binarias en el tiempo $\boldsymbol{x_j} = \{X_{jt} = i_t : t = 1, ..., T\}$, $j = 1, ..., N$ y $i_1, ..., i_T \in \{0, 1\}$. Asumimos

que, condicionado a la variable latente C_{jt} con soporte discreto y finito (estados latentes), las variables aleatorias $\{X_{jt} : t = k+1, ..., T\}$ siguen una cadena de Markov de orden k. Concretamente, se propone modelizar

$$h_{j,1|c} := P(X_{jt} = 1|C_{jt} = c, X_{j,t-k} = i_{t-k}, ..., X_{j,t-1} = i_{t-1}),$$

con $c = 1, 2$ y $h_{j,1|c}$ basada en una función logística como sigue,

$$h_{j,1|c} = \frac{\exp\{\psi(i_{t-k}, ..., i_{t-1}; \boldsymbol{\beta_c})\}}{1 + \exp\{\psi(i_{t-k}, ..., i_{t-1}; \boldsymbol{\beta_c})\}}. \tag{1.6}$$

Se asumen también que los procesos subyacentes en la ecuación anterior pertenecen a la clase de predictores lineales SLMC$_k(r)$,

$$\psi(X_{j,t-k}, ..., X_{j,t-1}; \boldsymbol{\beta_c}) = \beta_{0c} + \sum_{s=1}^{r} \beta_{sc} X_{j,t-s} + \vartheta_j(r, \tilde{\boldsymbol{\beta}}_{\boldsymbol{c}}) + \gamma_c \sum_{s=r+1}^{k} X_{j,t-s},\ c = 1, 2$$

donde $\boldsymbol{\beta_c}$ es el vector de parámetros del predictor lineal para cada estado latente. De nuevo, $\vartheta_j(r, \tilde{\boldsymbol{\beta}}_{\boldsymbol{c}})$ es la función que representa las interacciones hasta orden r y $\tilde{\boldsymbol{\beta}}_{\boldsymbol{c}}$ el vector de parámetros correspondiente a estas interacciones.

Sea u_j una variable aleatoria continua asociada con el j-ésimo individuo y sean $\{u_j\}_{j=1}^{N}$ independientes e idénticamente distribuidas. Asumimos que, condicionadas a una realización de la variable aleatoria u_j, las variables latentes $C_{j1}, ..., C_{jT}$ son independientes. Se propone modelizar las probabilidades condicionadas $w_{j,c|u} := P(C_{jt} = c|u_j)$, $c = 1, 2$, como

$$w_{j,1|u} = \Phi^{-1}(\alpha + u_j), \tag{1.7}$$

donde Φ es la distribución normal estándar acumulada, α es un parámetro común para todos los individuos y $w_{j,2|u} = 1 - w_{j,1|u}$.

Nótese que, condicionada al efecto aleatorio u_j, la secuencia de variables aleatorias $\{X_{jt} : t = k+1, ..., T\}$ sigue una mixtura de Bernoullis de dos componentes:

$$X_{jt}|u_j \sim \begin{cases} Bern\left(h_{j,1|1}\right), & \text{con probabilidad } w_{j,1|u} \\ Bern\left(h_{j,1|2}\right), & \text{con probabilidad } w_{j,2|u} \end{cases} \tag{1.8}$$

Aquí, $w_{j,c|u}$, $c = 1, 2$, son las probabilidades de la mixtura (pesos) tal cual definidas en (1.7) y que asumimos independientes entre los individuos. Para cada componente de la mixtura, las probabilidades $h_{j,1|c}$, $c = 1, 2$, no son constantes, sino que cambian de

acuerdo a las probabilidades de transición de dos cadenas de Markov como definidas en (1.6).

Para el j-ésimo individuo, las probabilidades de transición condicionadas al efecto aleatorio u_j se denota como $p_{i_{t-k},\ldots,i_{t-1}|u_j}$. Estas pueden obtenerse tomando la esperanza condicional de (1.8), y se tiene,

$$P(X_{jt} = 1|X_{j,t-k} = i_{t-k}, \cdots, X_{j,t-1} = i_{t-1}; u_j; \boldsymbol{\xi}) = \sum_{c=1}^{2} w_{j,c|u} h_{j,1|c}, \qquad (1.9)$$

donde $\boldsymbol{\xi}$ es el vector que contiene todos los parámetros del modelo. Las probabilidades de transición no condicionadas $\hat{p}^{(j)}_{i_{t-k},\ldots,i_{t-1}}$ estimadas para el j-ésimo individuo, se obtienen integrando (1.9) sobre la distribución a posteriori del efecto aleatorio u_j. Para un mayor detalle véase Capítulo 4 (Baena-Mirabete et al., 2019).

La función de verosimilitud, para el j-ésimo individuo, $t \geq k+1$, viene dada por,

$$L_j\left(\boldsymbol{\xi}; \boldsymbol{x_j}\right) = \int_{u_j} \left(\prod_{t=k+1}^{T} f_j\left(i_t|i_{t-k}, ..., i_{t-1}; u_j, \alpha, \boldsymbol{\beta_1}, \boldsymbol{\beta_2}\right) \right) g\left(u_j|\sigma\right) du_j, \qquad (1.10)$$

donde

$$f_j\left(i_t|i_{t-k}, ..., i_{t-1}; u_j, \alpha, \boldsymbol{\beta_1}, \boldsymbol{\beta_2}\right) = \sum_{c=1}^{2} w_{j,c|u} \left(h_{j,1|c}\right)^{i_t} \left(1 - h_{j,1|c}\right)^{1-i_t},$$

y $g\left(u_j|\sigma\right)$ es la función de densidad del efecto aleatorio, que asumimos que sigue una distribución Normal con desviación típica σ.

Dado que los individuos se asumen independientes entre sí, la función de verosimilitud global viene dada por la expresión $L(\boldsymbol{\xi}; \boldsymbol{x_1}, ..., \boldsymbol{x_N}) = \prod_{j=1}^{N} L_j(\boldsymbol{\xi}; \boldsymbol{x_j})$. Estimadores máximo-verosímiles de los parámetros son obtenidos maximizando numéricamente dicha función y, para ello, es necesario evaluar la integral de la expresión (1.10), por ejemplo, usando el método clásico de cuadratura gaussiana adaptativa, como descrito en Pinheiro & Bates (1995).

El modelo definido en (1.8) puede extenderse para considerar un mayor número de estados latentes ($c > 2$). Básicamente, las ecuaciones mostradas previamente serían idénticas y sólo los pesos $w_{j,c|u}$ deberían reformularse. Para un mayor detalle véase Capítulo 4 (Baena-Mirabete et al., 2019). El modelo propuesto en este apartado ha sido implementado usando el procedimiento NLMIXED en SAS v9.3 (SAS Institute, 2011).

1.2.4

Aplicación: estudio de la aleatoriedad de secuencias binarias generadas mentalmente por individuos

Hace casi 70 años, Hans Reichenbach[10], uno de los más importantes filósofos de la ciencia del siglo XX, sostuvo que las personas somos incapaces de generar secuencias aleatorias (Reichenbach, 1949). Investigaciones posteriores han apoyado esta afirmación y numerosos experimentos se han llevado a cabo con el fin de estudiar cómo la mente humana produce secuencias aleatorias y cómo esta juzga si una particular secuencia es o no aleatoria. Véase Bar-Hillel & Wagenaar (1991) y Nickerson (2002) para un resumen de diferentes estudios y experimentos al respecto. En general, en los diferentes experimentos reportados en la literatura (Budescu (1985);Lopes & Oden (1987);Falk & Konold (1997)), las personas tendemos a evitar patrones repetitivos en nuestras respuestas *aleatorias*. Sin embargo, algunos experimentos han señalado también, en algunos individuos, cierta tendencia a la repetición (e.g. Neuringer (1986);Budescu (1987)). En este tipo de experimentos, la tendencia a repetir la respuesta inmediatamente anterior es conocida en la literatura como 'efecto positivo de la inmediatez' y la tendencia a evitar patrones repetitivos se conoce como 'efecto negativo de la inmediatez'[11].

Hemos llevado a cabo un experimento en el que participaron 262 estudiantes del grado de Biología de la Universitat Autònoma de Barcelona. Después de una sesión informativa, a cada uno de los estudiantes se les retaba a simular mentalmente una moneda justa generando una serie de resultados binarios (cara-cruz) lo más aleatorios posibles. El experimento se llevó a cabó en diferentes días y en grupos de alrededor de 30 estudiantes con una edad comprendida entre los 18 y los 20 años. Cada estudiante debía generar una serie de 50 resultados de cara-cruz usando un aplicativo implementado para tal fin en Visual Basic para Aplicaciones (VBA). Para ello, los estudiantes debían pulsar, repetidamente,

[10]Hans Reichenbach (26 de septiembre de 1891, Hamburgo, – 9 de abril de 1953, Los Ángeles) fue un físico, filósofo y lógico alemán. Hizo importantes contribuciones a la teoría de la probabilidad y a las interpretaciones filosóficas de la relatividad, de la mecánica cuántica y de la termodinámica. Fundó el Círculo de Berlín, cuyos miembros participaron en muchas de las discusiones del Círculo de Viena, por lo que a veces se les considera como representantes del positivismo lógico.

[11]En inglés, *positive and negative recency effects*.

el ratón del ordenador y elegir entre las siguientes dos opciones mostradas en pantalla: 'CARA' o 'CRUZ'. Los resultados generados se guardaban en un 'worksheet' oculto de tal forma que los individuos no podían ver sus resultados previos generados.

A continuación, se resumen las principales conclusiones obtenidas en la estimación de los diferentes modelos expuestos. Para un mayor detalle de las mismas consúltese el Capítulo 4 (Baena-Mirabete et al., 2019). Se han estimado, en primer lugar, modelos basados en cadenas de Markov (MC_k) con dos estados posibles para diferentes valores de k. Estos modelos, típicamente usados en este campo, asumen que los resultados de 'cara-cruz' producidos por los individuos dependen directamente de sus últimos resultados. Sin embargo, uno podría plantearse que la decisión de 'cara-cruz' no puede ser explicada únicamente en base a sus últimos resultados sino que podría existir algún otro factor no observable entre individuos afectando también la variable respuesta, tal y como asumen los modelos descritos en la sección 1.2. En este sentido, los Modelos Latentes de Markov descritos en la sección 1.2.1 asumen que las respuestas 'cara-cruz' dependen indirectamente de los resultados previos solo a través de un proceso latente (no observable), lo que introduce una heterogeneidad adicional no observable entre individuos. En contraste, los modelos MTD (sección 1.2.2) asumen que los individuos generan una cierta respuesta 'cara-cruz' dependiendo solo de una de las respuestas previamente producidas, que es elegida de acuerdo a un mecanismo no observable. Los pobres resultados obtenidos para los modelos MTD, en términos del índice de ajuste BIC[12], sugieren que los individuos, en su proceso mental de aleatorización, tienen en cuenta todas sus respuestas previas, no solo una de ellas, al producir un cierto resultado 'cara-cruz'. Entre los modelos ajustados, una mixtura de dos cadenas de Markov, como propuesta en la sección 1.2.3, obtuvo los mejores resultados. Desde este último enfoque, mostramos que los individuos se desvían de la aleatoriedad cuando intentan generar una serie aleatoria debido a la combinación de dos procesos: el primero, produce un exceso de respuestas alternadas y una tendencia a equilibrar la distribución de resultados 'cara-cruz' en fragmentos cortos. El segundo pro-

[12]Para comparar los distintos modelos estimados, usamos el índice de bondad de ajuste BIC, siglas del inglés *Bayesian Information Criterion*. Este índice se define como $BIC = -2\hat{l} + q\log(n)$, donde $\hat{l}$ es el logaritmo de la función de verosimilitud, q es el número de parámetros y n es el número de componentes en la función de verosimilitud. El mejor modelo en términos de bondad de ajuste es aquel que presenta un valor más bajo en este índice.

ceso produce demasiadas repeticiones. Además, el papel que juega cada uno de estos dos procesos en el experimento depende del individuo.

Finalmente, los estudiantes que participaron en el experimento son, al menos vagamente, capaces de recordar sus últimos previos seis resultados producidos. Esta conclusión es consistente con aquella discutida en Miller (1956) quien sostuvo que el número de elementos que un individuo medio puede recordar es 7 ± 2. Estudios posteriores han corroborado este límite superior a nuestra capacidad de almacenar información (p.ej. Saaty & Ozdemir (2003)).

1.3

Relaciones de recurrencia para el cálculo de la función de probabilidad de distribuciones con soporte en $\mathbb{Z}$

Recientemente, muchos métodos estadísticos se han aplicado al análisis de resultados de fútbol, véase por ejemplo, Karlis & Ntzoufras (2003) y Gómez-Déniz et al. (2019). Considérese la variable aleatoria $Z = G_1 - G_2$ definida como la diferencia de goles marcados por dos equipos en un cierto partido. El número de goles marcados por cada equipo puede expresarse como $G_h = X_h + U$, para $h = 1, 2$, donde la variable aleatoria de recuento U, común para ambos equipos, refleja las condiciones ambientales o factores inherentes al propio partido, como por ejemplo, la permisividad del arbitraje. Por tanto, G_1 y G_2 están correlacionadas, pero X_1 y X_2 pueden asumirse independientes, y la diferencia de goles $Z = X_1 - X_2$ es simplemente la diferencia de dos variables aleatorias de recuento independientes. Karlis & Ntzoufras (2003) estudiaron el caso particular donde las variables X_h siguen una distribución de Poisson. El cálculo eficiente de las probabilidades de Z es crucial para evaluar la función de verosimilitud y ajustar modelos con covariables capaces de explicar la diferencia de goles. Pero ¿cómo calcular eficientemente las probabilidades de la diferencia (o la suma) de variables aleatorias independientes? Este será el principal objetivo de esta sección. Para un mayor detalle véase Capítulo 5 (Baena-Mirabete & Puig, 2020).

Sea X una variable aleatoria (va) de recuento que toma valores $n = 0, 1, ...$ y se denota la probabilidad $P(X = n)$ como p_n. Para algunas familias de distribuciones de

probabilidad, existe una simple relación de recurrencia entre p_n y p_{n-1}, como es el caso de la conocida familia de Katz-Panjer de distribuciones (Katz, 1965; Panjer, 1981). Sundt (1992) extendió esta familia considerando distribuciones de recuento cuya función de masa de probabilidad (fmp) satisfacía la siguiente relación de recurrencia,

$$p_n = \sum_{j=1}^{r}(\alpha_j + \beta_j/n)p_{n-j}, \qquad n = 1, 2, ... \tag{1.11}$$

para algún entero positivo r y constantes α_j, $\beta_j \in \mathbb{R}$, $j = 1, ..., r$, con $p_n = 0$ para $n < 0$. Esta clase de distribuciones contiene la familia de Katz-Panjer como un caso especial, tomando $r = 1$. Teorema 1 de Sundt (1992) muestra que la función generatriz de probabilidad (fgp), denotada como $f(z)$ con $z \in \mathbb{C}$, para una distribución de recuento surgiendo de (1.11), satisface

$$\frac{\mathrm{d}}{\mathrm{d}z}\ln f(z) = \frac{\sum_{j=1}^{r}(j\alpha_j + \beta_j)z^{j-1}}{1 - \sum_{j=1}^{r}\alpha_j z^j}. \tag{1.12}$$

Nótese que el término de la derecha en la ecuación (1.12) es el cociente entre un polinomio de grado hasta $r - 1$ y un polinomio de grado hasta r con término constante no nulo.

En el Capítulo 5 (Baena-Mirabete & Puig, 2020), extendemos algunos de los resultados descritos en Sundt (1992), generalizando la condición (1.12) a polinomios de cualquier grado en z. Por tanto, proveemos relaciones de recurrencia para el cálculo de la fmp de una gran familia de distribuciones de recuento, incluyendo aquellas con soporte entero en $\mathbb{Z}$.

1.3.1 Marco teórico

Sea X una va con soporte en los enteros y sea $p_n := P(X = n)$ su fmp definida en $n \in \mathbb{Z}$ (incluyendo enteros positivos y negativos). La fgp de X, definida como $f(\omega) = \mathbb{E}(\omega^X)$, $\omega \in (0, 1)$, puede ser analíticamente extendida al dominio complejo $0 < |z| < 1$. Consideramos la clase de va's con fgp, $f(z)$, que satisface

$$\frac{\mathrm{d}}{\mathrm{d}z}\ln f(z) = \frac{G(z)}{T(z)}, \tag{1.13}$$

donde $G(z) = \sum_{j=0}^{r}\psi_j z^j$ y $T(z) = \sum_{j=0}^{s}\eta_j z^j$ son polinomios de grado r y s respectivamente, con $z \in \mathbb{C}$. La forma funcional (1.13) es satisfecha por una amplia familia

de distribuciones de recuento, tales como la familia de distribuciones de Katz-Panjer, la Skellam o la distribución de Hermite. Nótese que la clase de distribuciones que satisfacen (1.12) está incluida en aquellas que satisfacen (1.13).

La fgp de una va definida en $\mathbb{Z}$ puede ser expresada como una serie de Laurent alrededor del cero,

$$f(z) = \mathbb{E}(z^X) = \sum_{n=-\infty}^{\infty} p_n z^n. \tag{1.14}$$

con $0 < |z| < 1$. Para va's definidas en $\mathbb{N}$, la derivada n-ésima de la fgp en $z = 0$, dividida por $n!$, da la probabilidad p_n. Sin embargo, las derivadas en $z = 0$ no tienen sentido para fgp's de va's definidas en $\mathbb{Z}$. De forma alternativa, las probabilidades pueden calcularse usando la fórmula de la integral de Cauchy,

$$p_n = \frac{1}{2\pi i} \oint_{\gamma} \frac{f(z)}{z^{n+1}} \, \mathrm{d}z, \tag{1.15}$$

donde $i = \sqrt{-1}$ and γ es un entorno cerrado cuyo interior contiene $z = 0$. Cuando $f(z)$ es analítica en una región que contiene el disco unidad, considerando el entorno γ como el círculo unidad, $|z| = 1$, parametrizado por $z(\theta) = \exp(i\theta)$, $0 \leq \theta \leq 2\pi$, $\mathrm{d}z = i \exp(i\theta) \mathrm{d}\theta$, el contorno integral (1.15) puede ser evaluado como,

$$p_n = \frac{1}{2\pi} \int_0^{2\pi} \exp(-i\theta n) f[\exp(i\theta)] \, \mathrm{d}\theta. \tag{1.16}$$

La expresión (1.16) es especialmente útil para la evaluación numérica de las probabilidades de muchas distribuciones con soporte en los enteros (positivos y negativos) con un bajo coste computacional. El siguiente resultado muestra que las probabilidades de va's definidas en los enteros (p.ej. va's de recuento), con una fgp cumpliendo (1.13), pueden ser obtenidas recursivamente de forma eficiente.

Proposición 1. La fmp de una va con fgp cumpliendo la condición (1.13) satisface la siguiente relación de recurrencia,

$$\sum_{j=0}^{s} (n-j)\eta_j p_{n-j} = \sum_{j=0}^{r} \psi_j p_{n-j-1}. \tag{1.17}$$

Para la demostración de la proposición anterior véase el Capítulo 5 (Baena-Mirabete & Puig, 2020).

1.3.2

Ejemplos de aplicación

1.3.2.1

Suma de variables aleatorias independientes de recuento

En muchos estudios estadísticos, uno puede estar interesado en estudiar la suma de va's de recuento. Por ejemplo, el total de siniestros en una cartera de seguros se modeliza como la suma de los siniestros en cada una de las diferentes pólizas que conforman dicha cartera, que se asumen independientes. Sean X_h, $h = 1, ..., k$, va's independientes, y se define $X = X_1 + ... + X_k$. Entonces, la fgp de la va X viene dada por $f_X(z) = \prod_{h=1}^{k} f_{X_h}(z)$. Si cada $f_{X_h}(z)$ satisface (1.13), $f_X(z)$ también cumple (1.13) y entonces, por la Proposición 1, la fmp de X puede ser recursivamente obtenida por (1.17). En el Capítulo 5 (Baena-Mirabete & Puig, 2020), se muestran algunos ejemplos concretos de recurrencias obtenidas para la suma de va's de recuentos independientes. Entre estas, de especial interés son aquellas relaciones de recurrencia relativas a los modelos INAR (Al-Osh & Alzaid, 1987) y sus extensiones, que son modelos muy utilizados en el análisis de series temporales discretas. Un INAR(1) se define por la siguiente ecuación:

$$X_t = \alpha \circ X_{t-1} + W_t,$$

donde el parámetro $0 < \alpha < 1$ es fijo y la notación $\circ$ hace referencia al conocido operador *binomial thinning* tal que $[\alpha \circ X_{t-1} | X_{t-1} = x_{t-1}] =_d \sum_{i=1}^{x_{t-1}} \{\mathrm{B}_i(\alpha)\}$, donde $\{\mathrm{B}_i(\alpha)\}$ es una secuencia de va's de Bernoulli(α) independientes e idénticamente distribuidas. Las innovaciones del modelo, $\{W_t\}$, siguen una distribución de Poisson(λ) y también son independientes e idénticamente distribuidas. La esperanza y la varianza de $\{X_t\}$ son $\mathbb{E}(X_t) = \mathbb{V}(X_t) = \lambda/(1-\alpha)$. Nótese, por tanto, que el clásico modelo INAR(1) implica la convolución de dos distribuciones, una Binomial y otra de Poisson. La exacta, pero computacionalmente ineficiente, fmp de la suma de va's de Poisson y Binomial, ambas independientes, puede consultarse, por ejemplo, en Freeland & McCabe (2004). Las probabilidades necesarias para el cálculo de la función de verosimilitud en un INAR(1) pueden obtenerse de forma más eficiente a partir de la relación de recurrencia mostrada en Baena-Mirabete & Puig (2020). El cálculo de la función de verosimilitud para modelos

INAR(p) de mayor orden ($p > 1$) son calificados como numéricamente intratables en la literatura y diferentes métodos han sido propuestos (p.ej. Pedeli et al. (2015) y Lu (2018)). Nuestra metodología es también apropiada para este tipo de problemas. Por ejemplo, la fmp condicionada de un modelo INAR(2) es la convolución de dos distribuciones Binomial con parámetros $v_h > 0$, $q_h \in (0, 1)$, $h = 1, 2$, y la distribución de Poisson de las innovaciones con parámetro $\lambda > 0$. Cálculos directos muestran que la fgp en este caso cumple la ecuación (1.13) y entonces, por la Proposición 1, la fmp satisface la siguiente relación de recurrencia,

$$p_{n+1} = \frac{1}{(n+1)(1-q_1-q_2+q_1q_2)} \left[\lambda q_1 q_2 p_{n-2} + A(n)p_{n-1} + B(n)p_n\right], \quad n = 2, 3, ...$$

donde $A(n) = q_1q_2(v_1 + v_2) + (q_1 + q_2 - 2q_1q_2)\lambda - (n-1)q_1q_2$ y $B(n) = v_1q_1 + v_2q_2 - q_1q_2(v_1 + v_2) + (1 - q_1 - q_2 + q_1q_2)\lambda - n(q_1 + q_2 - 2q_1q_2)$.

Un estudio del comportamiento numérico de algunas de la recurrencias obtenidas para la suma de va's puede consultarse en el Capítulo 5 (Baena-Mirabete & Puig, 2020) en el apartado de material suplementario relativo a dicho artículo, disponible también online en: https://doi.org/10.1016/j.spl.2020.108719.

1.3.2.2

Diferencia de variables aleatorias independientes de recuento

Las distribuciones discretas definidas en $\mathbb{Z}$ han despertado la atención de muchos investigadores. Una forma simple de construir distribuciones con soporte en los enteros es tomar la diferencia de dos va's de recuento, X_h, $h = 1, 2$, independientes, cada una con fgp $f_{X_h}(z)$. Por tanto, considerando $X = X_1 - X_2$, cálculos directos muestran que la fgp de X es,

$$f_X(z) = f_{X_1}(z) f_{X_2}(z^{-1}).$$

De nuevo, si $f_X(z)$ satisface la ecuación (1.13), entonces, por la Proposición 1, la fmp de la va X puede ser obtenida, de forma recursiva, por (1.17).

En el Capítulo 5 (Baena-Mirabete & Puig, 2020) se muestran algunos ejemplos concretos de recurrencias obtenidas para la diferencia de dos va's de recuentos, ambas independientes. Es remarcable la simple relación de recurrencia encontrada para el cálculo de

las probabilidades de la distribución Skellam[13], que usualmente son obtenidas usando la función de Bessel modificada de primer tipo. Concretamente, sea $X \sim Skellam(\lambda_1, \lambda_2)$, con $\lambda_1, \lambda_2 > 0$, entonces las probabilidades $p_n = P(X = n)$, $n \in \mathbb{Z}$, pueden obtenerse recursivamente como sigue,

$$\begin{cases} p_{n+1} = \frac{\lambda_1}{\lambda_2} p_{n-1} - \frac{n}{\lambda_2} p_n, & n = 0, 1, ... \\ p_{n-1} = \frac{n}{\lambda_1} p_n + \frac{\lambda_2}{\lambda_1} p_{n+1}, & n = 0, -1, ... \end{cases}$$

Como se describe en el Capítulo 5 (véase material suplementario correspondiente a Baena-Mirabete & Puig (2020)), el tiempo de ejecución obtenido en la evaluación de las probabilidades de la Skellam usando la recurrencia es claramente inferior a la evaluación directa a partir de la función de Bessel[14]. Además, como se muestra en Baena-Mirabete & Puig (2020), la recurrencia anterior encontrada para la distribución Skellam es fuertemente estable de acuerdo a Panjer & Wang (1993) siempre y cuando se haga un uso combinado de ambas direcciones según el valor que tome n[15].

1.4 Discusión

El objetivo de esta tesis es proveer algunos avances en el análisis de series temporales discretas así como relevantes aplicaciones de la metodología propuesta. Concretamente, proponemos: 1) modelos multiestados basados en cadenas de Markov de larga persistencia (Capítulo 3); modelos de mixturas de cadenas de Markov con dos estados posibles

[13] La distribución más frecuente definida en $\mathbb{Z}$ es la distribución Skellam, obtenida como la diferencia de dos va's independientes de Poisson. Véase, por ejemplo, Koopman et al. (2017) para un estudio de los cambios en los precios de cuatro activos financieros usando dicha distribución.

[14] El tiempo de ejecución para el cálculo de las probabilidades p_n de una distribución de Skellam con parámetros $\lambda_1 = 10$ y $\lambda_2 = 15$ con $n = -25, ..., 25$ ha sido de 123.73 ($s \times 10^6$) usando la relación de recurrencia y de 1575.13 ($s \times 10^6$) por evaluación directa usando la función de Bessel, por tanto, este último es aproximadamente 13 veces más lento.

[15] Teoremas 7 y 9 por Panjer & Wang (1993) son remarcables porque establecen la estabilidad (en un sentido fuerte) de recurrencias lineales con coeficientes no negativos. En el caso de la relación de recurrencia para la fmp de la Skellam, nótese que los coeficientes de la recursión 'hacia adelante' son todos no negativos, para $n < 0$, y por tanto, estable de acuerdo con Panjer & Wang (1993). Análogamente, los coeficientes de la recursión 'hacia atrás' son todos no negativos para $n > 0$.

para datos longitudinales (Capítulo 4) y, finalmente, un resultado teórico para el cálculo de probabilidades mediante relaciones de recurrencia para una gran familia de distribuciones de recuento, incluyendo aquellas con soporte entero en $\mathbb{Z}$ (Capítulo 5). Este último resultado es de especial interés para el análisis de series temporales de datos definidos en $\mathbb{Z}$ ya que permite evaluar, de forma eficiente, la función de verosimilitud que, en muchos casos, es calificada en la literatura como numéricamente intratable.

Los modelos descritos en el Capítulo 3, basados en cadenas de Markov de alto orden para estados ordenados, son especialmente apropiados para cadenas donde las transiciones entre estados son poco frecuentes y, cuando se producen, tienen lugar, generalmente, entre estados cercanos. Para este tipo de cadenas, que hemos denominado persistentes, no es necesario especificar todas sus transiciones de forma detallada sino que un resumen de las mismas es suficiente, reduciendo así el número de parámetros a estimar. Hemos aplicado estos modelos en el ámbito del riesgo de crédito. Como se ha comentado, el *Valor-en-Riesgo* (VaR) de una cartera crediticia, con un nivel de confianza α (%), es el cuantil estimado de la distribución de pérdidas correspondiente al valor que será excedido una fracción $1 - \alpha$ (%) de ocasiones por un inversor que invierta en dicha cartera. En la medición del riesgo, es de especial interés obtener estimaciones fiables de la probabilidad de transición entre *ratings* dado que estas estimaciones tienen un impacto directo tanto en la pérdida esperada como en el VaR de la cartera, en un cierto horizonte de tiempo fijado (generalmente, 1 año). Se ha analizado una muestra de 4747 emisores de bonos municipales en USA calificados por la agencia Standard & Poor's en el periodo comprendido entre los años 1990 y 2013 y se han obtenido estimaciones de las migraciones de *ratings* a partir de diferentes modelos propuestos basados en cadenas de Markov para estados persistentes. Se han simulado las exposiciones de una hipotética cartera de bonos a fin de obtener el VaR de dicha cartera usando estas probabilidades de transición estimadas. La idea es que si tenemos una cartera de bonos, cada uno evaluado en un cierto *rating*, y en el transcurso de 1 año dicha cartera evoluciona hacia *ratings* peores entonces el valor de la cartera disminuirá, produciéndose una pérdida. A un año vista, la pérdida[16] de la

[16]Generalmente, se suele calcular la pérdida de una cartera en valor positivo, es decir, una pérdida negativa significará una ganancia. Por tanto, por lo que acontece al cálculo del VaR, nos preocupará la cola derecha de la distribución de pérdidas estimada.

cartera es una variable aleatoria cuya distribución se estima a partir de las probabilidades de transición entre *ratings*. En la muestra[17] de emisiones utilizada, se ha obtenido que el VaR estimado por un modelo de Markov de orden 1, ampliamente usado en la literatura, es inferior a aquel obtenido por modelos de mayor orden. Este resultado es especialmente relevante ya que sugiere que en condiciones económicas adversas, el uso de modelos de Markov de orden 1 podría infraestimar el VaR de una cartera. Los resultados mostrados en el Capítulo 3 muestran, a modo de ejemplo, que el VaR estimado a partir de un modelo de orden 3 es aproximadamente un 32 % superior que el obtenido por un modelo de Markov de orden 1. Ello es debido a que un modelo de orden 1 no puede tener en cuenta efectos *momentum* que se producen en las migraciones de *ratings* y que sí son considerados en modelos de mayor orden.

En el Capítulo 4, se proponen diferentes modelos de mixturas basados en cadenas de Markov con dos estados posibles, enmarcado en el análisis de datos longitudinales. Los modelos presentados derivan de un estudio observacional realizado a un grupo de estudiantes a quienes les propusimos simular mentalmente las realizaciones de una moneda justa. La generación de secuencias aleatorias mentalmente producidas por individuos es una tarea muy común en el campo de la psicología experimental. Budescu (1987) propuso modelos de Markov (MC_k) para el análisis de secuencias aleatorias generadas por individuos. Sin embargo, uno de los principales resultados que hemos obtenido del experimento llevado a cabo es que las decisiones de 'cara-cruz' no pueden explicarse únicamente en base a los previos resultados observados. Modelos de mixturas de cadenas de Markov, que tienen en cuenta una heterogeneidad adicional (no observable) entre individuos, obtienen mejores resultados que los modelos estándar de cadenas de Markov (MC_k). Concretamente, hemos propuesto una extensión del trabajo de Park & Basawa (2002) para el caso de datos longitudinales. Desde este enfoque, mostramos que los individuos se desvían de

[17]La muestra utilizada de 4747 emisores de bonos municipales en USA, entre los años 1990 y 2013, calificados por la agencia Standard & Poor's, incluye los peores años de la 'crisis financiera de 2008' que se desató en Estados Unidos debido al colapso de la burbuja inmobiliaria en el año 2006, que provocó aproximadamente en octubre de 2007 la llamada crisis de las hipotecas *subprime*. Las repercusiones de la crisis hipotecaria comenzaron a manifestarse de manera extremadamente grave desde inicios de 2008, contagiándose primero al sistema financiero estadounidense, y después derivando en una crisis económica a escala internacional.

la aleatoriedad cuando intentan generar una serie aleatoria debido a la combinación de dos procesos: el primero, produce un exceso de respuestas alternadas y una tendencia a equilibrar la distribución de resultados 'cara-cruz' en fragmentos cortos. El segundo proceso produce demasiadas repeticiones. Además, el papel que juega cada uno de estos dos procesos en el experimento depende del individuo.

Finalmente, en el Capítulo 5, generalizamos los resultados propuestos por Sundt (1992) y proveemos relaciones de recurrencia para el cálculo de la fmp de una gran familia de distribuciones de recuento, incluyendo aquellas con soporte entero en $\mathbb{Z}$. En relación con el objeto principal de esta tesis, que como su título indica, es proponer avances en el análisis de series temporales discretas, este resultado es especialmente relevante ya que permite obtener, de forma eficiente, las probabilidades necesarias para el cálculo de la función de verosimilitud en muchos problemas considerados computacionalmente intratables en la literatura. Por ejemplo, en los últimos años, los modelos INAR han sido extensamente utilizados en el análisis de series temporales de recuento. El clásico modelo INAR(1) resulta de la convolución suma de una distribución Binomial y una Poisson. La exacta, pero computacionalmente ineficiente, fmp para un INAR(1) puede encontrarse, por ejemplo, en Freeland & McCabe (2004). El resultado obtenido en el Capítulo 5 permite obtener relaciones de recurrencia para el cálculo de las propabilidades de un modelo INAR(p) para cualquier orden p. Por otra parte, el análisis de series temporales con soporte entero en $\mathbb{Z}$ ha recibido mucho interés en años recientes. Por ejemplo, en el ámbito financiero, el análisis de los cambios (discretos) en el precio de ciertos activos financieros es especialmente de interés para el estudio de la volatilidad estocástica intradía en los mercados financieros. Koopman et al. (2017) utilizaron la distribución de Skellam para tal fin, mientras que Barndorff-Nielsen et al. (2012) propusieron la distribución de la diferencia de dos va's Binomiales Negativas, ambas independientes, para modelizar cambios de precios, también en un contexto financiero. Las relaciones de recurrencia encontradas en el Capítulo 5 permiten obtener de forma eficiente la fmp de la diferencia de va's independientes con una fgp satisfaciendo cierta forma funcional. Adicionalmente, en el mismo Capítulo 5 (Baena-Mirabete & Puig, 2020), en el apartado de material suplementario relativo a dicho artículo, se ha realizado un estudio del comportamiento numérico de algunas de las recurrencias consideradas. Los teoremas 7 y 9 por Panjer & Wang (1993)

son remarcables porque establecen la estabilidad (en un sentido fuerte) de recurrencias lineales con coeficientes no negativos. Nosotros mostramos que para algunas recurrencias, como por ejemplo la obtenida para la fmp de una distribución de Skellam, es posible obtener una estabilidad numérica fuerte combinando ambas direcciones (‘hacia adelante’ y ‘hacia atrás’) en la evaluación de la recurrencia.

CAPÍTULO 2

Conclusiones y líneas futuras de investigación

En los últimos años, han aparecido multitud de técnicas enfocadas al análisis de series temporales categóricas mediante modelos de cadenas de Markov. Uno de los mayores retos en estos modelos consiste en simplificar el número de parámetros ya que este aumenta exponencialmente con el orden de la cadena. Numerosos estudios se han centrado en este hecho desde enfoques bien diversos (Raftery (1985); Fokianos & Kedem (2003); Kharin & Petlitskii (2007); Kharin & Maltsau (2014)). En esta tesis hemos mostrado que, en algunas aplicaciones concretas, puede existir una natural reducción del número de parámetros al explorar cadenas de Markov de alto orden. Por ejemplo, en lo que se refiere al funcionamiento de la memoria humana, los individuos tienen una capacidad limitada de procesamiento (véase Miller (1956); Saaty & Ozdemir (2003)). Tan solo hay que pensar qué sucede cuando intentas replicar mentalmente, y de la forma más fidedigna posible, los resultados de una moneda justa. Recuerdas, seguramente con exactitud, tus últimos dos o tres resultados simulados pero más allá, solo recuerdas vagamente si simulaste un mayor número de caras o de cruces. Por tanto, es asumible pensar que los individuos pueden recordar de forma exacta los últimos resultados de una cierta secuencia pero únicamente un resumen de los mismos para retardos más alejados. Como hemos visto, esto nos ha llevado, en este contexto, a modelos basados en cadenas de Markov donde

no ha sido necesario especificar de forma detallada todas las transiciones de la cadena. En otro contexto, el de la medición del riesgo de crédito, la dinámica de las migraciones de *ratings* nos ha permitido de forma natural reducir el número de parámetros en modelos de cadenas de Markov de alto orden. Las transiciones entre *ratings* son poco comunes y cuando se producen, suelen tener lugar entre *ratings* cercanos, lo que nos ha permitido resumir ciertas transiciones entre estados poco frecuentes. Finalmente, se ha comentado que en algunas aplicaciones una combinación de más de una cadena de Markov puede ser una interesante alternativa. En el experimento de generación aleatoria llevado a cabo, hemos constatado que los individuos se desvían de la aleatoriedad cuando intentan generar mentalmente secuencias aleatorias debido a la combinación de dos procesos: el primero, produce un exceso de respuestas alternadas y, el segundo, un exceso de repeticiones. Como hemos visto, modelos de mixturas de cadenas de Markov como el planteado aquí obtienen mejores resultados que otros modelos más sencillos.

El objetivo principal de esta tesis ha sido introducir metodologías novedosas para el análisis de series temporales categóricas. La primera y segunda parte de la tesis exponen las metodologías propuestas, que han sido ilustradas con dos aplicaciones relevantes en ámbitos distintos. La tercera parte de la tesis expone un resultado teórico que es de gran utilidad para el cálculo de probabilidades de sumas y diferencias de variables aleatorias de recuento por medio de relaciones de recurrencia. Este resultado es de gran interés para calcular eficientemente la función de verosimilitud en muchas aplicaciones que usan estas distribuciones. En el ámbito de las series temporales, hemos comentado, por ejemplo, el uso de este resultado teórico para la estimación de los modelos INAR, extensamente utilizados para el análisis de series temporales de recuento.

A continuación, se exponen diferentes líneas de investigación futura, centradas principalmente en las aplicaciones descritas en esta tesis.

Una de las aplicaciones presentadas en esta tesis ha sido en el ámbito de la medición del riesgo financiero (véase Capítulo 3 (Baena-Mirabete & Puig, 2018)). La estimación de las probabilidades de transición entre *ratings* es crucial para las entidades financieras ya que estas son utilizadas para el cálculo del *Valor-en-Riesgo* (VaR) de una cartera de títulos (p.ej. bonos) y también en la asignación del capital económico, entre otras aplicaciones. El enfoque planteado aquí, extensamente utilizado en la literatura, ha sido estimar

la probabilidad de transición en el horizonte de 1 año, es decir, considerando únicamente el estado del *rating*, para un cierto título, al inicio y al final de cada año natural. Esto implica obviar las transiciones entre *ratings* que se producen intra-año. Una interesante aportación al trabajo presentado en esta tesis podría ser considerar la evolución de los *ratings* de forma continua en el tiempo, teniendo en cuenta así las posibles transiciones que pudieran producirse dentro de un mismo año natural. Un buen punto de partida podría ser el trabajo descrito en Bellotti & Crook (2014) donde los autores propusieron un modelo de análisis de la supervivencia discreta para estimar cambios en la probabilidad de morosidad bajo escenarios macroeconómicos adversos. Por otra parte, resultaría novedoso explorar, en este ámbito de aplicación, los modelos de mixturas de cadenas de Markov propuestos en el Capítulo 4 (Baena-Mirabete et al., 2019). En esta línea, es interesante el enfoque descrito en Boreiko et al. (2019) para estimar la dinámica de las migraciones de *ratings* a partir de modelos ocultos de cadenas de Markov.

Finalmente, en el Capítulo 4 (Baena-Mirabete et al., 2019) se muestran los resultados de un estudio observacional realizado a un grupo de estudiantes del grado de Biología de la Universitat Autònoma de Barcelona. En dicho experimiento, les propusimos simular mentalmente las realizaciones de una moneda justa de la forma más fidedigna posible. La generación de secuencias *aleatorias* mentalmente producidas por individuos es una tarea muy común en el campo de la psicología experimental. Tradicionalmente, este tipo de estudios se han realizado sobre adultos sanos, como el presentado en esta tesis. Sin embargo, existe también algunos estudios al respecto realizados sobre pacientes con algún tipo de trastorno neurológico. Los pacientes con enfermedad de Alzheimer (EA) leve muestran dificultades para realizar actividades cotidianas debidas a la afectación del control atencional, las cuales describen diferencias cualitativas significativas entre el envejecimiento normal y el patológico. Las tareas de generación aleatoria permiten valorar esta capacidad. García-Viedma et al. (2015) diseñaron una tarea de generación aleatoria que fue aplicada un grupo de personas mayores sanas y a un grupo de pacientes con EA inicial. Los resultados revelaron diferencias significativas entre ambos grupos, mostrando los pacientes más dificultades para evitar secuencias estereotipadas y cambiar de estrategias. Actualmente, estamos trabajando con este equipo en un experimento que se está realizando a adultos sanos (grupo de control) y a pacientes con Alzheimer o con deterioro

cognitivo leve. El objetivo ambicioso es utilizar los modelos descritos en el Capítulo 4 para desarrollar una herramienta que pueda ser aplicada en la detección temprana de la EA.

CAPÍTULO 3

Parsimonious higher-order Markov models for rating transitions

This Chapter corresponds to the contents of (Baena-Mirabete & Puig, 2018).

Abstract

We propose several parsimonious models for higher-order Markov chains, applied to the study of municipal rating migrations in credit risk. In full-parameterized Markov chain models, the number of parameters increases very rapidly as the order in the Markov chain grows and this can yield biased estimates when certain sequences of states are rare. For some processes, as in the case of credit ratings, this problem is accentuated because the transitions between distant states are unlikely (*persistent* transitions). We introduce the *Short* and *Long-persistence* models and compare them with the full-parameterized Markov chain, achieving a better fit with a lower number of parameters. Furthermore, downgrade *momentum* effects are found in the rating process, which are consistent with recent empirical findings.

3.1 Introduction

The main objective of rating agencies, such as Moody's, Standard and Poor's (S&P) or Fitch is to evaluate the capacity of a debt issuer to pay back this debt and also to estimate the probability of this issuer falling into default. To this end, rating agencies provide a risk scale that classifies the debt into homogeneous discrete risk classes based on the solvency of the issuer. The rating classes assigned by these agencies are expressed on an ordinal scale, where the position in the scale reflects the ability of the issuer to pay its obligations. For instance, Standard and Poor's rate borrowers on a scale from AAA to D, where AAA is the highest and D is the lowest score. The rating AAA indicates a strong financial capacity to cope with debts and the rating D indicates the lowest quality firm, which is usually in default and with a low likelihood of recovering principal or interest on their debt. Obligations rated from AAA to BBB are called *investment* grade and obligations with ratings lower than BBB are considered *speculative* investments. Local governments often issue bonds to the public in order to borrow resources for capital investment. Creditworthiness, as reflected in bond ratings, is of great interest to municipalities as it directly affects the cost and ability to borrow money. A lower rating assigned to its debt increases borrowing costs and limits fund-raising options. Municipalities that experience a decline in their economic health will be especially concerned about how these developments will impact their future bond ratings (Palumbo & Shick, 2006). On the other hand, municipal bonds are a staple of many fixed-income portfolios in banks and other financial institutions. The evolution of a bank's bond portfolio towards lower ratings has a negative impact on the value of this portfolio because the price of the bonds depends not only on the current prime rate but also on the risk of default on the debt (Kliger & Sarig, 2000). Hence, the analysis and study of rating transitions over time is a crucial matter for most financial entities (European Banking Authority, 2014). In addition, an understanding of the particularities of migration ratings together with reliable estimates can also be useful inputs for more sophisticated financial products such as credit derivatives (Acharya et al., 2002).

Numerous empirical studies have specially focused on corporate bonds, which pro-

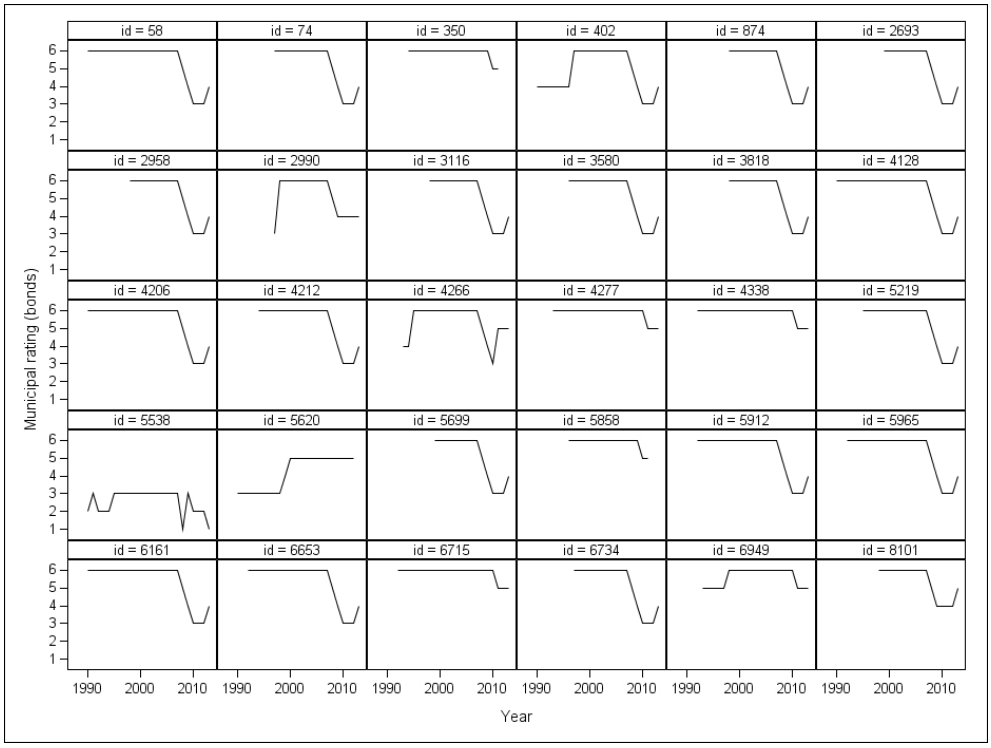

Figure 3.1: Sequences of ratings over time for 30 randomly selected debt issuers from the municipal rating sample.

vide some insight into the factors that may affect bond credit rating transitions. In Nickell et al. (2000), a 1-order Markov chain was considered, and it was found that the transition probabilities depend on the state of the economy, the type of industry or the issuer's country. However, research and practice in this field have increasingly converged towards a general consensus: knowledge of the previous rating is insufficient to explain the dynamic of rating migrations. The simple idea is that under a 1-order Markov chain, two debt issuers with an identical previous rating should have the same probability of being downgraded or upgraded in the current state. But several empirical studies have demonstrated quite the opposite. An interesting contribution can be seen in Lando & Skϕdeberg (2002) who consider the estimation of credit rating transitions based on continuous-time observations. The main advantage of this approach, which falls within survival analysis, is that the estimates are based on the full story of the rating transitions. Basically, two conclusions can be drawn from this study: the first is known in the literature as *rating momentum*, meaning that a previous downgrade (upgrade) increases the probability of

being downgraded (upgraded) again. The second finding of this study concerns the relationship between the time since the last entry in the current rating and the transition probability. The authors found that the longer the *persistence* in a rating, the less likelihood there is of a change of state in the immediate future. Other studies, such as Fuertes & Kalotychou (2007), support these results by finding evidence of duration dependence and downgrade *momentum* effects in the rating process. In line with these latest findings, latent factor models have been used in such a way as to allow for serial correlations in the unobserved risk factors and hence, implicitly, *momentum* effects. For instance, in McNeil & Wendin (2006), a class of generalized linear mixed models was fitted for ordered polytomous responses by assuming a Markov process for the temporal random effects. In this model, the authors take implicitly serial correlations into account in migration counts. However, the probability of transition in their model is not directly dependent on previous rating migrations. A similar study can be seen in Stefanescu et al. (2009) where the authors propose a Bayesian hierarchical model that describes the credit rating process and take into account ratings migration dependence through unobserved systematic macroeconomic factors. Their model, in contrast to McNeil & Wendin (2006), takes into account the initial rating to capture rating transition patterns that cannot be accounted for in the latent factor model.Kim & Young (2008) proposes a non-Markovian random effects multinomial model to estimate transition probabilities of credit ratings. In this model, not only are characteristic factors of the rating taken account, but also a random effect due to uncertainty within each rating that cannot be explained by such factors. However, as pointed out by the authors, the estimation results of the random effects model differs somewhat from the observed transitions. More sophisticated models have appeared, such as Korolkiewicz & Elliott (2008), in which credit quality is modeled by a hidden Markov model and the credit transition probabilities are estimated using the EM algorithm. In Wozabal & Hochreiter (2012), a coupled Markov chain model is proposed to model rating transitions. However, as noted by the authors, to use this model in practice, more sophisticated estimation routines are required in order to obtain realistic parameter estimates.

However, the analysis of higher-order Markov chain models in the study of rating transitions has not been addressed. This is because, as we shall see, in a full-parameterized

Markov chain model, the number of parameters to be estimated increases very rapidly as the order in the Markov chain grows. Moreover, in *persistent* multi-states, such as credit rating, in which the probability of each asset moving from one rating to another is small, the sparseness of data often leads to convergence failures in the estimation and unrealistic transition probabilities. This paper addresses this issue and makes two main contributions. First, we propose a parsimonious higher-order Markov chain to estimate rating transitions for *persistent* multi-states. In particular, the probability of a transition in our model depends not only on the initial rating but also on previous ratings or a summary thereof. Our second contribution consists of proposing a partial proportional odds model for sparse and diagonally dominant matrix migration in order to relax the proportional odds assumption of most previous studies on credit rating transitions.

This paper is structured as follows: Section 3.2 offers a brief description of the data. In Section 3.3, full-parameterized higher-order Markov chain models and some reduced forms are introduced and fitted. In Section 3.4, a *Long-persistence* model is proposed to estimate a higher-order Markov chain more parsimoniously. In Section 3.5, the main results of the estimation and *out-of-sample* forecast are presented. Section 3.6 discusses the economic significance of the model in order to generate accurate Value at Risk (VaR) measures. Section 3.7 concludes the article.

3.2 Municipal ratings data

The source of the dataset that we analyze here was the Standard and Poor's Ratings History Disclosure files that cover the credit ratings produced by Standard and Poor's Rating Services. The disclosure files are provided in an eXtensible Business Reporting Language (XBRL) format that is posted monthly on the Standard and Poor's website. The data refer to 4747 municipal bonds issuers in the United States rated by the agency at some time during the period from 1990 through 2013. The sample covers state and local governments, public hospitals, colleges and universities, and other councils that have a long-term debt rating. Municipal bonds are usually issued, for instance, by a state or local government in order to finance their infrastructure needs or raise capital for civil projects. The issuer of a

municipal bond receives a cash payment at the time of issuance with the promise to repay the money (capital and interest) to their investors in the form of periodical payments (for example, annually). The ability of the issuer to pay back the debt is determined by the rating agencies, such as Standard and Poor's, who assign a rating that is made available to potential investors. As described in Standard and Poor's (2013), the municipal bond rating is based on the assessment and scoring of several components among which economy receives the highest weight (30%). With all this information, Standard and Poor's assign an initial indicative rating that results from a weighted average of these key components. The final rating may differ from this initial indicative rating due to the adjustment factors based on trends and comparisons with similarly rated peers.

There are two reasons for the focus on municipal bonds. On the one hand, the literature on municipal bond ratings, which is less intensive than that on corporate bond ratings, has focused on the economic factors that affect the municipal rating and how the characteristics of the local area or municipality may be determinants. From this approach, an interesting study can be seen in Sarmiento (2006), which estimates a spatial logistic model to evaluate the statistical effect of conditions of communities on municipal bonds rated by Standard and Poor's. In Richard (2011), a study based on a survey of all US municipalities with a population of over 10,000 shows that microvariables, such as the quality of personnel recruited and others variables that are also related to the public institution's human capital are predictors of higher municipal bond ratings. However, municipal rating migration and consequently the way that previous ratings might affect the current rating has been less widely studied. The second reason for focusing on the analysis of municipal bonds rating is the fact that these are less likely to be updated than corporate bonds, which have more volatile ratings (Cornaggia et al., 2015), thus accentuating the problem relating to the few observations in the off-diagonal matrix migration that makes it difficult to obtain reliable estimates. This leads to the need for parsimonious models in order to explore higher-order dependencies without exponentially increasing the number of parameters.

Table 3.1 reports for the entire sample of US Municipal Ratings: i) total rating assignments and migrations, ii) rating duration in months, and iii) proportion of up/downgrades. Most of the assignments are to AAA, AA, A or BBB ratings, i.e investment grade. Rating assignments to speculative grade (lower than BBB) are less frequent. Downgrades are

Table 3.1: Summary Statistics of US Municipal Ratings. N_i denotes the total assignments to state i. $N_{i\to j, i\neq j}$ is the total number of migrations from state i to state j. T_i is the average time (measured in months) spent in each state. $N_{i,u}$ and $N_{i,d}$ are respectively the proportion of upgrades and downgrades relative to total migrations from rating i.

State i	N_i	$N_{i\to j, i\neq j}$	T_i	$N_{i,u}\%$	$N_{i,d}\%$
AAA	4325	4283	59	0	100
AA	5780	2893	9	24.51	75.49
A	5132	3456	10	65.63	34.38
BBB	1719	1618	20	96.79	3.21
BB	66	43	13	60.47	39.53
B	23	15	17	53.33	46.67
CCC	18	12	22	75.00	25.00
D	6	0	-	-	-
Total	17069	12320	28	37.22	62.78

more likely than upgrades in AAA and AA ratings because of a deterioration in municipal credit quality since 2008, which is a reflection of the global economic crisis, leading to downgrades in these ratings. As illustration, Figure 3.1 shows plots of rating paths over the time for a set of issuers that were drawn randomly from the municipal rating sample to make it easier to view. In contrast, for the rest of ratings, upgrades are more likely than downgrades, particularly for BBB ratings (i.e the boundary between the investment grade and the speculative grade), the proportion of upgrades is around 97%. The average time spent in each rating exceeds one year, with the exception of the AA and A ratings, with 9 and 10 months respectively. The average of duration in AAA ratings is 59 months because, historically, municipal ratings have been highly rated and very stable over time, showing low transitions between risk classes. We consider for constructing the estimation sample, the issuer's rating at the end of year and therefore, rating transitions within a year, less common in municipal bonds, are not taken into account. For each issuer, we consider its history of annual ratings from the first credit rating until the last rating assignment. Therefore, right censored ratings occurring between the last rating update and the end of the sample windows (i.e year 2013) are excluded from the history of an issuer. Figure 3.2

shows, for the estimation sample, the number of rated issuers per year: a large proportion of ratings cover the period between 2008 and 2011, specifically 44% of the total rated issuers.

Municipal rating bonds are grouped into six risk classes (ordered from highest quality to lowest): AAA, AA, A, BBB, BB/B, CCC/CC/C/D. There are also other increments to provide more gradations, including + or - symbols (for instance AA+), but because of their low frequency, these qualifiers have been eliminated. For example, in our data, basis AA corresponds to any of the ratings AA+, AA or AA-. This classification is also used in McNeil & Wendin (2006) and in Stefanescu et al. (2009) for corporate bonds rating. We have merged BB and B into a single rating class, BB/B, due to the low observed frequencies of ratings in the sample. In the same way, the ratings CCC/CC/C and the default state D were merged into a single rating class due to the low number of observations. Over time, the risk of default occurring on a municipal bond has been extremely rare (Standard and Poor's, 2014) and indeed, while the risk of default is a crucial consideration for corporate bonds, it is less significant in the municipal bond market. To facilitate the analysis, the rating classes were recoded into a numerical format from 1 to 6, 1 being the lowest rating (from CCC to D) and 6 the highest (AAA).

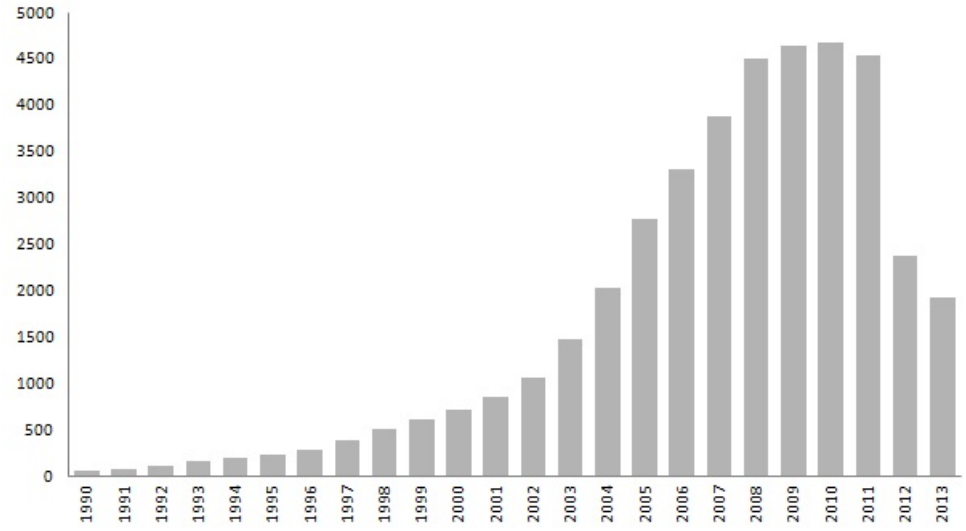

Figure 3.2: Number of rated issuers per year for the sampled ratings.

3.3

Higher-order Markov chain and some common parameter reductions

The Markov chain is a well-known probabilistic model used to explore dependencies between successive observations. The usual 1-order Markov chain model considers that, given the present, the future is conditionally independent from the past. However, in some situations the future depends not only on the present but on the last observations.

3.3.1 Full-parameterized Markov chain model

Consider a discrete-time random variable X_t taking values in a finite set of states $1, ..., m$. From now on and without loss of generality, subscript t refers to time point (measured in years), thus X_t refers to the state in the current year t and X_{t-k} is the state k years before year t with $k \leq t$. We have a k-order Markov chain if,

$$P(X_t = i_t | X_0 = i_0, X_1 = i_1, ..., X_{t-1} = i_{t-1}) =$$
$$P(X_t = i_t | X_{t-k} = i_{t-k}, ..., X_{t-1} = i_{t-1})$$

where $i_0, i_1, ..., i_t \in \{1, 2, ..., m\}$. We are interested in multi-state series, specifically the number of states that we will consider is $m = 6$. Each k-order Markov chain model is completely characterized by its transition probabilities. For instance, for $k = 3$ these transition probabilities can be arranged according to the matrix form in Table 3.2, in which the sum of each row is equal to one. In this specific case, this model has 1080 parameters to be estimated. Unfortunately, the number of parameters of these full-parameterized models increases very rapidly: a k-order Markov chain model with 6 states has $6^k(6-1)$ parameters. From now on, MC_k will denote a full-parameterized k-order Markov chain model. MC_0 model will correspond to a series of outcomes where the states at different times are independent events.

Direct calculations show that the log-likelihood function of an MC_k model, given the first k observations, is

$$l = \sum_{i_k,...,i_1} \sum_{s=1}^{6} n_{s;i_k,...,i_1} \log(p^s_{i_k,...,i_1}),$$

where $\sum_{j=1}^{6} p^j_{i_k,...,i_1} = 1$; $n_{s;i_k,...,i_1}$ indicates the number of observed transitions of the form $X_{t-k} = i_k$, ..., $X_{t-1} = i_1$ and $X_t = s$ and $p^s_{i_k,...,i_1}$ its transition probability. It imme-

Table 3.2: Transition probabilities in tabulated form for a full-parameterized 3-order Markov chain model.

			X_t					
X_{t-3}	X_{t-2}	X_{t-1}	1	2	3	4	5	6
1	1	1	p^1_{111}	p^2_{111}	p^3_{111}	p^4_{111}	p^5_{111}	p^6_{111}
1	1	2	p^1_{112}	p^2_{112}	p^3_{112}	p^4_{112}	p^5_{112}	p^6_{112}
	...		...	...	...	...	...	...
1	1	6	p^1_{116}	p^2_{116}	p^3_{116}	p^4_{116}	p^5_{116}	p^6_{116}
1	2	1	p^1_{121}	p^2_{121}	p^3_{121}	p^4_{121}	p^5_{121}	p^6_{121}
1	2	2	p^1_{122}	p^2_{122}	p^3_{122}	p^4_{122}	p^5_{122}	p^6_{122}
	...		...	...	...	...	...	...
1	2	6	p^1_{126}	p^2_{126}	p^3_{126}	p^4_{126}	p^5_{126}	p^6_{126}
	...		...	...	...	...	...	...
1	6	6	p^1_{166}	p^2_{166}	p^3_{166}	p^4_{166}	p^5_{166}	p^6_{166}
2	1	1	p^1_{211}	p^2_{211}	p^3_{211}	p^4_{211}	p^5_{211}	p^6_{211}
	...		...	...	...	...	...	...
6	6	6	p^1_{666}	p^2_{666}	p^3_{666}	p^4_{666}	p^5_{666}	p^6_{666}

diately follows that the maximum likelihood estimate (MLE) of the transition probability $p^s_{i_k,\ldots,i_1}$ is $\hat{p}^s_{i_k,\ldots,i_1} = n_{s;i_k,\ldots,i_1}/(\sum_{j=1}^6 n_{j;i_k,\ldots,i_1})$. So, the MLE can easily be calculated by cross-tabulation and the standard errors can be estimated using the usual expression for the variance of an estimated proportion, $se^2(\hat{p}^s_{i_k,\ldots,i_1}) = \hat{p}^s_{i_k,\ldots,i_1}(1-\hat{p}^s_{i_k,\ldots,i_1})/(\sum_{j=1}^6 n_{j;i_k,\ldots,i_1})$. When k increases, many of the cells of the cross-tabulation may be empty ($n_{j;i_k,\ldots,i_1} = 0$) and so it is not possible to make inference. Consequently, the number of parameters must be reduced in order to explore a higher order dependence.

We will focus on Markov chains in which the probability of remaining unchanged at the same state is larger than the one of moving to any other state. We define a *persistent* chain as one where its one-step transition matrix is diagonally dominant, that is $P(X_t = s|X_{t-1} = s) > 0.50$. The larger this probability, the more unlikely the transitions between states are and we will say that the process is highly *persistent*. From now on, we will refer to the states of a *persistent* chain as *persistent* states.

As in Berchtold & Raftery (2002), to assess and compare the *in-sample* performance of several k-order Markov models, we use the Bayesian Information Criterion (BIC). This is defined by $BIC = -2\hat{l} + q\log(n)$, where $\hat{l}$ is the estimated log-likelihood, q the number of parameters in the model and n the sample size minus the order k of the model (for a full MC_3 model, $q = 1080$ and n is the number of observations minus 3). A model with a lower BIC value should be preferable to a model with a higher BIC value.

3.3.1.1

Estimation of the MC_k models

For the discrete state series of municipal ratings with $m = 6$ possible states, we have fitted the MC_k models, for $k = 0, 1, 2, 3$ and we have selected the best model according to the BIC statistic. A two-way table for the consecutive ratings is shown in Table 3.3, detailing the estimated transition probabilities for an MC_1 model and their corresponding standard errors and the observed frequencies (in brackets). Note that the 1-year transition matrix is diagonally dominant as a common feature of credit rating transitions. As outlined in Fuertes & Kalotychou (2007), migrations are typically towards neighboring ratings and transitions between distant ratings are less likely. We will refer to this property of credit ratings as the *persistence* of ratings. Consequently, the off-diagonal probabilities are poorly estimated because of the few observations in their corresponding cells: the greater the distance to the diagonal, the lower the number of observations. For instance, note in Table 3.3 that the transition probability from rating 1 to rating 6 is estimated with only one observation. This becomes more relevant in an MC_k model with a large k because the number of empty (or close to zero) cells increases and their estimation is unreliable. For this reason, for our dataset we have only explored the full-parameterized k-order Markov chain models from $k = 0$ until $k = 3$. Table 3.4 shows the goodness of fit results. Note that, according to the BIC criterion, the best MC_k model is obtained for $k = 2$. MC_0, the independence model, produces one of the worst results. Even when adding a random effect by issuer, MC_0 was worse than the higher order Markov chain models explored.

However, as mentioned previously, the number of parameters of these fully parameterized models increases very rapidly. The following sections present some ways to reduce

Table 3.3: 1-year estimated probabilities for the MC_1 model of the 36 possible pairs of ratings $\{X_{t-1}, X_t\}$. Observed frequencies and standard errors for the estimates (in brackets).

	X_t					
X_{t-1}	1	2	3	4	5	6
1	0.7586	0.0690	0.0345	0.0690	0.0345	0.0345
	(22;0.0795)	(2;0.0471)	(1;0.0339)	(2;0.0471)	(1;0.0339)	(1;0.0339)
2	0.1053	0.8316	0.0316	0.0105	0.0211	0.0000
	(10;0.0315)	(79;0.0384)	(3;0.0179)	(1;0.0105)	(2;0.0147)	(0;0.0000)
3	0.0009	0.0081	0.6735	0.3136	0.0033	0.0006
	(3;0.0005)	(27;0.0016)	(2238;0.0081)	(1042;0.0080)	(11;0.0010)	(2;0.0004)
4	0.0003	0.0021	0.2709	0.6638	0.0594	0.0035
	(1;0.0003)	(8;0.0007)	(1013;0.0073)	(2482;0.0077)	(222;0.0039)	(13;0.0010)
5	0.0000	0.0000	0.0064	0.2203	0.7645	0.0088
	(0;0.0000)	(0;0.0000)	(37;0.0010)	(1276;0.0054)	(4429;0.0056)	(51;0.0012)
6	0.0001	0.0002	0.0010	0.0137	0.2101	0.7750
	(1;0.0001)	(3;0.0001)	(14;0.0003)	(198;0.0010)	(3042;0.0034)	(11222;0.0035)

the number of parameters in an MC_k model. We will also show that in processes where states are *persistent* over time, there is no need to use the detailed sequence of states and a summary of these may be sufficient. Standard optimization packages like the VGLM function in R (Yee, 2010) or the NLMIXED procedure in SAS v9.3 (SAS Institute Inc., Cary, NC, USA), can be used to fit these models.

3.3.2 A reduced-form Markov chain model

One way to reduce the number of parameters in an MC_k, when the states are ordered, is to consider a multinomial cumulative (logistic) regression model as follows,

$$P(X_t \leq s | X_{t-k} = i_{t-k}, ..., X_{t-1} = i_{t-1}) = \frac{\exp\{\psi(s; i_{t-k}, ..., i_{t-1})\}}{1 + \exp\{\psi(s; i_{t-k}, ..., i_{t-1})\}} \qquad (3.1)$$

Table 3.4: BIC statistics and the number of parameters (q) for several MC_k fitted models. $\hat{l}$ is the estimated log-likelihood. Sample size $n = 27459$. The first 3 observations for each subject are removed from the sample. MC_0^* refers to the multinomial model with subject-specific random effect for each issuer.

Model	No.params.(q)	$-2\hat{l}$	BIC
MC_0	5	72469	72520
MC_0^*	6	67894	67945
MC_1	30	35286	35592
MC_2	180	32030	33870
MC_3	1080	30487	41525

where the linear predictor has the form,

$$\psi(s) := \psi(s; X_{t-k}, ..., X_{t-1}) = \beta_{s,0} + \sum_{i=1}^{k} \sum_{j=1}^{m-1} \beta_{s,j}^{(i)} I_{\{X_{t-i}=j\}}, \tag{3.2}$$

where $\beta_{s,0}$ and $\beta_{s,j}^{(i)}$ are the intercepts and the slope parameters respectively, and $I_{\{X_{t-i}=j\}}$ are indicator functions taking the value 1 if $X_{t-i} = j$ and 0 otherwise. Therefore, $m-1$ *logit* equations are defined for $s = 1, 2, ..., m-1$ where the coefficients are different across the same. We find it convenient to note the above linear system in matrix form as

$$\Psi = \eta_0 + B^{(1)} \cdot Y_1 + ... + B^{(k)} \cdot Y_k \tag{3.3}$$

where $\Psi = (\psi(1), ..., \psi(m-1))^T$ is the vector of *logit* probabilities, $\eta_0 = (\beta_{1,0}, ..., \beta_{m-1,0})^T$ is the vector containing the intercepts and $Y_i = (I_{\{X_{t-i}=1\}}, ..., I_{\{X_{t-i}=m-1\}})^T$ for each lagged i-period $i = 1, ..., k$. Finally, $B^{(i)}$ is the square matrix of parameters as follows,

$$B^{(i)} = \begin{bmatrix} \beta_{1,1}^{(i)} & \beta_{1,2}^{(i)} & \beta_{1,3}^{(i)} & ... & \beta_{1,m-1}^{(i)} \\ \beta_{2,1}^{(i)} & \beta_{2,2}^{(i)} & \beta_{2,3}^{(i)} & ... & \beta_{2,m-1}^{(i)} \\ & & ... & & \\ \beta_{m-1,1}^{(i)} & \beta_{m-1,2}^{(i)} & \beta_{m-1,3}^{(i)} & ... & \beta_{m-1,m-1}^{(i)} \end{bmatrix}$$

Therefore, the transition probability $p_{i_k,...,i_1}^s$ is calculated, for $s > 1$, as

$$p_{i_k,...,i_1}^s = \frac{\exp\{\psi(s; i_{t-k}, ..., i_{t-1})\}}{1 + \exp\{\psi(s; i_{t-k}, ..., i_{t-1})\}} - \frac{\exp\{\psi(s-1; i_{t-k}, ..., i_{t-1})\}}{1 + \exp\{\psi(s-1; i_{t-k}, ..., i_{t-1})\}}$$

Note that the transition probabilities for $s = 1$ are calculated directly from equation (3.1). The previous model, called the non-proportional odds model (NPOM) in Bender & Grouven (1998), has $k(m-1)^2 + (m-1)$ parameters, far less than the corresponding saturated k-order Markov chain, which has $m^k(m-1)$ parameters. However, note that for $k = 1$ both the saturated Markov chain and the NPOM model have the same number of parameters to be estimated.

The Markov chain based on multinomial models is widely used in many practices and there is extensive literature on this issue. For example, Masson & Ruge-Murcia (2005) analyze the transition between exchange rate regimes using a Markov chain model with *time-varying* transition probabilities. In another field of application, Salazar et al. (2007) considers a multi-state Markov chain in a longitudinal study of transitions to dementia using a multinomial logistic model with a *subject-specific* random effects. Alternatively, a categorical log-linear model as in Avery & Henderson (1999) has also been used.

A more parsimonious model can be considered by imposing the proportional odds assumption, which establishes that the cumulative odds ratio for any two values of the covariates is constant across the response states (or categories) and, therefore, that the coefficients are identical across dependent variable categories, except for the constant term, i.e $\beta_{1,j}^{(i)} = \beta_{2,j}^{(i)} =, ..., \beta_{m-1,j}^{(i)} = \beta_j^{(i)}$ for all $j = 1, 2, ..., m-1$, leading to a lower number of parameters and thus to more parsimonious models. To understand this, consider a chain X_t characterized by the previous state values, say $X_{t-1} = j$ and $X_{t-1} = j'$. For the MC_1 model defined in (3.1) and under the proportional odds assumption, the cumulative odds take the form,

$$\frac{P(X_t \leq s|X_{t-1} = j)/P(X_t > s|X_{t-1} = j)}{P(X_t \leq s|X_{t-1} = j')/P(X_t > s|X_{t-1} = j')} = \exp\left(\beta_{s,j}^{(1)} - \beta_{s,j'}^{(1)}\right) = \exp\left(\beta_j^{(1)} - \beta_{j'}^{(1)}\right).$$

Thus the comparison of populations in terms of cumulative odds does not depend on the set of the response states s. In credit rating transitions, that means that if, for example, the cumulative odds for a previous rating $X_{t-1} = j$ is twice the cumulative odds for a previous rating $X_{t-1} = j'$, this holds for all the rating categories at X_t. In this context, the violation of this assumption can lead to model misspecification in the estimation of rating transitions probabilities which may affect the credit risk assessment. Hence, before estimating the proportional odds model (POM), it is important to verify whether the assumption of

proportionality is satisfied. As we shall see, this assumption is not satisfied in the municipal rating transitions. The only reference, to our knowledge, to the analysis of rating transitions in which this assumption is tested can be seen in Liu (2012), who examine the relationship between municipal bond insurance premia and the future credit migration of insured bonds. They also find that the proportional odds assumption is violated and estimate several binary regressions for each rating migration. Intermediate solutions are the Partial proportional odds models (PPOM), which allow the non-proportional odds property for a subset of variables as described in Bercedis & Harrell (1990). In Yee & Hastie (2003) a reduced-rank regression was proposed to reduce the number of parameters by replacing the matrix of regression coefficients in (3.3) with an approximation of lower rank r ($r < k$). A particular case of the reduced-rank regression was proposed by Anderson (1984), and also in subsequent papers like Kuss (2006), where it is called the *stereotype model*.

The Mixture Transition Distribution (MTD) models in Raftery (1985) and Berchtold & Raftery (2002) are higher-order Markov chain with fewer parameters and so they could be an excellent alternative. However, the meaning of the parameters for this model is not always clear and the efficiency for MTD estimations may be problematic on account of the large number of constraints on the parameters. Hence, there may be a need for more sophisticated optimization techniques (Lèbre & Bourguignon, 2008) but not even these algorithms always guarantee convergence. The Markov regression models in Zeger & Qaqish (1988) are another way to reduce the number of parameters.

3.3.3 A partial proportional odds model for persistent states

As noted previously, a multinomial cumulative model (NPOM) as defined in (3.3) can be used to estimate a k-order Markov chain model. However, even for an order $k = 1$ and $m = 6$ states, this model has 30 parameters to be estimated. In addition, in a process with *persistent* states, as shown in Table 3.3 for the ratings data, the off-diagonal probabilities are poorly estimated due to the low number of observations and could even occasionally lead us to convergence failures. In Stuart et al. (2013), the authors show that when the number of outcome categories is relatively large, the sample size is relatively

small or certain categories are rare, maximum likelihood can yield biased estimates of the regression parameters. This highlights the need for parsimonious models and a common procedure is to impose constraints on the parameters. As already noted in this regard, the POM model is highly restrictive and in many practices is an unrealistic assumption. We propose a partial proportional odds model that falls somewhere between the NPOM and the POM model, to fit a k-order Markov chain model for *persistent* states. This model, which will be denoted PPOM$_k(B^*)$, has the same specification as in (3.3) but the matrix $B^{(i)}$ is replaced with the matrix $B^{*(i)}$ defined as follows,

$$B^{*(i)} = \begin{bmatrix} \beta_{1,1}^{(i)} & \beta_2^{(i)} & \beta_3^{(i)} & \dots & \beta_{m-1}^{(i)} \\ \beta_1^{(i)} & \beta_{2,2}^{(i)} & \beta_3^{(i)} & \dots & \beta_{m-1}^{(i)} \\ & & \dots & & \\ \beta_1^{(i)} & \beta_2^{(i)} & \beta_3^{(i)} & \dots & \beta_{m-1,m-1}^{(i)} \end{bmatrix} \tag{3.4}$$

$B^{*(i)}$ is the *persistent-diagonal* matrix in which the diagonal parameters are not constrained, unlike in the POM model. Note that, in *persistent* processes the diagonal parameters may be reliably estimated since the transitions between states are diagonally dominant. It is plain to see that the number of parameters to estimate a k-order Markov chain using the PPOM$_k(B^*)$ model is $m - 1 + 2(m - 1)k$. For instance, for an order $k = 1$ and $m = 6$ states, this model has 15 parameters to be estimated. Unfortunately, the parameters increase linearly as k increases and, for example, for $k = 2$ the number of parameters is 25 and for $k = 3$ there are 35. In the following section, we propose a model for *long-persistence* states to estimate a k-order Markov chain. Before, the k-order Markov chain PPOM$_k(B^*)$ model for $k = 1, 2$ will be applied to estimate the transition probabilities for the ratings data. The results will be compared with the full-parameterized MC$_1$ and MC$_2$ models.

3.3.3.1 Estimation of the PPOM$_k(B^*)$ model

We fit a k-order Markov chain using the PPOM$_k(B^*)$ model for $k = 1, 2$ described previously, in order to estimate the rating transition probabilities in a 1-year time horizon. Before, we estimate the 1-order Markov chain as defined in (3.3) under the proportional

odds assumption (POM), obtaining a BIC equal to 36102. The total number of estimated parameters was 10. The comparison of this result with the BIC obtained for the full-parameterized MC_1 model in Table 3.4 leads to a worse goodness of fit for the POM model. In addition, the test of the proportional odds assumption (Rao et al., 1998) is significant (p-value < 0.001), indicating that the proportional odds assumption does not hold, and suggesting that non-equal parameters are needed across the *logit* equations in (3.3).

BIC for the fitted 1-order Markov chain $PPOM_1(B^*)$ model was equal to 35551, which is better than both the 1-order Markov chain using the POM model and the full-parameterized MC_1. Several F-statistics were carried out to test the linear null hypothesis $H_0 : \beta_{s,s}^{(1)} - \beta_s^{(1)} = 0$ in the matrix defined in (3.4), leading us to reject them. Only for the coefficients $\beta_3^{(1)}$ and $\beta_4^{(1)}$, were the null hypothesis not rejected and the proportional odds assumption maintained. Table 3.5 presents the estimated coefficients (b_{ij}), their standard errors and the t-student statistic. Note that the number of parameters to be estimated for the proposed model was 13 in contrast to the full-parameterized MC_1, where the number was 30. A *subject-specific* random effect for each debt issuer was tested but did not prove to be statistically significant, unlike the independence model MC_0^* in Table 3.4, which was found to be significant. Table 3.6 shows the estimated transition probabilities and their standard errors, for the 1-order Markov chain using the $PPOM_1(B^*)$ model. These results can be compared with those obtained previously for the full-parameterized MC_1 model in Table 3.3. Note that the greater the number of observations in each cell, the closer the two estimations are. Note also how the $PPOM_1(B^*)$ model smooths the estimations producing zero estimates for cells (transitions) having very few observations.

Similarly, we fit a 2-order Markov chain using the $PPOM_2(B^*)$ model and the BIC obtained was 33817, which is better than those obtained for the 1-order Markov chain models and the full-parameterized MC_2 model. Table 3.7 presents the estimated coefficients and their standard errors for the $PPOM_2(B^*)$ model. The proportional odds assumption was imposed on some lagged variables ($b_{ij}^{(2)}$) due to convergence failures in the estimation because of the high number of parameters to estimate, and the reduced number of observations for low credit ratings. Table 3.8 shows the estimated transition probabilities for some rating migrations using the $PPOM_2(B^*)$ and MC_2 models. Note that the

Table 3.5: Estimated parameters for a 1-order Markov chain using the PPOM$_1(B^*)$ model.

Parameter	Estimate	Standard Error	t-statistic	p-value
b_{10}	-14.88	0.35	-42.30	<.0001
b_{20}	-13.07	0.17	-77.63	<.0001
b_{30}	-7.92	0.09	-88.16	<.0001
b_{40}	-4.18	0.07	-63.02	<.0001
b_{50}	-1.24	0.02	-62.16	<.0001
$b_1^{(1)}$	14.46	0.52	27.61	<.0001
$b_{11}^{(1)}$	15.91	0.56	28.17	<.0001
$b_2^{(1)}$	12.42	0.47	26.35	<.0001
$b_{22}^{(1)}$	16.39	0.53	30.95	<.0001
$b_3^{(1)}$	8.68	0.10	89.98	<.0001
$b_4^{(1)}$	6.93	0.09	77.35	<.0001
$b_5^{(1)}$	2.95	0.07	40.25	<.0001
$b_{55}^{(1)}$	5.96	0.14	41.98	<.0001

Table 3.6: 1-year estimated probabilities for a 1-order Markov chain using the PPOM$_1(B^*)$ model. Approximate standard errors for the estimates using the delta method (in brackets).

	X_t					
X_{t-1}	1	2	3	4	5	6
1	0.7335	0.0667	0.1984	0.0014	0.0000	0.0000
	(0.0856)	(0.0434)	(0.0786)	(0.0007)	(0.0000)	(0.0000)
2	0.0848	0.8724	0.0331	0.0094	0.0002	0.0000
	(0.0263)	(0.0267)	(0.0152)	(0.0049)	(0.0001)	(0.0000)
3	0.0019	0.0103	0.6703	0.3065	0.0104	0.0006
	(0.0007)	(0.0016)	(0.0080)	(0.0076)	(0.0007)	(0.0001)
4	0.0003	0.0019	0.2693	0.6683	0.0570	0.0032
	(0.0001)	(0.0003)	(0.0071)	(0.0070)	(0.0032)	(0.0003)
5	0.0000	0.0000	0.0069	0.2196	0.7646	0.0089
	(0.0000)	(0.0000)	(0.0005)	(0.0053)	(0.0056)	(0.0012)
6	0.0000	0.0000	0.0004	0.0147	0.2099	0.7750
	(0.0000)	(0.0000)	(0.0000)	(0.0010)	(0.0034)	(0.0035)

2-order Markov chain model allow us to take into account rating *momentum* effects in the rating process: for example, it is interesting to compare the estimated probabilities for the sequences of ratings $(X_{t-2}, X_{t-1}, X_t) = (6, 5, 4)$ and $(X_{t-2}, X_{t-1}, X_t) = (5, 5, 4)$, respectively 0.2722 and 0.1421. The estimated probability of migrating from rating 5 to 4 at time t when the previous rating was downgraded is larger than that obtained when the rating has remained unchanged in the last two previous states.

Table 3.7: Estimated parameters for a 2-order Markov chain using the PPOM$_2(B^*)$ model.

Parameter	Estimate	Standard Error	t-statistic	p-value
b_{10}	-16.34	0.37	-43.53	<.0001
b_{20}	-14.49	0.19	-76.16	<.0001
b_{30}	-8.59	0.10	-85.57	<.0001
b_{40}	-4.19	0.07	-62.96	<.0001
b_{50}	-1.23	0.02	-61.90	<.0001
$b_1^{(1)}$	19.31	0.82	23.62	<.0001
$b_{11}^{(1)}$	20.76	0.84	24.65	<.0001
$b_2^{(1)}$	17.26	0.67	25.88	<.0001
$b_{22}^{(1)}$	21.51	0.62	34.66	<.0001
$b_3^{(1)}$	13.80	0.18	76.25	<.0001
$b_4^{(1)}$	9.44	0.12	76.66	<.0001
$b_5^{(1)}$	3.23	0.07	42.85	<.0001
$b_{55}^{(1)}$	7.04	0.22	31.35	<.0001
$b_1^{(2)}$	-3.44	0.90	-3.82	<.0001
$b_2^{(2)}$	-3.28	0.59	-5.50	<.0001
$b_3^{(2)}$	-4.91	0.13	-36.57	<.0001
$b_4^{(2)}$	-3.05	0.10	-31.62	<.0001
$b_5^{(2)}$	-0.82	0.06	-12.74	<.0001
$b_{55}^{(2)}$	-1.71	0.23	-7.23	<.0001

3.4

Long-persistence multi-state models

As previously discussed, in *persistent* states, transitions between distant states are unlikely. Therefore, even though the use of some of the previous states to estimate the transition probabilities seems to be a realistic assumption, the need to use the detailed history is less probable. We introduce models that include the immediately previous observations (*short-persistence*) and also a summary of what happened many times before a certain instant (*long-persistence*).

Given a *persistent* chain, we define

$$E_r^k(t) := E_r^k(t; X_{t-r}, ..., X_{t-k}) = \#\{X_{t-j} = X_{t-r}; j = r+1, ..., k\}$$

$$U_r^k(t) := U_r^k(t; X_{t-r}, ..., X_{t-k}) = \#\{X_{t-j} < X_{t-r}; j = r+1, ..., k\}$$

where k and r are integers, $k > r$. $E_r^k(t)$ counts how many realizations were equal to that of the state at a time equal to $t-r$, between the instants $t-r-1$ and $t-k$. $U_r^k(t)$ represents the number of realizations that was below (in the considered order) than that of the state at time $t-r$, between $t-r-1$ and $t-k$. Note that the number of realizations in the interval $t-r-1$ and $t-k$ greater than that of the state at $t-r$, denoted by $D_r^k(t)$, is given by $D_r^k(t) = k-r-E_r^k(t)-U_r^k(t)$. As illustration, consider the following sequences of states: $(X_{t-3}, X_{t-2}, X_{t-1}, X_t) = (6, 5, 5, 4)$ and $(X_{t-3}, X_{t-2}, X_{t-1}, X_t) = (5, 5, 5, 4)$. For the first sequence, $E_1^3(t) = 1$, $U_1^3(t) = 0$ and $D_1^3(t) = 1$. For the second one, $E_1^3(t) = 2$, $U_1^3(t) = 0$ and $D_1^3(t) = 0$. Note that, in this last case, these variables together with the information about the state at time $t-1$, provide full information on the exact sequence of states.

We define the k-order *Long-persistence* Markov chain model of degree r, which will be denoted LPM$_k(r)$, by means of the relationship:

$$P(X_t = i_t | X_0 = i_0, X_1 = i_1, ..., X_{t-1} = i_{t-1}) =$$
$$P(X_t = i_t | E_r^k(t) = n_e, U_r^k(t) = n_u, D_r^k(t) = n_d, X_{t-r} = i_{t-r}, ..., X_{t-1} = i_{t-1})$$

where n_e,n_u,$n_d \in \{0, 1, ..., k-r\}$ such that $n_e + n_u + n_d = k-r$. In the conditional part, there are the *long-persistence* terms, $E_r^k(t)$,$U_r^k(t)$,$D_r^k(t)$ and the *short-persistence* terms represented by the short memory observations $X_{t-r} = i_{t-r}$,...,$X_{t-1} = i_{t-1}$.

Table 3.8: 1-year estimated probabilities for a 2-order Markov chain using the PPOM$_2$(B^*) model and their approximate standard errors using the delta method (in brackets). The MC$_2$ estimator is also shown (in square brackets).

		X_t					
X_{t-2}	X_{t-1}	1	2	3	4	5	6
...	...			...	...		
2	2	0.0860	0.8908	0.0187	0.0045	0.0001	0.0000
		(0.0329)	(0.0308)	(0.0127)	(0.0024)	(0.0000)	(0.0000)
		[0.1111]	[0.8095]	[0.0476]	[0.0159]	[0.0159]	[0.0000]
...	...			...	...		
3	3	0.0006	0.0031	0.5700	0.4173	0.0085	0.0005
		(0.0002)	(0.0005)	(0.0101)	(0.0099)	(0.0007)	(0.0000)
		[0.0004]	[0.0084]	[0.5467]	[0.4400]	[0.0035]	[0.0009]
...	...			...	...		
4	4	0.0000	0.0003	0.0988	0.8010	0.0942	0.0057
		(0.0000)	(0.0000)	(0.0057)	(0.0067)	(0.0053)	(0.0005)
		[0.0000]	[0.0029]	[0.0492]	[0.8653]	[0.0782]	[0.0043]
4	5	0.0000	0.0000	0.0002	0.0175	0.9227	0.0595
		(0.0000)	(0.0000)	(0.0000)	(0.0016)	(0.0127)	(0.0131)
		[0.0000]	[0.0000]	[0.0000]	[0.0370]	[0.9444]	[0.0185]
...	...			...	...		
5	4	0.0004	0.0024	0.5034	0.4820	0.0103	0.0015
		(0.0001)	(0.0004)	(0.0127)	(0.0123)	(0.0009)	(0.0003)
		[0.0000]	[0.0000]	[0.6594]	[0.3091]	[0.0293]	[0.0023]
5	5	0.0000	0.0000	0.0020	0.1421	0.8395	0.0163
		(0.0000)	(0.0000)	(0.0002)	(0.0069)	(0.0073)	(0.0024)
		[0.0000]	[0.0000]	[0.0177]	[0.0436]	[0.9181]	[0.0206]
...	...			...	...		
6	5	0.0000	0.0000	0.0046	0.2722	0.7201	0.0030
		(0.0000)	(0.0000)	(0.0003)	(0.0070)	(0.0071)	(0.0006)
		[0.0000]	[0.0000]	[0.0000]	[0.3242]	[0.6739]	[0.0019]
6	6	0.0000	0.0000	0.0002	0.0147	0.2109	0.7742
		(0.0000)	(0.0000)	(0.0000)	(0.0009)	(0.0034)	(0.0035)
		[0.0001]	[0.0002]	[0.0010]	[0.0137]	[0.2103]	[0.7747]

In order to estimate the transition probabilities, we generalize the model described in

(3.3) to the LPM$_k(r)$. Thus, we assume the following multinomial cumulative model:

$$P(X_t \leq s | E_r^k(t) = n_e, U_r^k(t) = n_u, X_{t-r} = i_{t-r}, ..., X_{t-1} = i_{t-1}) = \frac{\exp\{\psi(s; n_e, n_u, i_{t-r}, ..., i_{t-1})\}}{1 + \exp\{\psi(s; n_e, n_u, i_{t-r}, ..., i_{t-1})\}}$$

where the linear predictor is similarly defined as in (3.2),

$$\psi(s) := \psi(s; E_r^k(t), U_r^k(t), X_{t-r}, ..., X_{t-1}) = \beta_{s,0} + \sum_{i=1}^{r} \sum_{j=1}^{m-1} \beta_{s,j}^{(i)} I_{\{X_{t-i}=j\}} + \gamma_s^e E_r^k(t) + \gamma_s^u U_r^k(t)$$

where γ_s^e and γ_s^u are the coefficients for the *long-persistent* variables. Once again, $m-1$ *logit* equations are defined for $s = 1, 2, ..., m-1$. We note the above linear system in matrix form as in (3.3). However, we can replace the matrix $B^{(i)}$ by the *persistent-diagonal* matrix $B^{*(i)}$ defined in (3.4) in order to reduce the number of parameters and we have,

$$\Psi = \eta_0 + B^{*(1)} \cdot Y_1 + ... + B^{*(r)} \cdot Y_r + \Gamma \cdot L_r^k \tag{3.5}$$

where L_r^k is the vector of the *long-persistent* variables $L_r^k = [E_r^k(t), U_r^k(t)]^T$, and Γ is the matrix containing their corresponding coefficients, that is $\Gamma = [\gamma^e, \gamma^u]$ where $\gamma^e = (\gamma_1^e, ..., \gamma_{m-1}^e)^T$ and $\gamma^u = (\gamma_1^u, ..., \gamma_{m-1}^u)^T$. It is plain to see that the number of parameters for the LPM$_k(r)$ model, considering m states, is $(m-1)(3+2r)$, far lower than the MC$_k$ model, which is $m^k(m-1)$. Note that the number of parameters in the LPM$_k(r)$ model does not depend on k, the order of the *long-persistence* term. For instance, the number of parameters for the LPM$_3(1)$ model is equal to 25 in contrast to the MC$_3$ model, which is 1080. Note that the model defined in (3.5) can be directly extended to account for measurable covariates that may have an effect on the state probabilities.

3.5 Results

3.5.1 Estimation of the LPM$_k(r)$ models

We have fitted the LPM$_k(r)$ models for $k = 2, 3, ..., 6$ and $r = 1, 2$ to the ratings data by previously removing the first 6 observations of each issuer, in order to obtain a good

comparison between them. We have selected the best model according to the BIC statistic. The results are summarized in Table 3.9. According to the BIC criterion, the best model is the LPM$_3(1)$, followed by the LPM$_3(2)$. This means that the state of the rating at instant t depends on the 3 previous states. Additionally, for all the other LPM$_k(2)$ models shown in Table 3.9, the proportional odds assumption was imposed on the lagged indicator variables due to convergence failures in the estimation because of the high number of parameters to estimate. Again, the off-diagonal coefficients are poorly estimated because these correspond to cells with few observations. Hence, models with a large order r will be less reliable and more likely to suffer from convergence problems. Therefore, the chosen model is the LPM$_3(1)$, which has the best goodness of fit statistic, being in some sense, the most parsimonious.

Table 3.9: BIC statistics for several LPM$_k(r)$ models. $\hat{l}$ is the estimated log-likelihood. Sample size $n = 15208$. The first 6 observations for each subject are removed from the sample.

Model	$-2\hat{l}$	BIC
LPM$_6(1)$	17990	18134
LPM$_5(1)$	17678	17823
LPM$_4(1)$	17269	17414
LPM$_3(1)$	16481	16644
LPM$_2(1)$	17510	17654
LPM$_6(2)$	17643	17826
LPM$_5(2)$	17478	17661
LPM$_4(2)$	17182	17365
LPM$_3(2)$	16529	16712

Table 3.10 presents the estimated coefficients, their standard errors and the t-student statistics of the LPM$_3(1)$ model. As in previous sections, the first 3 observations for each issuer were removed from the sample. Note that all the parameters are statistically significant (p-value $< .0001$). Hypothesis tests on the parameters were carried out to validate the odds proportional assumption. First, the F-statistics to test the linear null

Table 3.10: Estimated parameters for the LPM$_3(1)$ model.

Parameter	Estimate	Standard Error	t-statistic	p-value
b_{10}	-14.14	0.37	-37.75	<.0001
b_{20}	-12.24	0.18	-67.99	<.0001
b_{30}	-5.85	0.10	-58.06	<.0001
b_{40}	-2.07	0.11	-18.89	<.0001
b_{50}	0.88	0.09	9.78	<.0001
b_1	15.50	0.61	25.35	<.0001
b_{11}	17.01	0.65	26.10	<.0001
b_2	12.95	0.48	26.72	<.0001
b_{22}	18.02	0.51	35.02	<.0001
b_3	8.93	0.12	76.59	<.0001
b_4	6.09	0.09	65.27	<.0001
b_5	1.28	0.11	12.09	<.0001
b_{55}	4.96	0.14	34.20	<.0001
γ^e_{13}	-1.84	0.05	-39.72	<.0001
γ^e_{45}	-1.06	0.04	-24.16	<.0001
γ^u_{13}	-1.65	0.26	-6.28	<.0001
γ^u_{45}	-0.98	0.10	-9.56	<.0001

hypotheses $H_0 : \beta^{(1)}_{s,s} - \beta^{(1)}_s = 0$ are calculated as in the estimated 1-order Markov chain under the PPOM$_1(B^*)$ model described in an earlier section. The results of the test lead us to reject the null hypotheses with the exception of that involving the coefficients β_3 and β_4. In this case, the proportional odds assumption is maintained. With regard to the parameters γ^e_s for $s = 1, ..., 5$ the results of the tests lead us to reject the odds proportional assumption. In order to improve the model, we have relaxed this assumption by estimating two parameters: one for low ratings ($s \leq 3$) denoted γ^e_{13} and other for high ratings ($s > 3$) denoted γ^e_{45}. We use this cut boundary because it corresponds to the division between *investment* grade and *speculative* grade and can be economically interpretable. The linear null hypothesis $H_0 : \gamma^e_{13} - \gamma^e_{45} = 0$ is rejected (p-value $< .0001$). The same procedure is carried out for the parameters γ^u_s with identical results. In this case, the null hypothesis $H_0 : \gamma^u_{13} - \gamma^u_{45} = 0$ is rejected with a p-value equal to 0.0168.

The BIC obtained for the LPM$_3(1)$ model is 32532, which is lower than for the previously estimated models. Comparing this value with the BIC obtained for the full-parameterized MC$_3$ model in Table 3.4, we note that this is around 22% lower. Additionally, the number of parameters is far lower for the LPM$_3(1)$ model, which has 17 estimated parameters compared to 1080 parameters in the MC$_3$ model. Compared to both the PPOM$_1(B^*)$ model and to the PPOM$_2(B^*)$ model, the BIC obtained by both models, 35551 and 33817 respectively, was worse than the one obtained by the LPM$_3(1)$ model.

Table 3.11 shows the 1-year estimated transition probabilities obtained from the LPM$_3(1)$ model and their corresponding standard errors (in brackets), conditioned to some fixed values of the *short* and *long persistence* variables. It is interesting to note that the estimated probabilities to be downgraded are larger for ratings with previous downgrades. For instance, the estimated transition probability of migrating in one year from rating 4 to rating 3, when the rating was previously downgraded ($D_1^3(t) = 2$) is 0.5595 and the observed probability (sample proportion) is 0.5698. In contrast, the estimated probability for a transition from rating 4 to rating 3 when the rating has remained unchanged in the past ($E_1^3(t) = 2$) is far smaller, i.e 0.0314 (the observed probability is 0.0401). This result is consistent with many findings that provide empirical evidence that ratings with previous downgrades are more prone to further downgrades than others (*rating momentum*). For the border rating 3 that, as mentioned, represents the boundary between the *investment* grade and the *speculative* grade, the behavior is somewhat different: when the rating 3 has remained unchanged in the last 3 previous states, then the probability of being upgraded increases. In this case, the estimated transition probability of migrating from rating 3 to 4 is 0.6350 and the observed probability is 0.7313. However, the estimated probability of a rating 3 being downgraded to 2 in one year, when it has previously been downgraded, is small (0.0298) and it is most likely to remain as rating 3 (0.9210).

As described in section 3.2, the municipal rating assigned by the agency results from a weighted average of several factors, among which economy score receives the highest weight. Consequently, a deterioration in economy over time should be expected to have an impact on the rating system in terms of downgrades. In our model, these downgrades would directly affect the predicted ratings in a 3-year horizon, by making further downgrades more likely. Since there may be a gap between the economic shock and the

Table 3.11: 1-year probabilities estimated by the $\text{LPM}_3(1)$ model conditioned to some fixed values of the variables X_{t-1}, $E_1^3(t)$, $U_1^3(t)$ and $D_1^3(t)$ which indicate the number of times that the realizations of the ratings in the interval $t-2$ and $t-3$ were higher than those of the state at $t-1$ and calculated as $D_1^3(t) = 2 - E_1^3(t) - U_1^3(t)$. Approximate standard errors for the estimates using the delta method (in brackets).

			X_t					
$E_1^3(t)$	$U_1^3(t)$	X_{t-1}	1	2	3	4	5	6
0	0	1	0.9461	0.0168	0.0370	0.0001	0.0000	0.0000
			(0.0269)	(0.0131)	(0.0207)	(0.0000)	(0.0000)	(0.0000)
0	0	2	0.2321	0.7648	0.0023	0.0008	0.0000	0.0000
			(0.0648)	(0.0646)	(0.0014)	(0.0004)	(0.0000)	(0.0000)
0	0	3	0.0055	0.0298	0.9210	0.0427	0.0010	0.0000
			(0.0019)	(0.0046)	(0.0056)	(0.0033)	(0.0001)	(0.0000)
0	0	4	0.0003	0.0018	0.5595	0.4208	0.0166	0.0010
			(0.0001)	(0.0003)	(0.0112)	(0.0101)	(0.0014)	(0.0001)
0	0	5	0.0000	0.0000	0.0103	0.3023	0.6845	0.0029
			(0.0000)	(0.0000)	(0.0009)	(0.0073)	(0.0074)	(0.0004)
	...				...	...		
0	2	4	0.0000	0.0000	0.0449	0.8416	0.1068	0.0067
			(0.0000)	(0.0000)	(0.0226)	(0.0295)	(0.0186)	(0.0014)
	...				...	...		
2	0	3	0.0001	0.0008	0.3555	0.6350	0.0082	0.0005
			(0.0001)	(0.0001)	(0.0117)	(0.0117)	(0.0008)	(0.0001)
2	0	4	0.0000	0.0000	0.0314	0.8396	0.1213	0.0077
			(0.0000)	(0.0000)	(0.0026)	(0.0076)	(0.0070)	(0.0007)
2	0	5	0.0000	0.0000	0.0003	0.0518	0.9243	0.0236
			(0.0000)	(0.0000)	(0.0000)	(0.0041)	(0.0048)	(0.0033)
2	0	6	0.0000	0.0000	0.0001	0.0149	0.2097	0.7753
			(0.0000)	(0.0000)	(0.0000)	(0.0010)	(0.0034)	(0.0035)
	...				...	...		

updating of the rating system, in many studies, especially those focused on corporate ratings, macroeconomic factors have been analyzed in order to predict rating migrations. McNeil & Wendin (2006) analyzes, for example, the validity of the Chicago Fed National

Activity Index (CFNAI) as an explanatory variable but, as noted by the authors, the results were slightly contradictory. In this same way, additional covariates could be directly included in the specified model in (3.5). This would imply the assumption that all obligors are exposed to the same systematic risk in a specific period of time. However, in local governments this assumption is questionable, and recent and previously cited studies have suggested that local conditions affecting an area or municipality could prove to be more explanatory. As also suggested in Standard and Poor's (2013), an overall economic shock can threaten the paying ability of a company more directly than a local government, because the company may find it difficult to raise prices or reduce costs. In our opinion, more in-depth analysis of other factors affecting municipal rating migrations is required, but this is beyond the scope of this study.

3.5.2 Forecasting the marginal rating probabilities

To assess the performance of the $\text{LPM}_3(1)$ model, we use the sample of municipal bond issuers rated by Standard and Poor's in 2014. Therefore, these ratings have not been used in the estimation model and will only be used for the purpose of evaluating the predictive capability of the model. As commented earlier, when a financial entity holds a portfolio of rated assets, it is a crucial to analyze the deterioration in the quality of that portfolio. For instance, if an entity, in a 1-year time horizon, predicts an evolution of its portfolio towards lower ratings, this will have an immediate impact on its expected loss and capital at risk. Therefore, the prospects for changes in the rating distribution will directly affect this entity's results. Since a portfolio of N assets in the current year t of which, for example, n_{BB} are rated BB, we can directly work out the probability of selecting an asset rated BB, i.e n_{BB}/N. A key point, in practice, is to determine the probability of choosing, in the next year $t+1$, an asset that is, for example, rated BB from this portfolio. This will depend on the estimated probabilities of rating transition for 1-year horizon.

The $\text{LPM}_k(r)$ model can be used to predict the marginal state probabilities 1-period ahead as follows:

$$P(X_{t+1}=s)=\sum_{n_e=0}^{k-r}\sum_{n_u=0}^{k-r}\sum_{i_{t-r+1}=1}^{m}...\sum_{i_t=1}^{m}P(X_{t+1}=s|\xi)P(\xi)$$

where $\xi = \{E_r^k(t+1) = n_e, U_r^k(t+1) = n_u, X_{t-r+1} = i_{t-r+1}, ..., X_t = i_t\}$ is the history of states or a summary of the same, between instants t and $t-k+1$. Note that the *long-persistence* terms, $E_r^k(t+1)$ and $U_r^k(t+1)$, only depend on the known sequence of states between t and $t-k+1$. $P(X_{t+1} = s|\xi)$ are the estimated transition probabilities resulting from the LPM$_k(r)$ model and $P(\xi)$ the estimated probabilities for each possible occurrence of ξ based on the sample proportions. The procedure is exactly the same as for the MC$_k$ model but instead of the exact sequence of states $\{X_{t-k}, ..., X_t\}$, we have a summary of states contained in the *long-persistence* variables. Table 3.12 shows the predicted marginal rating probabilities for the year 2014, when knowing the rating probabilities for 2013. Once again, the results are also shown for the estimated 1-order Markov chain under the PPOM$_1(B^*)$ model and the full-parameterized MC$_1$. According to the *chi-square* statistic, which is reported in a merely descriptive manner, the best prediction results were obtained by the 3-order Markov chain LPM$_3(1)$ model.

Table 3.12: *Out-of-sample* observed (p_i) and predicted (p_i^e) marginal rating probabilities in year 2014 for the 3-order Markov chain LPM$_3(1)$ model and the estimated 1-order Markov chain models: PPOM$_1(B^*)$ and MC$_1$. Observed (N_i) and expected (N_i^e) counts are also presented. *Chi-square* statistic (χ^2) for each model is shown below.

Rating i	Observed		LPM$_3(1)$		PPOM$_1(B^*)$		MC$_1$	
	N_i	p_i	N_i^e	p_i^e	N_i^e	p_i^e	N_i^e	p_i^e
1	2	0.0004	3	0.0005	3	0.0006	4	0.0007
2	22	0.0041	16	0.0031	20	0.0037	19	0.0036
3	148	0.0277	105	0.0197	402	0.0754	402	0.0754
4	961	0.1802	1136	0.2130	1540	0.2888	1539	0.2885
5	3720	0.6974	3617	0.6781	2970	0.5568	2971	0.5570
6	481	0.0902	456	0.0856	399	0.0748	399	0.0748
χ^2		-		**50.95**		**585.48**		**584.52**

Notes: *Chi-square* statistic is calculated as $\chi^2 = \sum_i (N_i - N_i^e)^2 / N_i^e$.

3.6

Economic relevance: credit portfolio risk

Understanding the relative risk of different types of credit exposure is an important objective for regulators designing regulatory capital requirements, and for banks themselves engaged in the allocation of economic capital. An excellent discussion about the use of credit risk models as a basis for banks' calculations and their regulatory implications can be seen in Jackson & Perraudin (2000) and Crook & Bellotti (2010). A measure of relative risk of credit exposures is the Value-at-Risk (VaR) statistic. For a given portfolio, the VaR with confidence level α is the estimated quantile of the credit loss distribution at a probability level α, corresponding to the loss that will be exceeded on a fraction α of occasions by an investor holding this portfolio. Because the rating is assumed to reflect credit risk, a change in credit rating has direct performance implications. Credit risk may be defined as the risk of losses due to credit events, i.e default or a change in the quality of the rating (rating transition). For high quality portfolios as in the case of municipal bonds, the credit risk in default is very low and the risk lies particularly in the possibility to migrate towards lower ratings (European Central Bank, 2007). Estimates of rating transitions are used to quantify the changes in value of portfolio of loans or bonds at a given time horizon, usually 1 year.

There is extensive literature on the economic relevance of accounting for the detailed dynamics of the bond rating process. Altman & Kao (1992) and Altman (1998) reveal a tendency for a downgrade in rating to be followed by a second downgrade (downward *momentum*) and argue that investors may be able to use this information to enhance their bond portfolios' expected returns. Bangia et al. (2002) shows the importance of accounting for business cycle effects when assessing the credit risk of portfolio. By separating the economy into two states or regimes, expansion and contraction, and conditioning the rating transitions on these states, they show that the loss distribution of credit portfolios can differ greatly. Specifically, they obtain that the maximum loss in value at a 99.9% confidence level during a economic contraction is more than two times that in expansion. In an interesting study, Fuertes et al. (2012) consider a Mixture of Markov Chain models to account for stochastic business cycle effects in credit rating migration risk and show their impact on the capital requirements. At the level 99.9%, they obtain a risk capital,

in a context of economic slowdown, a 65% higher than that suggested by the classical MC_1 model. Guttler & Raupach (2008) analyze the impact of the downward *momentum* effect on credit portfolio risk. In their model, the rating transition probabilities are conditioned on previous downgrades and they find that investors who account for the downward *momentum* perceive higher VaR compared with investors who do not.

The valuation of a bond is derived from the one-year forward zero-curve corresponding to the rating of the issuer. A detailed overview can be found in Gupton et al. (1997) and in Crouhy et al. (2000). Spread data for several maturities may be obtained from sources such as CreditMetrics or Bloomberg. Let us assume, as a simple example, the bond yield, for each rating, shown in Table 3.13. These values are arbitrary, used only for illustrative purposes.

Table 3.13: Example one-year forward zero curve (year 4) by credit rating category (%).

Rating	1	2	3	4	5	6
%	15.00	8.00	7.00	5.32	5.17	4.50

Note that the riskier the credit the higher the yield, that is, compensation (spread) for the possibility of not receiving future coupons or the principal. It is well known and commonly used, for example in Kiesel et al. (2001), that the 1-year forward price of a zero-coupon bond rated s at time t and with maturity in T years, denoted $V_s^t(r^s(T), T)$, may be written as

$$V_s^t(r^s(T), T) = C \exp\{-r^s(T)T\},$$

where C is the bond's value at maturity (face amount) and $r^s(T)$ is the T-maturity bond yield rated s. As an illustration, we calculate credit-VaR at time $t+1$ for a zero-coupon bond maturing exactly in 5 years with a face amount of \$926,923 and showing the historical sequence of ratings $(X_{t-2}, X_{t-1}, X_t) = (5, 5, 4)$. The 1-year forward price of the bond, if the obligor stays unchanged in rating 4, is equal to $V_4^{t+1}(0.0532, 4) = \$749,248$. The value of the bond yield rated as $s = 4$, 0.0532, is obtained from Table 3.13. Note that if the bond is downgraded, the bond yield will rise, and thus the 1-year forward price will fall. In this example, if the bond is downgraded from rating 4 to 3 at time $t+1$ then the 1-year forward price is equal to $V_3^{t+1}(0.07, 4) = \$700,553$ and so, the change in the bond

value (ΔV) is $\Delta V = \$700,553 - \$749,248 = -\$48,695$. Conversely, if the bond gets upgraded from rating 4 to 5 the price is $V_5^{t+1}(0.0517, 4) = \$753,757$ and the change in the bond value is $\Delta V = \$753,757 - \$749,248 = \$4,509$. If we replicate these calculations for each rating category we obtain the values shown in Table 3.14. We can use the estimated transition probabilities resulting from the LPM$_3(1)$ model or from any of the models previously estimated to derive the forward distribution of the changes in the bond value, at the 1-year horizon, due to an eventual change in credit quality. For example, the estimated probability of having a loss equal to $-\$48,695$, at time $t+1$, according to the LPM$_3(1)$ model, is 0.5595 which corresponds to the probability of migrating from rating 4 to rating 3 when the rating was previously downgraded (see also Table 3.11). The percentile of the empirical distribution of ΔV (last column in Table 3.14), corresponding to credit-VaR at the 99.9% confidence level, is $-\$76,164$. Table 3.14 shows the probabilities of the states coming from both the MC$_1$ and LPM$_3(1)$ models. Note that, in this illustrative example, the LPM$_3(1)$ model, which takes into account downgrade *momentum* effects, gives a higher probability of loss ($\Delta V < 0$) than the MC$_1$ model.

Table 3.14: Distribution, at time $t+1$, of the changes in the value of a bond, with historical sequence of ratings $(X_{t-2}, X_{t-1}, X_t) = (5, 5, 4)$, resulting both from the MC$_1$ and LPM$_3(1)$ model.

Year-end rating	Probability of state		Forward price: V ($)	Change in value: ΔV ($)
	MC$_1$	LPM$_3(1)$		
1	0.0003	0.0003	508,706	-240,542
2	0.0021	0.0018	673,084	-76,164
3	0.2709	0.5595	700,553	-48,695
4	0.6638	0.4208	749,248	0
5	0.0594	0.0166	753,757	4,509
6	0.0035	0.0010	774,231	24,982

A similar procedure is applied to calculate credit-VaR at time $t+1$ for a portfolio which consists of more than one bond, but with the additional complexity that asset returns and therefore rating migrations may be correlated. We simulate a credit portfolio of 100 zero-coupon bonds maturing in 5 years, assign to each a random historical sequence of corre-

lated ratings (X_{t-2}, X_{t-1}, X_t) randomly drawn from the state space $S = \{2, 3, ..., 6\}$. The face amount for each bond is randomly drawn from an uniform distribution with range \$1 to \$1m summing up a total of \$50,286,406 for the whole portfolio. We assume a scenario of economic downturn for constructing this portfolio, simulating a negative shock at time t, which leads to an impact in terms of rating downgrades. The objective is to quantify the impact of the downward rating *momentum* on portfolio risk for one-year horizon. Similar to CreditMetrics (Gupton et al., 1997), we use the Vasicek's model (Vasicek, 2002), which generates rating transitions on a portfolio of credits in a simple version of the single asset model of Merton (Merton, 1974). The Vasicek's model is written as,

$$A_i = \sqrt{\rho}Z + \sqrt{1-\rho}\xi_i$$

where A_i is an asset return of firm i, Z is a common factor which affects all firms equally, ξ_i is the idiosyncratic credit risk of firm i, and ρ is the asset correlation between any two firms in the portfolio. Random variables Z and ξ_i are assumed to be i.i.d. standard normal. Here, we use a value for the asset correlation of 0.10, as found in Dullmann et al. (2007). According to the Vasicek's model, the firm's asset return determines its evolution of credit ratings: a very sharp decrease in the company's asset value makes the company default; a sufficient decrease leads to a downgrade, and a sufficient increase to an upgrade. Otherwise, the company remains in the current rating class. The simulation procedure used here, described in detail in Bangia et al. (2002) and Guttler & Raupach (2008), is as follows: each simulation step j consists of drawing an asset return A_i from Vasicek's model for each obligor i, followed by assigning a new rating at time $t+1$ according to the transition probabilities resulting from the models previously estimated, calculating individual change in the bond value $\Delta V_i^{(j)}$ and aggregating them to a single random of the portfolio's present value as $L^{(j)} = \sum_i \Delta V_i^{(j)}$, which is the decrease or increase in the value of portfolio due to rating transitions. By simply repeating this procedure a sufficiently large number of iterations, it yields to a Monte Carlo approximation of the loss distribution and its quantiles. Table 3.15 reports the VaR estimates at confidence levels of 99.0% and 99.9%, although we focus on the 99.9% VaR because it is standard under Basel III (Basel Committee on Banking Supervision, 2011), and in many bank internal calculations of economic capital as well. Note that the VaR suggested by the $\text{LPM}_3(1)$

model at \$2,77m is about 32% larger than that suggested by the MC_1 model (\$2,10m). In fact, estimated models for higher-order Markov chains, which account for the downward *momentum*, show higher VaR compared with those models for 1-order Markov chain (MC_1 and $PPOM_1(B^*)$), which ignore rating *momentum* effects in the rating process.

Table 3.15: VaR for One-Year Risk Horizon for a credit portfolio consisting of 100 bonds with random exposure between \$1 and \$1m. Relative VaR for each model with respect to VaR calculated by the MC_1 model (in brackets).

Level	MC_1	$PPOM_1(B^*)$	$PPOM_2(B^*)$	$LPM_2(1)$	$LPM_3(1)$
99.0%	\$1,876,135	\$1,869,456	\$2,675,962	\$2,183,324	\$2,135,647
	-	(99.64%)	(142.63%)	(116.37%)	(113.83%)
99.9%	\$2,102,498	\$2,219,029	\$3,143,806	\$2,905,930	\$2,766,526
	-	(105.54%)	(149.53%)	(138.21%)	(131.58%)

3.7 Conclusions

We have presented a statistical model to fit higher-order Markov chain models with *long-persistence*, i.e models for which state transitions are rare and typically take place between states that are close to each other. One of the main problems with higher-order Markov chain models is the large number of parameters to be estimated. In our model, the *long-persistence* property of rating transitions is used to fit more parsimonious high-order Markov chain models in order to estimate the 1-year transition probabilities between rating states. We find that, in municipal ratings, information on the previous state is insufficient to explain the rating at the present time. Specifically, the probability of observing a particular rating at the current instant depends on the observed states over the last 3 years, i.e the order of the Markov chain for the best fitted model among those analyzed. One economic explanation for this result may be that the current rating of a public organization or local government affects its finance cost and limits its fund-raising options and therefore, is expected to affect the rating in the future. We also show that it is not necessary to specify all of the information concerning the previous history and a summary

of the same has led us to construct a simpler and more explanatory model with fewer parameters. Our model is a partial proportional model that becomes less restrictive than the POM, obtaining better results according to the BIC criterion.

One of the aims of this study is to understand the mechanism of rating migrations. The obtained results confirm the hypothesis for the *rating momentum* that was found in the extensive literature on this topic. We find that the estimated probability of being downgraded is higher in ratings with previous downgrades. Moreover, the longer the *persistence* of a rating, the more likely it is to remain unchanged in the following period. However, the results are different for rating 3, which behaves as a rating *barrier* in the sense that the probability of being downgraded is low even when the previous rating was downgraded. Indeed, the results show that when rating 3 has remained unchanged in the previous states, it is more likely to be upgraded, because rating 3 is the boundary between the *investment* grade and the *speculative* grade, and a debt issuer that falls into this rating may be expected to implement such measures as increasing tax or decreasing expenditure in order to improve its solvency and to avoid being downgraded.

Finally, an application to credit risk assessment indicates that VaR suggested by models for higher-order Markov chains is higher than the one obtained by models for 1-order Markov chain, in an adverse economic scenario. The model here proposed is able to identify rating *momentum* effects in the rating process without incurring high number of parameters. Thus by estimating parsimonious higher-order Markov models for rating transitions that account for the downward *momentum* can help to generate more accurate risk estimates of great practical relevance in many financial applications. An interesting line of research may be use this proposed approach to other ratings type showing *persistent* states, for example sovereign or corporate ratings, where the risk of default is more significant than in municipal ratings. It is likely that the differences in VaR between 1-order and higher-order Markov chain models are more accentuated for these populations.

CAPÍTULO 4

Exploring the randomness of mentally generated head-tail sequences

This Chapter corresponds to the contents of (Baena-Mirabete et al., 2019).

Abstract

It is well-known that people deviate from randomness as they attempt to mentally generate head-tail sequences as randomly as possible. This deviation from randomness is quantified by an excess of repetitions or alternations between successive responses more than would be expected by chance. We conducted an experiment in which a sample of students was asked to mentally simulate a sequence as if it is produced by a fair coin. We propose several models based on Markov chains for analyzing the dynamic of head-tail outcomes in these sequences. First, we explore observed Markov chains and suggest some practical solutions to reduce the number of parameters. However, there is a need for more sophisticated models and, in this case, we propose latent Markov models and mixture of Markov chains to analyze these head-tail sequences. A generalization of the so-called *mixture transition distribution model* (MTD) is also considered.

4.1 Introduction

Almost 70 years ago, Reichenbach (1949) claimed that untrained people would be unable to mentally generate random sequences and would instead produce series with too many alternations. Subsequent research has provided support for this claim and numerous experiments have been performed in order to study how the human mind attempts to generate random series and how the mind judges whether or not a particular series is produced at random. See the interesting review by Bar-Hillel & Wagenaar (1991) about different researches, approaches and experiments related to the study of the human perception of randomness over time. Other interesting review of findings from experiments with randomness production and randomness perception tasks can be seen in Nickerson (2002). A tendency to repeat the immediately preceding response is known in the literature as a *positive recency effect* and a tendency to avoid repeating it as a *negative recency effect*. Binary sequences that people mentally produce when they are trying to generate random ones have been reported typically to have more alternations (negative recency) than would be expected by chance (Budescu (1985);Lopes & Oden (1987);Falk & Konold (1997)). Negative recency has been the more common finding in the literature. However, some studies have also noticed some bias towards repetition (positive recency) in sequences mentally produced by some individuals (e.g. Neuringer (1986);Budescu (1987)).

The generation of *random* human sequences is a common task in experimental psychology. Some authors have even explored the capacity for generating sequences of random responses, relating it to the central executive component of working memory (Baddeley et al. (1998);Towse (1998)). When subjects are required to produce mentally random sequences they have to select new strategies to keep the sequence as random as possible, avoiding the occurrence of stereotyped responses. Before producing the response, one must then decide if it is sufficiently random given the previous responses, and, if not, change the strategy. All these selection and control functions correspond exactly to the role that is assigned by Baddeley (1986) to the central executive system of working memory. Typically, random generation studies have been conducted on normal young adults (e.g., college students). A number of studies have also examined random genera-

tion in neuropsychological patient groups, such as patients suffering from a dysexecutive syndrome, usually due to brain damage (Spatt & Goldenberg, 1993), patients with dementia of the Alzheimer type (Brugger et al., 1996) or in schizophrenia patients (Salamé & Danion, 2007). Random generation tasks have even been proposed in early diagnosis of Alzheimer's disease (García-Viedma et al., 2015). All these studies showed that sequences generated by brain-damaged patients were more stereotyped than those generated by their respective control group.

We present an application involving generation of *random* binary sequences by human subjects. We conducted an experiment in which Biology degree students were asked to mentally simulate a fair coin. To that end, the 262 participating students were each to produce a single sequence of 50 head-tail outcomes, by not considering the previous choices. The observational study here presented is framed in the context of longitudinal data analysis in which a single binary response, for a same student, is repeated at 50 different time points. We propose several parsimonious models based on Markov chains and their mixtures for exploring the randomness of these mentally generated binary sequences.

The Markov chain of order k ($1 \leq k < \infty$), taking values in a finite set of states $A = \{0, 1, ..., m-1\}$, $m \geq 2$, is a well-known probabilistic model that is usually used to describe short-memory series. An introduction to the Markov chain model for longitudinal data is given in Section 2.7 of the book by Bartolucci et al. (2013). However, Markov modelling in the context of categorical time series can be problematic because the number of parameters increases exponentially as the order k grows. Then, it is not possible to make reliable inference with a high number of parameters. Numerous studies have specifically focused on this issue, and some of them provide solutions to overcome it (Raftery (1985); Kharin & Petlitskii (2007); Kharin & Maltsau (2014)). Fokianos & Kedem (2003) showed the flexibility of regression methods for dealing with categorical time series, allowing for more parsimonious models and the incorporation of covariates, as opposed to other procedures. Likewise, Baena-Mirabete & Puig (2018) proposed several parsimonious models for higher-order Markov chains to analyze ordered multistate series.

For more sophisticated processes, a combination of more than one Markov chain could be an alternative. In activities concerning the production of randomness by hu-

man subjects and, in light of the aforementioned findings on this issue, one would expect the deviation from randomness to be explained by a finite mixture that captures population heterogeneity in transition probabilities. Mixture models based on Markov chains have been commonly used in the literature from different approaches. In Park & Basawa (2002), the authors proposed finite mixtures of Markov processes in which, conditional on an unobserved variable (latent process), the response variables followed a Markov process and, the latent process was assumed to be independent and identically distributed. Another approach that is widely used is the so-called latent Markov model which also assumes the existence of an unobservable (latent) process affecting the distribution of the response variables. In this case, the latent process is assumed to follow a Markov chain and, given this process, the response variables are assumed to be conditionally independent. From this latter approach, an interesting study can be seen in Albert (1991) who proposed a Markov mixture model for a time series of epileptic seizure counts. A similar study can be seen in Leroux & Puterman (1992) where the authors applied a Markov-dependent mixture of Poissons to model the number of fetal lamb movements. Vermunt et al. (1999) proposed a flexible version of the latent Markov model where both the initial and the transition probabilities of the latent process were depending on time-constant and time-varying covariates. Research and practice on latent Markov models have been increasingly considered in the literature. A detailed overview of this literature is provided in Section 1.2 of the book by Bartolucci et al. (2013). In Bartolucci et al. (2009), the authors presented a method for estimating the effect of nursing homes on the transition probabilities between states in a latent Markov model. Jackson et al. (2013) and Jackson et al. (2015), in a longitudinal study of teen driving, proposed the use of latent Markov models to relate risky driving behavior with the occurrence of a crash event. In an interesting study, Bartolucci et al. (2014a) proposed a model based on a mixture of latent auto-regressive AR(1) to analyze aspects related to retirement and health among elderly individuals in the USA. Latent Markov models may be also formulated to take into account certain types of unobserved heterogeneity. In some applications, it is unrealistic to assume that the process under study is equal for all individuals. Van de Pol & Langeheine (1990) proposed the so-called *mixed latent Markov model* in which the parameters are allowed to vary in different discrete latent subpopulations. In Bartolucci et al. (2017),

an illustration of the mixed latent Markov model by using the R (R Core Team, 2017) package LMest is provided. An extension of this approach was proposed by Altman (2007), which incorporates both covariates and continuous random effects to account for between-subject differences in the latent process in a more parsimonious way.

This paper addresses the analysis of mentally generated binary sequences using Markov chain models and their mixtures. We begin with exploring observed Markov chains parameterized as logistic models and proposing some solutions to reduce the number of parameters. Then, we propose mixture models based on Markov chains that, to our knowledge, have never been used before to analyze this type of data. These mixture models are analyzed from different approaches. First, we fit latent Markov models, allowing for the transition probabilities of the latent process to depend on lagged response variables as in Bartolucci & Pennoni (2007). Second, we fit the models described in Bartolucci & Farcomeni (2010) that are related to the so-called *mixture transition distribution model* (MTD) of Raftery (1985). In a third approach, we propose a mixture of Markov chains that accounts for an additional heterogeneity by adding a *subject-specific* random effect. This latter model extends the previously cited work of Park & Basawa (2002) to longitudinal data.

The paper is structured as follows: in Section 4.2 there is a brief overview of full parameterized higher-order Markov chain models for binary series. Some more parsimonious models are presented and fitted. In Section 4.3, several mixture models are introduced and studied. In Section 4.4, the main results are shown. Section 4.5 concludes the paper.

4.2 Higher-order binary Markov chains and some practical parameter reductions

4.2.1 Full parameterized binary Markov chain model

Let $\{X_t : t = 1, ..., T\}$ be discrete-time random variables taking values in a finite set of states $\{0, 1, ..., m-1\}$. Given that subscript t refers to a certain point in time, X_t is the

corresponding state at time t, and X_{t-k} is the state k periods before ($k < t \leq T$). Then, we have a k-order Markov chain if,

$$P(X_t = i_t | X_1 = i_1, \cdots, X_{t-1} = i_{t-1}) = P(X_t = i_t | X_{t-k} = i_{t-k}, \cdots, X_{t-1} = i_{t-1})$$

where $i_1, ..., i_T \in \{0, 1, ..., m-1\}$. This is the general definition but, in this paper, we will only be interested in binary series ($m = 2$).

Unfortunately, the number of parameters of these full parameterized models increases very rapidly: a k-order binary Markov chain model with two states has 2^k parameters. For example, for an order $k = 3$, the model has eight parameters to be estimated. From now on, MC_k will denote a full parameterized k-order binary Markov chain model. Thus, MC_0 corresponds to a series of independent outcomes.

Direct calculations show that, given the first k observations, the log-likelihood function of a MC_k model is

$$l(\mathbf{\Lambda}) = \sum_{\substack{\nu_k,...,\nu_1 \\ \nu_k,...,\nu_1 \in \{0,1\}}} n_{1;\nu_k,...,\nu_1} \log(p_{\nu_k,...,\nu_1}) + n_{0;\nu_k,...,\nu_1} \log(1 - p_{\nu_k,...,\nu_1}),$$

where $\mathbf{\Lambda}$ is the unknown 2^k x 2 transition matrix with row elements $[p_{\nu_k,...,\nu_1}, 1 - p_{\nu_k,...,\nu_1}]$ for all $\nu_k, ..., \nu_1 \in \{0, 1\}$, $n_{1;\nu_k,...,\nu_1}$ is the number of observed transitions from $X_{t-k} = \nu_k$, ..., $X_{t-1} = \nu_1$ to $X_t = 1$, and $p_{\nu_k,...,\nu_1}$ is its corresponding transition probability. Similarly, $n_{0;\nu_k,...,\nu_1}$ is the number of transitions from $X_{t-k} = \nu_k$, ..., $X_{t-1} = \nu_1$ to $X_t = 0$, and $1 - p_{\nu_k,...,\nu_1}$ is its transition probability. Note that each k-order Markov chain model is completely characterized by its transition probabilities. Hence, the maximum likelihood estimates of the transition probabilities are

$$\hat{p}_{\nu_k,...,\nu_1} = \frac{n_{1;\nu_k,...,\nu_1}}{n_{1;\nu_k,...,\nu_1} + n_{0;\nu_k,...,\nu_1}},$$

which can easily be obtained by cross-tabulation. The standard errors can be estimated using the usual expression for the standard deviation of an estimated proportion:

$$\widehat{SE}\left(\hat{p}_{\nu_k,...,\nu_1}\right) = \sqrt{\frac{\hat{p}_{\nu_k,...,\nu_1}(1 - \hat{p}_{\nu_k,...,\nu_1})}{n_{1;\nu_k,...,\nu_1} + n_{0;\nu_k,...,\nu_1}}}.$$

It is interesting to note that these results for the full parameterized MC_k model can also be obtained using an equivalent logistic model (e.g. Zeger & Qaqish (1988)) with

$$P(X_t = 1 | X_{t-k} = i_{t-k}, ..., X_{t-1} = i_{t-1}) = \frac{\exp\{\psi(i_{t-k}, ..., i_{t-1}; \boldsymbol{\beta})\}}{1 + \exp\{\psi(i_{t-k}, ..., i_{t-1}; \boldsymbol{\beta})\}}, \qquad (4.1)$$

where $\boldsymbol{\beta}$ is the vector of parameters of the linear predictor given by

$$\psi := \psi(X_{t-k}, ..., X_{t-1}; \boldsymbol{\beta}) = \beta_0 + \sum_{s=1}^{k} \beta_s X_{t-s} + \vartheta(k, \tilde{\boldsymbol{\beta}}), \tag{4.2}$$

$\vartheta(k, \tilde{\boldsymbol{\beta}})$ being the k-order function of the interaction terms and defined as

$$\vartheta(K, \tilde{\boldsymbol{\beta}}) := \vartheta(K; \tilde{\boldsymbol{\beta}}; X_{t-K}, ..., X_{t-1}).$$

Integer K is the order of the interactions and $\tilde{\boldsymbol{\beta}}$ denotes the vector of parameters for the interaction terms. In general, for $K > 1$, it may be written as

$$\vartheta(K, \tilde{\boldsymbol{\beta}}) = \sum_{j=1}^{K-1} \left(\sum_{s_1=1}^{K-j} \sum_{s_2=s_1+1}^{K-j+1} \sum_{\substack{s_3=s_2+1 \\ j \geq 2}}^{K-j+2} \cdots \sum_{\substack{s_K=s_{K-1}+1 \\ j \geq K-1}}^{K} \beta_{s_1 \ldots s_{j+1}} X_{t-s_1} \cdot \ldots \cdot X_{t-s_{j+1}} \right).$$

For example, taking $K = 2$ and $K = 3$ we have,

$$\vartheta(2, \tilde{\boldsymbol{\beta}}) = \beta_{12} X_{t-1} \cdot X_{t-2},$$

$$\vartheta(3, \tilde{\boldsymbol{\beta}}) = \beta_{12} X_{t-1} \cdot X_{t-2} + \beta_{13} X_{t-1} \cdot X_{t-3} + \beta_{23} X_{t-2} \cdot X_{t-3} + \beta_{123} X_{t-1} \cdot X_{t-2} \cdot X_{t-3}.$$

It is plain to see that the model defined above is equivalent to the full parameterized MC_k model. In the following section, we propose some practical solutions for constructing higher-order binary Markov chains with fewer parameters.

4.2.2 Parsimonious higher-order binary Markov chain models

As noted previously, the full parameterized MC_k model, as formulated in (4.2), requires 2^k parameters to be estimated. Raftery (1985) suggested the so-called *mixture transition distribution* (MTD) *model* to reduce the number of parameters from 2^k to $k+1$. However, the parameters of this model are not always easy to interpret. The BAR(k) model, i.e. the k-order binary autoregressive model proposed by Cox (1981), is another way to reduce the number of parameters. This model is obtained from equation (4.2) by removing the interaction terms, considering therefore $\vartheta(k, \tilde{\boldsymbol{\beta}}) = 0$. Consequently, only main effects are taken into account in this restrictive model. Similarly, Slud & Kedem (1994) proposed a logistic autoregressive model for binary time series that takes into account stochastic

time dependent covariates. See Section 2.7.2 of the book by Bartolucci et al. (2013) for different types of parameterization to reduce the number of parameters in the case of ordinal response variables.

We propose a less restrictive k-order binary Markov chain model that falls somewhere between the full parameterized model in (4.2) and the BAR(k) model by Cox (1981). Specifically, we suggest a MC$_k$ model as in (4.2), but which has interaction terms up to order r ($r < k$), as follows:

$$\psi = \beta_0 + \sum_{s=1}^{r} \beta_s X_{t-s} + \vartheta(r, \tilde{\boldsymbol{\beta}}) + \sum_{s=r+1}^{k} \beta_s X_{t-s}, \tag{4.3}$$

where $\vartheta(r, \tilde{\boldsymbol{\beta}})$ is the r-order function of the interaction terms, $\tilde{\boldsymbol{\beta}}$ being the vector of parameters for these interaction terms. Note that this model has $2^r + k - r$ parameters, fewer than the corresponding saturated MC$_k$ model, which has 2^k parameters. Thus, in the case of $r = 2$ and $k = 6$, this model has eight parameters to be estimated, the BAR(6) model has seven parameters and the MC$_6$ model has 64.

A more parsimonious model can be obtained from equation (4.3) under the hypothesis that $\beta_{r+1} = ... = \beta_k = \gamma$. That is,

$$\psi = \beta_0 + \sum_{s=1}^{r} \beta_s X_{t-s} + \vartheta(r, \tilde{\boldsymbol{\beta}}) + \gamma \sum_{s=r+1}^{k} X_{t-s}. \tag{4.4}$$

This model has $2^r + 1$ parameters to be estimated. From now on, it will be called the *k-order short-long-time memory* Markov chain model of degree r, which will be denoted SLMC$_k(r)$. The so-called short-time memory is represented by the outcomes of the r immediately previous observations $X_{t-r} = i_{t-r}, ..., X_{t-1} = i_{t-1}$. On the other hand, the long-time memory is represented by the influence of the states between the instants $t - r - 1$ and $t - k$, summarized by $\sum_{s=r+1}^{k} X_{t-s}$. SLMC$_k(0)$ refers to the k-order model that only takes into account the summary of states between $t - 1$ and $t - k$. It is plain to see that, in the particular case of $k = 1$, SLMC$_1(0)$ is equivalent to the MC$_1$ model.

Note that the SLMC$_k(r)$ model can be directly interpreted. Short-term memory is generally considered to have limited capacity. See Miller (1956) for an early study of the capacity limit associated with short-term memory. Therefore, it seems reasonable to believe that individuals are able to remember the immediately previous responses in an exact way, but only a summary of what happened many times before. In other contexts,

the model defined in (4.4) should be checked, especially by testing the hypothesis $\beta_{r+1} = ... = \beta_k = \gamma$.

To compare the *in-sample* performance of different models, in this paper we use the Bayesian information criterion (BIC). This index is defined by $BIC = -2\hat{l} + q\log(n)$, where $\hat{l}$ is the log-likelihood of the model, q is the number of parameters and n is the number of components in the log-likelihood. The model achieving the lowest BIC is chosen. Following the same procedure as in Raftery (1985) and Berchtold & Raftery (2002), we removed the first six observations of each mentally generated head-tail sequences to have the same number of components in the log-likelihood of each model, and thereby obtain suitable BIC comparisons between Markov chains up to order $k = 6$. See Katz (1981) for a discussion of the use of the BIC index in the context of Markov chains.

The following sections present mixture models to account for heterogeneity among subjects in transition probabilities. Before, the MC_k and $\text{SLMC}_k(r)$ models will be applied to the binary series produced by students in the random generation experiment carried out.

4.2.3 Application: the production of randomness by human subjects

We have conducted an experiment in which 262 Biology degree students at the Autonomous University of Barcelona were asked to mentally simulate a fair coin, providing a head-tail series as randomly as possible. The experiment was performed on different days and in groups of around 30 students with an age between 18 and 20 years. Each of these sessions was held early in the day and in the same classroom. After an informative session, each student was asked to mentally produce and record a series of 50 head-tail outcomes using a Windows dialog box designed with an Excel Visual Basic for Applications (VBA). To do this, the students were repeatedly required to click on the mouse and choose between two options shown on the screen: 'HEAD' or 'TAIL'. The results were recorded in a hidden worksheet so that the student could not see the previous choices. The effect of age on the random generation experiment has not been addressed and it is beyond the scope of this study. For a discussion on this issue, see for example Van der Linden et al. (1998).

For illustrative purposes, figure 4.1 shows the head-tail sequences generated by two individuals in the sample. It is interesting to note that the series generated by the second individual (id=222) seems to alternate responses more frequently than that produced by the first individual (id=91).

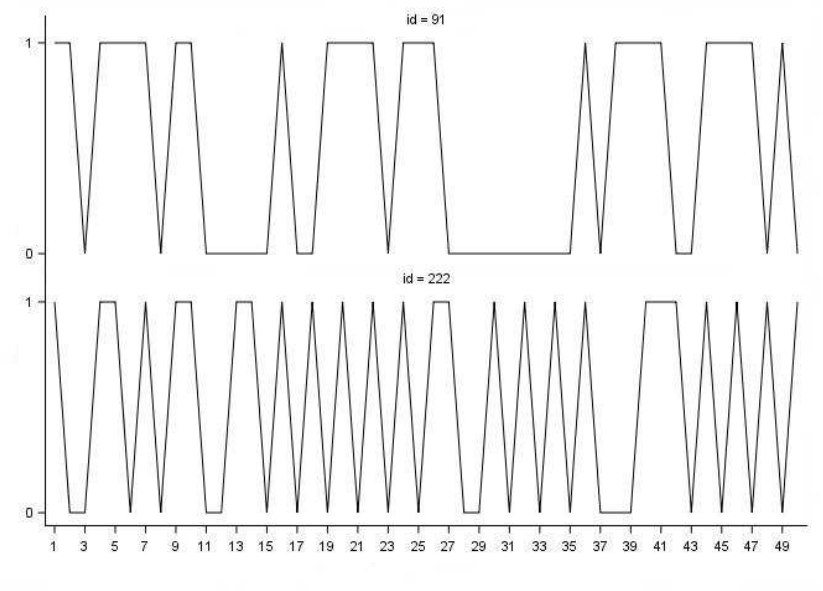

Figure 4.1: Head-tail series produced by two individuals in the sample.

The first approach to analyzing this dataset was to fit standard MC_k models to each subject's head-tail chain with different parameters for each individual. Because the number of parameters in these models increases rapidly with k, for each subject we have explored only those with $k = 0, 1, 2, 3$. The best model was selected according to the BIC index. Table 4.1 shows the frequency distribution of the students according to the best model. Note that MC_0, the independence model, is the best for most subjects (85%). For around 13% of the students the current state depends on the previous result and for only 3% of the students does the current state depend on the two previous outcomes.

Table 4.1: No. of students for which the corresponding MC_k model was the best.

Best model	MC_0	MC_1	MC_2	MC_3
No. students	222	33	7	0
% students	84.73	12.60	2.67	0.00

As already pointed out, the number of parameters of these full parameterized models increases very rapidly. This highlights the need for parsimonious models to explore higher-order dependencies. For this purpose, we estimate, for each subject, the parsimonious SLMC$_k(r)$ models, for $k = 0, \ldots, 6$ and $r = 0, \ldots, 3$ and also, the BAR(k) models. The parameters are estimated independently for each subject. Table 4.2 displays the frequency of students for which the corresponding Markov chain model was the best. For around 50% of students, the best model is one of order $k \geq 2$. Furthermore, note that for most of these students, the best model was an SLMC$_k(r)$ model where $r = 0$. That is, although individuals do not take the previous results exactly into account, they can remember a summary of what happened in the k previous outcomes before a certain instant. Comparing these results with those obtained in Table 4.1, we note that the percentage of subjects for which the independence model MC$_0$ is the best model decreases from 85% to 46%.

Table 4.2: No. of students for which the corresponding MC$_k$, BAR(k) or SLMC$_k(r)$ models was the best.

		MC$_k$	BAR(k)	SLMC$_k(r)$				Total	%
				$r:$ 0	1	2	3		
	0	120	*	-	-	-	-	120	45.8
	1	14	*	*	-	-	-	14	5.3
	2	0	2	20	*	-	-	22	8.4
$k:$	3	0	1	19	2	0	-	22	8.4
	4	0	0	15	0	1	0	16	6.1
	5	0	0	26	1	0	1	28	10.7
	6	0	0	35	5	0	0	40	15.3

$(*)$ Notes:

BAR(0) model is equivalent to MC$_0$ model.

SLMC$_1(0)$ and BAR(1) models correspond to MC$_1$ model.

SLMC$_2(1)$ model corresponds to BAR(2) model.

Table 4.3 summarizes the results from more parsimonious models by assuming common parameters across individuals (models 1 to 9). The best BIC is achieved by the model SLMC$_6(2)$ (model 9) which exactly takes into account the two previous outcomes, and a summary of the results between $t-3$ and $t-6$. As described, the SLMC$_6(2)$ model

Table 4.3: Fitting results and total number of parameters (q) for several estimated Markov chains with common parameters across individuals. $\hat{l}$ is the global log-likelihood.

Model number	Model	q	$\hat{l}$	BIC
1	MC_0	1	-7990	15989
2	MC_1	2	-7947	15914
3	MC_2	4	-7891	15820
4	$SLMC_3(1)$	3	-7891	15810
5	$SLMC_4(1)$	3	-7885	15798
6	$SLMC_5(1)$	3	-7857	15744
7	$SLMC_6(1)$	3	-7847	15721
8	BAR(6)	7	-7830	15726
9	$SLMC_6(2)$	5	-7831	15710

(equation (4.4)) is a restricted version of the model defined in (4.3) under the null hypothesis $H_0 : \beta_3 = \beta_4 = \beta_5 = \beta_6 = \gamma$. The Likelihood Ratio Test did not find enough statistical evidence to reject it. Note also that the estimated log-likelihood is almost the same as in model 8, the BAR(6). This suggests that the exact results prior to $t - 2$ do not produce a significant improvement in the fit of the data. This may be explained by the limitations of short-term memory (e.g. Miller (1956);Kareev (1992);Rapoport & Budescu (1997)), in the sense that individuals can remember some previous results in an exact way but only a summary of what happened many times before a certain instant. The results for the models reported in this Section can be obtained by using standard statistical packages, specifically we have used the GENMOD procedure in SAS v9.3 (SAS Institute, 2011).

4.3 Mixture models

In the following sections, we analyze several mixture models from different approaches that take into account the unobserved heterogeneity among subjects: (1) A latent Markov model which assumes the existence of an unobservable (latent) process affecting the distribution of the response variables. The latent process is assumed to follow a Markov chain and, given this process, the response variables are assumed to be conditionally in-

dependent. (2) A generalization proposed by Bartolucci & Farcomeni (2010) of the MTD model (Raftery, 1985) based on latent Markov models. (3) We propose a mixture of high-order Markov chains where the latent process, conditional on a *subject-specific* random effect, is assumed to be independent and identically distributed. From this latter approach, we consider that, conditional on the latent process, the response variables follow a high-order Markov chain. The dataset and the codes used in this Section are available upon request and can also be found through the link: http://www.statmod.org/smij/archive.html.

4.3.1 Latent Markov model

There is extensive literature on the analysis of longitudinal data. For a review see, among others, Fitzmaurice et al. (2008). In this literature, latent Markov (LM) models (Wiggins, 1973) represent an important milestone because they are widely used in many fields. A detailed description may be found in Vermunt (2010) and Bartolucci et al. (2014b). Let $\{X_{jt} = i_t : t = 1, ..., T\}$ denote the observed sequence of binary response variables over time for the jth individual with $j = 1, ..., N$ and $i_1, ..., i_T \in \{0, 1\}$. The LM model considers that given a particular value of the latent variable C_{jt} taking a number of finite discrete values (latent states), the observed variables over time are independent, which is often referred to as the local independence assumption. For a discussion on identifiability issues see Petrie (1969). Then, we model the conditional response probabilities $h_{1|c} := P(X_{jt} = 1|C_{jt} = c)$, with $c = 1, ..., z$ where z takes finite discrete values, by using a logistic function as follows,

$$h_{1|c} = \frac{\exp\{\alpha_c\}}{1 + \exp\{\alpha_c\}}. \tag{4.5}$$

In the above expression, the parameters α_c are the support points corresponding to each latent state. It is also assumed that $C_{j1}, ..., C_{jT}$ follow a 1-order Markov chain with state space $\{1, ..., z\}$ where transition probabilities are modeled according to a multinomial *logit* parameterization,

$$\log \frac{P(C_{jt} = c|C_{j,t-1} = \ddot{c}, X_{j,t-k} = i_{t-k}, ..., X_{j,t-1} = i_{t-1})}{P(C_{jt} = \ddot{c}|C_{j,t-1} = \ddot{c}, X_{j,t-k} = i_{t-k}, ..., X_{j,t-1} = i_{t-1})} = \psi(i_{t-k}, ..., i_{t-1}; \boldsymbol{\beta}_{\ddot{c}\boldsymbol{c}}), \tag{4.6}$$

for $t = k+2, ..., T$ and $\ddot{c}, c = 1, ..., z$ with $\ddot{c} \neq c$ and where $\boldsymbol{\beta}_{\ddot{c}c}$ is the vector of parameters. We also assume that the underlying processes belong to the SLMC$_k(r)$ class, that is,

$$\psi(X_{j,t-k}, ..., X_{j,t-1}; \boldsymbol{\beta}_{\ddot{c}c}) = \beta_{0\ddot{c}c} + \sum_{s=1}^{r} \beta_{s\ddot{c}c} X_{j,t-s} + \vartheta_j(r, \tilde{\boldsymbol{\beta}}_{\ddot{c}c}) + \gamma_{\ddot{c}c} \sum_{s=r+1}^{k} X_{j,t-s},$$

where $\vartheta_j(r, \tilde{\boldsymbol{\beta}}_{\ddot{c}c})$ is the r-order function of the interaction terms, $\tilde{\boldsymbol{\beta}}_{\ddot{c}c}$ being the vector of parameters for these interaction terms. Note that $z(z-1)$ linear equations are defined in (4.6) and, therefore, there are $z(z-1)(2^r+1)$ parameters contained in the set of vectors $\boldsymbol{\beta}_{\ddot{c}c}$. A natural way to allow the initial probabilities of the latent model, at time $k+1$, to depend on lagged responses is by adopting a multinomial *logit* parameterization as follows,

$$\log \frac{P(C_{j,k+1} = c | X_{j,t-k} = i_{t-k}, ..., X_{j,t-1} = i_{t-1})}{P(C_{j,k+1} = 1 | X_{j,t-k} = i_{t-k}, ..., X_{j,t-1} = i_{t-1})} = \bar{\psi}(i_{t-k}, ..., i_{t-1}; \boldsymbol{\beta}_{\boldsymbol{c}}), \qquad (4.7)$$

for c=2,...,z and again, we assume that the linear predictor equations $\bar{\psi}$ in (4.7) belong to the SLMC$_k(r)$ class. Specifically, we propose a parsimonious SLMC$_k(0)$ specification, that is, $\bar{\psi}(X_{j,t-k}, ..., X_{j,t-1}; \boldsymbol{\beta}_{\boldsymbol{c}}) = \beta_{0c} + \gamma_c \sum_{s=1}^{k} X_{j,t-s}$. From now on, z-LM_SLMC$_k(r)$ will denote a LM model, as described above, with z latent states.

It is plain to see that the number of parameters to be estimated for the z-LM_SLMC$_k(r)$ model is $3z + z(z-1)(2^r+1) - 2$. For instance, the number of parameters for the 2-LM_SLMC$_6(2)$ model is equal to 14. In constrast, the 3-LM_SLMC$_6(2)$ model has 37 parameters to be estimated. The EM algorithm (Dempster et al., 1977) is commonly used to obtain maximum likelihood estimates for the parameters of the latent Markov model, especially when based on discrete latent variables. An overview of the EM algorithm focused on the LM models can be found in Bartolucci et al. (2014b). The goodness-of-fit results for the z-LM_SLMC$_k(r)$ models reported here have been obtained using the *est_lm_cov_latent* function in the R (R Core Team, 2017) package LMest (Bartolucci et al., 2017).

4.3.2 Latent Markov extension of the mixture transition distribution model

The mixture transition distribution model (MTD), as mentioned earlier, was introduced by Raftery (1985) to approximate high-order Markov chains with far fewer parameters than

the fully parameterized model. In our particular case, given a sample of N individuals, for the jth individual, we observe realizations of a sequence of binary random variables over time $\{X_{jt} = i_t : t = 1, ..., T\}$, $j = 1, ..., N$ and $i_1, ..., i_T \in \{0, 1\}$. Let $\{C_{jt} : t = 1, ..., T\}$ denote independent and identically distributed *occasion-specific* latent variables with state space $\{1, ..., k\}$. The MTD of order k assumes that

$$P(X_{jt} = 1|C_{jt} = c, X_{j,t-k} = i_{t-k}, ..., X_{j,t-1} = i_{t-1}) = P(X_{jt} = 1|C_{jt} = c, X_{j,t-c} = i_{t-c}),$$

for $t = k + 1, ..., T$ and $c = 1, ..., k$. Thus, the response variable X_{jt} depends only on the lagged variable $X_{j,t-c}$, where the lag c is chosen by a random mechanism which is not directly observable.

Bartolucci & Farcomeni (2010) proposed a generalization of the MTD model based on the assumption that the latent variables were not independent and identically distributed, but followed a homogenous 1-order Markov chain. We adapt this model to the case of longitudinal data and binary response. Specifically, we model

$$h_{j,1|c} := P(X_{jt} = 1|C_{jt} = c, X_{j,t-c} = i_{t-c}),$$

with $c = 1, ..., k$ and where $h_{j,1|c}$ is based on a logistic function as follows,

$$h_{j,1|c} = \frac{\exp\{\psi(i_{t-c}; \boldsymbol{\beta_c})\}}{1 + \exp\{\psi(i_{t-c}; \boldsymbol{\beta_c})\}}. \tag{4.8}$$

The linear predictor in the above equation is given by:

$$\psi(X_{j,t-c}; \boldsymbol{\beta_c}) = \beta_{0c} + \beta_{1c} X_{j,t-c},$$

where $\boldsymbol{\beta_c} = (\beta_{0c}, \beta_{1c})$ is the vector of parameters of the linear predictor for each latent state. Here, the latent process $\{C_{jt} : t = 1, ..., T\}$ follows a homogenous 1-order Markov chain with initial probabilities $w_c = P(C_{j,k+1} = c)$, $c = 1, ..., k$, and transition probabilities $\phi_{c_1,c_2} = P(C_{jt} = c_2|C_{j,t-1} = c_1)$, $c_1, c_2 = 1, ..., k$ for $t > k + 1$. Following a similar notation to that of Bartolucci & Farcomeni (2010), this extended MTD model will be denoted LM_gMTD$_k$. It is plain to see that the number of parameters for the LM_gMTD$_k$ model is $k^2 + 2k - 1$. A reduction in the number of parameters can be obtained from equation (4.8) by imposing that $\boldsymbol{\beta_c} = \boldsymbol{\beta}$ and, therefore, that the coefficients are not lag-specific, leading to a more parsimonious model. This constrained model will be denoted

LM_MTD$_k$. Note that, in this case, the number of parameters is $k^2 + 1$. See Bartolucci & Farcomeni (2010) for details on the maximum likelihood estimation of the parameters in these models using an EM algorithm (Dempster et al., 1977). We have implemented a function in R (R Core Team, 2017) to estimate both the LM_gMTD$_k$ and the LM_MTD$_k$ models described in this Section.

4.3.3 Mixture of high-order Markov chains

The model specified in equation (4.1) considers that the probability of a certain event occurring is determined by a observed Markov chain of order k. However, in some contexts, more than one process might be involved. For example, in experiments concerning the production of randomness, people tend to produce too many alternations of outcomes. However, as mentioned earlier, some studies have also reported a certain bias towards repetition in some individuals. It could therefore be assumed that, when individuals have to mentally simulate the outcomes of a fair coin, they combine more than one process to produce head-tail sequences. However, one would expect the mixing weights of each process to be different among individuals.

Suppose we have a sample of N individuals and, for the jth individual, we observe realizations of a sequence of binary random variables over time $\boldsymbol{x_j} = \{X_{jt} = i_t : t = 1, ..., T\}$, $j = 1, ..., N$ and $i_1, ..., i_T \in \{0, 1\}$. We assume that, conditional on an *occasion-specific* latent variable C_{jt} taking a number of finite discrete values (latent states), the random variables $\{X_{jt} : t = k + 1, ..., T\}$ follow a k-order Markov chain. Specifically, we model

$$h_{j,1|c} := P(X_{jt} = 1|C_{jt} = c, X_{j,t-k} = i_{t-k}, ..., X_{j,t-1} = i_{t-1}),$$

with $c = 1, 2$ and where $h_{j,1|c}$ is based on a logistic function as follows,

$$h_{j,1|c} = \frac{\exp\{\psi(i_{t-k}, ..., i_{t-1}; \boldsymbol{\beta_c})\}}{1 + \exp\{\psi(i_{t-k}, ..., i_{t-1}; \boldsymbol{\beta_c})\}}. \tag{4.9}$$

We assume that the underlying processes in the above equation belong to the SLMC$_k(r)$ class. In other words, the linear predictor is given by:

$$\psi(X_{j,t-k}, ..., X_{j,t-1}; \boldsymbol{\beta_c}) = \beta_{0c} + \sum_{s=1}^{r} \beta_{sc} X_{j,t-s} + \vartheta_j(r, \tilde{\boldsymbol{\beta}}_{\boldsymbol{c}}) + \gamma_c \sum_{s=r+1}^{k} X_{j,t-s},\ c = 1, 2$$

where $\boldsymbol{\beta_c}$ is the vector of parameters of the linear predictor for each latent state. Once again, $\vartheta_j(r, \tilde{\boldsymbol{\beta}}_{\boldsymbol{c}})$ is the r-order function of the interaction terms, $\tilde{\boldsymbol{\beta}}_{\boldsymbol{c}}$ being the vector of parameters for these interaction terms.

Let u_j be a continuous random effect associated with the jth individual and let $\{u_j\}_{j=1}^{N}$ be independent and identically distributed. Then, we assume that, conditional on the *subject-specific* random effect u_j, the latent variables $C_{j1}, ..., C_{jT}$ are independent. We propose modeling the conditional probabilities $w_{j,c|u} := P(C_{jt} = c|u_j)$, $c = 1, 2$, as

$$w_{j,1|u} = \Phi^{-1}(\alpha + u_j), \tag{4.10}$$

where Φ is the cumulative standard normal distribution, α is a common parameter for all individuals and $w_{j,2|u} = 1 - w_{j,1|u}$.

Note that, conditional on the random effect u_j, the random variables $\{X_{jt} : t = k + 1, ..., T\}$ follow a 2-component mixture of Bernoullis given by:

$$X_{jt}|u_j \sim \begin{cases} Bern\left(h_{j,1|1}\right), & \text{with probability } w_{j,1|u} \\ Bern\left(h_{j,1|2}\right), & \text{with probability } w_{j,2|u} \end{cases} \tag{4.11}$$

Here, $w_{j,c|u}$, $c = 1, 2$, are the mixing probabilities (weights) as defined in (4.10) that we assume to be independent across individuals. For each Bernoulli component, parameter $h_{j,1|c}$, $c = 1, 2$, is not considered constant; rather, it changes in accordance with the transition probabilities of two Markov chains as defined in (4.9).

For the jth individual, the transition probabilities conditioned on the random effect u_j are denoted $p_{i_{t-k},...,i_{t-1}|u_j}$. These can be obtained by taking the conditional expectation of (4.11), that is,

$$P(X_{jt} = 1|X_{j,t-k} = i_{t-k}, \cdots, X_{j,t-1} = i_{t-1}; u_j; \boldsymbol{\xi}) = \sum_{c=1}^{2} w_{j,c|u} h_{j,1|c}, \tag{4.12}$$

where $\boldsymbol{\xi}$ is the vector of all model parameters.

The likelihood contribution, for the jth individual, $t \geq k + 1$, is given by,

$$L_j\left(\boldsymbol{\xi}; \boldsymbol{x_j}\right) = \int_{u_j} \left(\prod_{t=k+1}^{T} f_j\left(i_t|i_{t-k}, ..., i_{t-1}; u_j, \alpha, \boldsymbol{\beta_1}, \boldsymbol{\beta_2}\right) \right) g\left(u_j|\sigma\right) du_j, \tag{4.13}$$

where

$$f_j\left(i_t|i_{t-k}, ..., i_{t-1}; u_j, \alpha, \boldsymbol{\beta_1}, \boldsymbol{\beta_2}\right) = \sum_{c=1}^{2} w_{j,c|u} \left(h_{j,1|c}\right)^{i_t} \left(1 - h_{j,1|c}\right)^{1-i_t},$$

and $g\left(u_j|\sigma\right)$ is the density of the random effect, which is assumed to be normally distributed with standard deviation σ.

Since individuals are assumed to be independent, the global likelihood is given by the expression $L(\boldsymbol{\xi}; \boldsymbol{x_1}, ..., \boldsymbol{x_N}) = \prod_{j=1}^{N} L_j(\boldsymbol{\xi}; \boldsymbol{x_j})$. The maximum likelihood estimates of the parameters, which will be denoted by $\hat{\boldsymbol{\xi}}$, can be obtained by numerical maximization of the global likelihood. To do so, the integral of expression (4.13) needs to be evaluated. A classic way to compute the integral is by means of adaptive Gaussian quadrature, which approximates the integral by a weighted sum over predefined abscissas of the random effect, as described in Pinheiro & Bates (1995).

Empirical Bayes estimates can be used to assign values to the random effect u_j for each individual, which is a widely used method in the extensive literature on this issue (see Morris (1983) and Gaver & O'Muircheartaigh (1987)). An excellent discussion of the prediction of random effects and expected responses in generalized linear mixed models (Breslow & Clayton, 1993) can be found in Skrondal & Rabe-Hesketh (2009). For the model proposed in equation (4.12), the posterior distribution of the random effect would be,

$$v(u_j|\boldsymbol{x_j}; \hat{\boldsymbol{\xi}}) = \frac{g(u_j|\hat{\sigma}) f_j(\boldsymbol{x_j}|u_j; \hat{\alpha}, \hat{\boldsymbol{\beta}}_{\mathbf{1}}, \hat{\boldsymbol{\beta}}_{\mathbf{2}})}{L_j(\hat{\boldsymbol{\xi}}; \boldsymbol{x_j})},$$

where $L_j(\hat{\boldsymbol{\xi}}; \boldsymbol{x_j})$ is simply the likelihood contribution of the jth individual and,

$$f_j(\boldsymbol{x_j}|u_j; \hat{\alpha}, \hat{\boldsymbol{\beta}}_{\mathbf{1}}, \hat{\boldsymbol{\beta}}_{\mathbf{2}}) = \prod_{t=k+1}^{T} f_j(i_t|i_{t-k}, ..., i_{t-1}, u_j; \hat{\alpha}, \hat{\boldsymbol{\beta}}_{\mathbf{1}}, \hat{\boldsymbol{\beta}}_{\mathbf{2}})$$

is the conditional likelihood given a realization of the random variable u_j.

Then, the estimated transition probability $\hat{p}^{(j)}_{i_{t-k},...,i_{t-1}}$ of the jth individual is obtained by plugging the estimated parameters $\hat{\boldsymbol{\xi}}$ into equation (4.12), and integrating over the posterior density, for each individual:

$$\hat{p}^{(j)}_{i_{t-k},...,i_{t-1}} = \int_{-\infty}^{\infty} \left(\sum_{c=1}^{2} \hat{w}_{j,c|u} \hat{h}_{j,1|c} \right) v(u_j|\boldsymbol{x_j}; \hat{\boldsymbol{\xi}}) du_j.$$

Also, the empirical Bayes estimate of the random effect u_j can be calculated as the expected value of the empirical posterior distribution given by:

$$\hat{u}_j = \mathbf{E}\left[u_j|\boldsymbol{x_j}; \hat{\boldsymbol{\xi}}\right] = \int_{-\infty}^{\infty} u_j v(u_j|\boldsymbol{x_j}; \hat{\boldsymbol{\xi}}) du_j. \tag{4.14}$$

A more straightforward way to obtain estimates of the transition probability for the jth individual is to plug directly into equation (4.12) the empirical Bayes estimate of the random effect and the estimated parameters $\hat{\boldsymbol{\xi}}$. However, it is important to note that, for non-linear link functions, estimates obtained by both approaches are not the same, and the above expression should be solved by numerical integration as well (e.g. Gibbons et al. (1994)).

Similarly, we can obtain the estimated transition probability $\hat{p}_{i_{t-k},\ldots,i_{t-1}}$, for the population, by integrating the estimated conditional probabilities in equation (4.12) over the prior density as follows,

$$\hat{p}_{i_{t-k},\ldots,i_{t-1}} = \int_{-\infty}^{\infty} \left(\sum_{c=1}^{2} \hat{w}_{j,c|u} \hat{h}_{j,1|c} \right) g\left(u_j|\hat{\sigma}\right) du_j, \tag{4.15}$$

which can be solved by numerical integration. The delta method can be used to obtain the standard error of prediction (e.g. Rosen et al. (2000) and Thijs et al. (2002)).

The model defined in (4.11) can be extended to take into account a higher number of latent states ($c > 2$). Basically, the previously shown equations would be exactly identical and only the mixing weights $w_{j,c|u}$ should be rewritten to allow for more than two mixture components. We define the random mixing weights in terms of continuation-ratio (Fienberg, 1980) in a similar way to that described in Dayton & Macready (1988), in which mixing weights were modeled as functions of fixed factors based on continuation-ratio *logits*. In our model, for a mixture of three components, we would propose a continuation-ratio *probit* specification for the random mixing weights as follows,

$$\begin{aligned} w_{j,1|\mathbf{u}} &= \Phi^{-1}(\alpha_1 + u_{j1}) \\ w_{j,2|\mathbf{u}} &= (1 - w_{j,1|\mathbf{u}})\Phi^{-1}(\alpha_2 + u_{j2}) \\ w_{j,3|\mathbf{u}} &= 1 - w_{j,1|\mathbf{u}} - w_{j,2|\mathbf{u}}, \end{aligned} \tag{4.16}$$

where the vector of random effects follows a normal bivariate distribution $(u_{j1}, u_{j2}) \sim N(0, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma}$ being the corresponding covariance matrix. From now on, the mixture model described in this Section, with z latent states, will be called the z-component mixture model of *k-order short-long-time memory* Markov chains of degree r, which will be denoted z-SLMC$_k(r)$. The z-SLMC$_k(r)$ model, as it is proposed here, have been implemented using the NLMIXED procedure in SAS v9.3 (SAS Institute, 2011).

4.4

Application continued: analysis of head-tail sequences

We continue with the analysis of head-tail sequences mentally produced by a group of students in the random generation experiment carried out. Table 4.4 presents the goodness of fit results for several mixture models: z-LM_SLMC$_k(r)$, z-SLMC$_k(r)$ and extended MTD models as described, respectively, in Sections 4.3.1, 4.3.3 and 4.3.2. The comparison of these results with the BIC obtained for the observed Markov chains (without mixture) in Table 4.3 leads to a better goodness of fit for the mixture models, with the exception of those related to the MTD models. Among the fitted models, the 2-SLMC$_6(2)$ (models 32 and 33) yielded the smallest BIC index. The 3-SLMC$_k(r)$ mixture models given in (4.16) were also considered. These had a worse goodness of fit than those obtained by the 2-SLMC$_k(r)$ models: for the fitted 3-SLMC$_6(1)$ model we obtained a BIC=15717, while for the 2-SLMC$_6(1)$ model the BIC value was smaller. Extended MTD models, from 19 to 27, achieved the worst results of fitting.

Once the previous analysis comparing models is complete, we select the 2-SLMC$_6(2)$ (model 33) to fit the dataset. Parameter estimates, once the non-significant interaction term has been removed, are shown in Table 4.5. For the first mixture component, note that all the estimated parameters are statistically significant (p-value < 0.0001). The negative sign on the slope parameters indicates that the probability of alternating responses is higher than the probability of persistent responses by producing, therefore, alternations from heads to tails more often than really happen at random. The estimated negative effect for the long-memory variable ($\hat{\gamma_1}$) can be interpreted as a tendency to balance the distribution of responses in short sequence fragments. For the second mixture component, it is interesting to note the positive sign on the previous response variable (b_{12}), unlike in the first mixture component. This implies a higher probability of repeating a response, i.e. a clustering effect of identical responses. Again, the negative effect for the long-memory variable ($\hat{\gamma_2}$) indicates a certain tendency, although not significant, to balance the distribution of responses when the previous outcomes before instant $t-2$ present many repetitions of identical responses (0 or 1). The estimated random effect for individual is statistically significant with an estimated standard deviation equal to 1.16.

Table 4.4: Fitting results and total number of parameters (q) for several estimated mixture models. $\hat{l}$ is the global log-likelihood. Models 10 to 18 are z-LM_SLMC$_k(r)$ models (Section 4.3.1). Models 19 to 27 are related to the MTD models (Section 4.3.2). Models 28 to 33 are z-SLMC$_k(r)$ models (Section 4.3.3).

Model number	Model	q	$\hat{l}$	BIC
10	2-LM_SLMC$_3(1)$	10	-7862	15781
11	2-LM_SLMC$_4(1)$	10	-7849	15754
12	2-LM_SLMC$_5(1)$	10	-7831	15718
13	2-LM_SLMC$_6(1)$	10	-7823	15702
14	2-LM_SLMC$_4(2)$	14	-7840	15758
15	2-LM_SLMC$_5(2)$	14	-7824	15727
16	2-LM_SLMC$_6(2)$	14	-7813	15704
17	2-LM_SLMC$_6(2)^*$	12	-7813	15693
18	3-LM_SLMC$_6(2)^*$	31	-7759	15692
19	LM_gMTD$_2$	7	-7916	15898
20	LM_gMTD$_3$	14	-7914	15958
21	LM_gMTD$_4$	23	-7912	16040
22	LM_gMTD$_5$	34	-7910	16137
23	LM_MTD$_2$	5	-7919	15886
24	LM_MTD$_3$	10	-7918	15930
25	LM_MTD$_4$	17	-7918	15995
26	LM_MTD$_5$	26	-7911	16066
27	LM_MTD$_6$	37	-7907	16159
28	2-SLMC$_4(1)$	8	-7870	15783
29	2-SLMC$_5(1)$	8	-7835	15715
30	2-SLMC$_6(1)$	8	-7825	15695
31	2-SLMC$_5(2)$	12	-7817	15701
32	2-SLMC$_6(2)$	12	-7804	15674
33	2-SLMC$_6(2)^*$	9	-7804	15658

$(*)$ Notes:

Estimated model by removing non-significant interaction term.

Table 4.6 shows the conditional probabilities resulting from the estimated 2-SLMC$_6(2)$ model, for responses generated by each mixture component. By focusing on responses drawn from the first mixture component, note that the estimated probability of alternating response is higher than that of repeating the previous outcomes. For example, it is interesting to compare the estimated probability of $X_t = 1$ in the case that all previous responses until instant $t-6$ are equal to zero, and that obtained in the case that of all them being equal to one: these were 0.8426 and 0.1495, respectively. Note, however, that, for

Table 4.5: Estimated parameters for the 2-SLMC$_6(2)$ mixture model.

Parameter	Estimate	Standard Error	t-statistic	p-value
$\hat{\sigma}$	1.16	0.47	2.48	0.0137
$\hat{\alpha}$	0.68	0.45	1.51	0.1310
Component 1				
b_{01}	1.68	0.29	5.77	<0.0001
b_{11}	-1.05	0.23	-4.53	<0.0001
b_{21}	-0.83	0.13	-6.10	<0.0001
$\hat{\gamma_1}$	-0.38	0.07	-5.69	<0.0001
Component 2				
b_{02}	-0.15	0.23	-0.63	0.5283
b_{12}	0.61	0.30	2.07	0.0399
b^*_{22}	0.00	-	-	-
$\hat{\gamma_2}$	-0.10	0.07	-1.37	0.1723

$(*)$ Notes:

The estimate obtained for b_{22} is -0.02 with p-value equal to 0.9150. The model is re-estimated by removing this parameter.

a response coming from the second mixture component, the estimated conditional probabilities of repeating the previous outcomes are higher than those of alternating. Besides, it is also interesting to note that the observations produced by the second mixture component seem to be closer to randomness (probability equal to 0.50) than those coming from the process characterized by the first component. With the goal of testing the hypothesis that the second mixture component could be the independence model (MC$_0$), we establish the null hypothesis $H_0 : \beta_{12} = \beta_{22} = \gamma_2 = 0$, i.e. that all slope parameters are equal to zero. To this end, we use the Likelihood Ratio Test which leads us to reject H_0 (p-value < 0.01). Therefore, the process derived from the second mixture component is statistically different from MC$_0$, because of some bias towards repetition of the previous outcomes.

The estimated mixing weight for individual is calculated as defined in equation (4.10), that is $\hat{w}_{j,1|u} = \Phi^{-1}(\hat{\alpha} + \hat{u}_j)$, where $\hat{u}_j$ is the Empirical Bayes estimate of the *subject-specific* random effect given by equation (4.14). For an average individual with random effect $u_j = 0$, the estimated mixing weight is 0.75, which indicates a high probability of the first mixture component in his/her mental randomization. For illustrative purposes,

Table 4.6: Estimated probabilities of X_t for each mixture SLMC$_6(2)$ component, conditioned to the values of the variables X_{t-1}, X_{t-2} and $\sum_{s=3}^{6} X_{t-s}$. Approximate standard errors (s.e) for the estimates have been calculated using the delta method (in parentheses).

			Component 1			Component 2		
$\sum_{s=3}^{6} X_{t-s}$	X_{t-2}	X_{t-1}	X_t		s.e	X_t		s.e
			0	1		0	1	
0	0	0	0.1574	0.8426	(0.0386)	0.5371	0.4629	(0.0585)
1	0	0	0.2152	0.7848	(0.0394)	0.5608	0.4392	(0.0459)
2	0	0	0.2869	0.7131	(0.0373)	0.5843	0.4157	(0.0377)
3	0	0	0.3712	0.6288	(0.0340)	0.6074	0.3926	(0.0363)
4	0	0	0.4642	0.5358	(0.0336)	0.6300	0.3699	(0.0418)
0	1	0	0.3001	0.6999	(0.0382)	0.5371	0.4629	(0.0585)
1	1	0	0.3862	0.6138	(0.0293)	0.5608	0.4392	(0.0459)
	...		...	...	...			
2	0	1	0.5351	0.4649	(0.0200)	0.4322	0.5678	(0.0399)
3	0	1	0.6281	0.3719	(0.0276)	0.4559	0.5441	(0.0473)
4	0	1	0.7125	0.2875	(0.0358)	0.4798	0.5202	(0.0591)
0	1	1	0.5509	0.4491	(0.0338)	0.3859	0.6141	(0.0451)
1	1	1	0.6429	0.3571	(0.0328)	0.4089	0.5911	(0.0392)
2	1	1	0.7254	0.2746	(0.0352)	0.4322	0.5678	(0.0399)
3	1	1	0.7949	0.2051	(0.0368)	0.4559	0.5441	(0.0473)
4	1	1	0.8505	0.1495	(0.0359)	0.4798	0.5202	(0.0591)

for the head-tail series produced by individual j=91 shown in Figure 4.1, who exhibits clusters of identical responses, the estimated mixing weight for the first component is 0.16. Instead, for individual j=222, who shows a greater tendency to alternate responses, the estimated weight is 0.93. Figure 4.2 shows the histogram for the estimated probability of the first mixture component. Note that the distribution is skewed to the right, showing that, in general, individuals give a greater weight to the process characterized by the first mixture component, which is therefore the one that is primarily responsible for the deviation from randomness.

Table 4.7 shows the marginal probabilities estimated by the 2-SLMC$_6(2)$ model obtained by solving by Monte Carlo simulation the integral defined in equation (4.15). Here, it is interesting to note that individuals tend to balance the distribution of responses when the set of previous observations presents many identical responses.

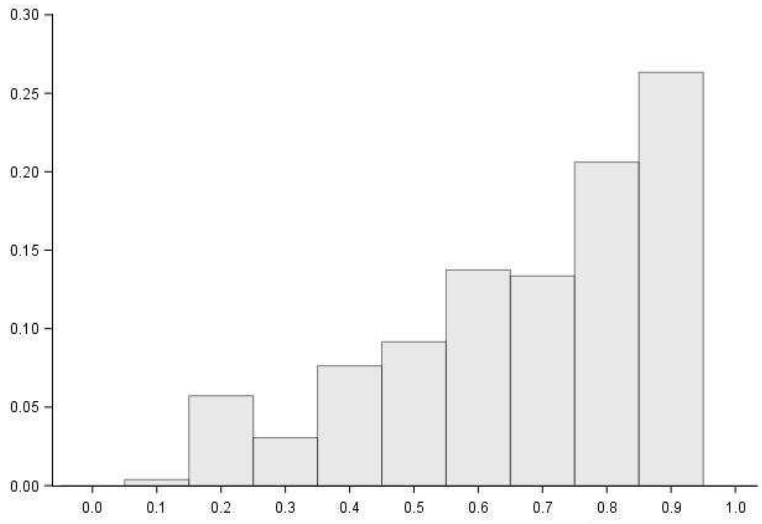

Figure 4.2: Histogram of the mixing weight ($\hat{w}_{j,1|u}$) estimates for individual.

4.5 Discussion

Differences in the random generation between individuals have been reported in the extensive literature on this issue. As mentioned, negative recency has been the more common finding, however, some studies have also noticed some bias towards repetition of responses (positive recency) in some individuals. We have presented several models based on Markov chains to analyze mentally generated head-tail sequences. Observed Markov chain models (Section 4.2) typically used in this field assume that head-tail outcomes depend directly on the immediately previous choices. We found that the results are improved when taking into account the unobserved heterogeneity between individuals that cannot be explained only on the basis of the previous observed choices. In this regard, latent Markov models, as proposed in Section 4.3.1, consider that head-tail responses depend indirectly on the immediately previous choices only through an unobservable (latent) process. In contrast, MTD models (Section 4.3.2) assume that individuals generate a certain head-tail response depending only on one of the immediately lagged responses, which is chosen according to an unobservable mechanism. The poor performance of MTD models suggests that individuals, in their mental randomization, keep in mind the whole set

Table 4.7: Probabilities for values of X_t estimated by the 2-SLMC$_6(2)$ mixture model conditioned to some values of the variables X_{t-1}, X_{t-2} and $\sum_{s=3}^{6} X_{t-s}$. Approximate standard errors (s.e) for the estimates are calculated using the delta method (in parentheses).

$\sum_{s=3}^{6} X_{t-s}$	X_{t-2}	X_{t-1}	X_t		s.e
			0	1	
0	0	0	0.2808	0.7192	(0.0318)
1	0	0	0.3328	0.6672	(0.0254)
2	0	0	0.3850	0.6150	(0.0194)
3	0	0	0.4454	0.5546	(0.0160)
4	0	0	0.5170	0.4830	(0.0178)
0	1	0	0.3726	0.6274	(0.0196)
1	1	0	0.4430	0.5570	(0.0140)
	...		...	...	...
2	0	1	0.5005	0.4995	(0.0105)
3	0	1	0.5712	0.4288	(0.0140)
4	0	1	0.6324	0.3676	(0.0194)
0	1	1	0.4932	0.5068	(0.0182)
1	1	1	0.5663	0.4337	(0.0164)
2	1	1	0.6293	0.3707	(0.0196)
3	1	1	0.6829	0.3171	(0.0254)
4	1	1	0.7335	0.2665	(0.0314)

of the immediately previous choices, not just one of them, to produce a certain head-tail outcome. Indeed, we show that models which take into account a summary of the six previous outcomes are among those producing the best results. Therefore, the individuals in the experiment are at least vaguely capable of remembering the six previous head-tail choices. This finding is consistent with that discussed by Miller (1956) who argued that the number of elements an average human can hold in working memory is 7 ± 2. Subsequent studies corroborate this upper limit on our capacity to process information (e.g. Saaty & Ozdemir (2003)). Among the fitted models, 2-component mixture models of high-order Markov chains, as proposed in Section 4.3.3, yielded the best results. From this approach, we show that individuals deviate from randomness when they try to generate sequences as randomly as possible due to the combination of two processes: the first, produces an excess of alternations and a tendency to balance the distribution of outcomes in short fragments. The second process produces too many repetitions. Moreover, the role

that each of these processes plays in this experiment depends on the individual.

Our random generation task was conducted considering a single head-tail sequence mentally produced by each individual. However, some studies have shown the sensitivity of the results to changes in the experimental conditions. For instance, Baddeley et al. (1998) analyzed the influence of generation rate, showing that, randomness decreased when subjects had less time at their disposal to complete the task. It could be interesting to consider more than one sequence produced by each individual by varying experimental conditions for a more precise understanding of the dynamic of head-tail outcomes. Finally, the proposed approach here provides a tool that may be useful for identifying differences in the mental generation of random numbers in brain-damaged patients, such as patients with dementia of the Alzheimer type. García-Viedma et al. (2015) designed a random generation task, with minimum memory requirements, to specifically evaluate attentional control in Alzheimer's disease patients and it was even proposed in early diagnosis of the disease. It was applied to both healthy elderly people and patients with Alzheimer's disease and the results revealed significant differences between both groups. This is an interesting topic to investigate in future work.

CAPÍTULO 5

Computing probabilities of integer-valued random variables by recurrence relations

This Chapter corresponds to the contents of (Baena-Mirabete & Puig, 2020).

Abstract

We derive a set of recurrence relations for the calculation of the probabilities of a large class of integer-valued random variables. We show that the probability mass function can be recursively computed for random variables with a probability generating function satisfying certain functional form.

5.1

Introduction

In recent times, many statistical methods have been applied for analysing football (soccer) results, see for instance Karlis & Ntzoufras (2003), Gómez-Déniz et al. (2019), and the references therein. Let us define the difference $Z = G_1 - G_2$ of the goals scored by two opposing teams. The number of goals scored by each team can be expressed as $G_h = X_h + U$, for $h = 1, 2$, where the count random variable U reflects game and environmental conditions. Therefore, G_1 and G_2 are correlated, but X_1 and X_2 can be assumed independent, and the difference of goals $Z = X_1 - X_2$ is just the difference of two independent count random variables. Karlis & Ntzoufras (2003) studied the particular case where X_h are Poisson distributed. The efficient computation of the probabilities of Z is essential for computing the likelihoods and fitting models with covariates able to explain the difference of goals. But how to efficiently compute the probabilities of the difference (or the sum) of independent random variables? This will be the main objective of the present research.

Let X be a count random variable (rv) taking values $n = 0, 1, ...$ and let $P(X = n)$ be denoted by p_n. For some families of probability distributions, there exists a simple recurrence relationship between p_n and p_{n-1}, involving n and the parameters of the distribution, like in the case of the Katz-Panjer family (Katz, 1965; Panjer, 1981). Sundt (1992) extended this family by considering count distributions whose probability mass function (pmf) satisfied a recurrence relation of the form,

$$p_n = \sum_{j=1}^{r} (\alpha_j + \beta_j/n) p_{n-j}, \qquad n = 1, 2, ... \tag{5.1}$$

for some positive integer r and constants α_j, $\beta_j \in \mathbb{R}$, $j = 1, ..., r$, with $p_n = 0$ for $n < 0$. This class of distributions contains the Katz-Panjer family as a special case, taking $r = 1$. Theorem 1 by Sundt (1992) shows that the probability generating function (pgf), denoted $f(z)$ with $z \in \mathbb{C}$, for a count distribution arising from (5.1), satisfies

$$\frac{\mathrm{d}}{\mathrm{d}z} \ln f(z) = \frac{\sum_{j=1}^{r} (j\alpha_j + \beta_j) z^{j-1}}{1 - \sum_{j=1}^{r} \alpha_j z^j}. \tag{5.2}$$

Note that the right term in (5.2) is the ratio between a polynomial of degree at most $r - 1$ and a polynomial of degree at most r with a non-zero constant term.

In this paper, we extend some of the results by Sundt (1992), by generalizing the condition (5.2) to polynomials of arbitrary degree in z. Thus, we provide recurrence relations for the calculation of the pmf of a large family of count distributions, including those with integer support on $\mathbb{Z}$. The paper is organized as follows: In Section 5.2, we define a general class of distributions obtaining recurrence relations for their probabilities in Proposition 1; relevant applications of Proposition 1 are presented in Section 5.3; Section 5.4 gives the conclusions. A study of the numerical behaviour of some of the recurrence relations is shown in the Online Supplement.

5.2

Recurrence relations for the probabilities of integer-valued and count distributions

Let X be an integer-valued rv, and let $p_n := P(X = n)$ be its pmf defined on $n \in \mathbb{Z}$ (including both positive and negative integers). The pgf of X, defined as $f(\omega) = \mathbb{E}(\omega^X)$, $\omega \in (0, 1)$, can always be analytically extended at least to the complex domain $0 < |z| < 1$. Let consider the class of rv's with a pgf, $f(z)$, satisfying

$$\frac{\mathrm{d}}{\mathrm{d}z} \ln f(z) = \frac{G(z)}{T(z)}, \tag{5.3}$$

where $G(z) = \sum_{j=0}^{r} \psi_j z^j$ and $T(z) = \sum_{j=0}^{s} \eta_j z^j$ are polynomials of degree r and s respectively, with $z \in \mathbb{C}$. The functional form (5.3) is satisfied by many well known families of count distributions, such as the Katz-Panjer family, the Skellam and the Hermite distribution. Note that the class of distributions satisfying (5.2) is contained in those satisfying (5.3).

The pgf of a rv defined on $\mathbb{Z}$ can be expressed as a Laurent series around zero,

$$f(z) = \mathbb{E}(z^X) = \sum_{n=-\infty}^{\infty} p_n z^n. \tag{5.4}$$

with $0 < |z| < 1$. For rv's defined on $\mathbb{N}$, the nth-derivative of the pgf at $z = 0$, divided by $n!$, provides the probability p_n. However, the derivatives at $z = 0$ do not have sense for pgf's of rv's defined on $\mathbb{Z}$. Alternatively, the probabilities can be computed by using

Cauchy's integral formula, that is,

$$p_n = \frac{1}{2\pi i} \oint_{\gamma} \frac{f(z)}{z^{n+1}} \, \mathrm{d}z, \tag{5.5}$$

where $i = \sqrt{-1}$ and γ is a closed contour whose interior contains $z = 0$. When $f(z)$ is analytic in a region containing the unit disc, taking the contour γ to be the unit circle, $|z| = 1$, parameterized by $z(\theta) = \exp(i\theta)$, $0 \leq \theta \leq 2\pi$, $\mathrm{d}z = i \exp(i\theta)\mathrm{d}\theta$, the contour integral (5.5) can be evaluated as,

$$p_n = \frac{1}{2\pi} \int_0^{2\pi} \exp(-i\theta n) f[\exp(i\theta)] \, \mathrm{d}\theta. \tag{5.6}$$

Expression (5.6) is especially useful for the numerical evaluation of the probabilities of many integer-valued distributions at a low computational cost. Zhu & Joe (2010) proposed to compute the probabilities by inverting the characteristic function of X, that is, $\varphi_X(t) = \mathbb{E}(e^{itX}) = f(e^{it})$, with $t \in \mathbb{R}$. The inversion formula provides the cumulative distribution function:

$$\Gamma(n) := P(X < n) = \frac{1}{2} - \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{Re}\left(\frac{\exp(-i\mu n) f[\exp(i\mu)]}{1 - \exp(-i\mu)} \right) \mathrm{d}\mu.$$

Therefore, $p_n = \Gamma(n+1) - \Gamma(n)$ and,

$$\begin{aligned} \Gamma(n+1) - \Gamma(n) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{Re}\left(\frac{[\exp(-i\mu n) - \exp(-i\mu(n+1))] f[\exp(i\mu)]}{1 - \exp(-i\mu)} \right) \mathrm{d}\mu = \\ &= \frac{1}{2\pi} \int_0^{2\pi} \mathrm{Re}\left(\exp(-i\mu n) f[\exp(i\mu)] \right) \mathrm{d}\mu, \end{aligned}$$

and this is just expression (5.6), because the imaginary part of the integral in (5.6) is equal to 0. Next result shows that the probabilities of integer-valued and count rv's with a pgf satisfying (5.3) can be recursively obtained in an efficient way.

Proposition 1 *The pmf of a rv with pgf fulfilling condition* (5.3) *satisfies the following recurrence relation,*

$$\sum_{j=0}^{s} (n-j)\eta_j p_{n-j} = \sum_{j=0}^{r} \psi_j p_{n-j-1}. \tag{5.7}$$

Proof: First, notice that solving the integral in (5.5) by integration by parts, taking $u = f(z)$ and $\mathrm{d}v = z^{-(n+1)}$, one has that

$$p_n = \frac{1}{2\pi i} \oint_{\gamma} \frac{f'(z)}{nz^n} \, \mathrm{d}z. \tag{5.8}$$

From equation (5.3), one has that $T(z)f'(z) = G(z)f(z)$. Therefore, dividing both sides by z^n and integrating, one obtains

$$\oint_{\gamma} \frac{\left(\sum_{j=0}^{s} \eta_j z^j\right) f'(z)}{z^n} \, \mathrm{d}z = \oint_{\gamma} \frac{\left(\sum_{j=0}^{r} \psi_j z^j\right) f(z)}{z^n} \, \mathrm{d}z. \tag{5.9}$$

Using equations (5.8) and (5.5) in the left and right terms of the expression (5.9), respectively, one directly obtains the recurrence relation (5.7), which ends the proof. □

5.3 Examples of application

5.3.1 Difference of two independent count random variables

Discrete distributions defined on $\mathbb{Z}$ have attracted the attention of many researchers. A common way to construct integer-valued distributions is taking the difference of two independent count rv's X_h, $h = 1, 2$, each of them with pgf $f_{X_h}(z)$. Therefore, considering $X = X_1 - X_2$, direct calculations show that the pgf of X is,

$$f_X(z) = f_{X_1}(z) f_{X_2}(z^{-1}). \tag{5.10}$$

If $f_X(z)$ satisfies (5.3), then, by Proposition 1, the pmf of the rv X can be recursively obtained by (5.7). Note that a sufficient condition for satisfying (5.3) is that this condition holds for both $f_{X_h}(z)$. Table 5.1 shows some examples of recurrence relations for the pmf of the difference of two frequently used count rv's.

The most frequent distribution on $\mathbb{Z}$ is the Skellam distribution, obtained as the difference of two independent Poisson rv's $X_h \sim Pois(\lambda_h)$, $h = 1, 2$, where $\lambda_h > 0$ are the parameters. See, for example, Koopman et al. (2017) for a study of tick-by-tick discrete price changes for four liquid stocks traded on the New York Stock Exchange, where the

authors used a zero-deflated Skellam distribution. Its pmf can be expressed in a closed-form in terms of the modified Bessel function of the first kind. As far as we know, the recurrence relation shown in Table 5.1 is new, and it can be very useful for practical work with this distribution.

An alternative to Skellam is the distribution of the difference of two independent Negative Binomial rv's $X_h \sim NB(v_h, q_h)$, $h = 1, 2$, with $v_h > 0$, $q_h \in (0, 1)$. Its probabilities can also be recursively computed through the recurrence relation shown in Table 5.1. An interesting application of this distribution can be seen in Song & Smith (2011), who used it in a study for identifying dispersed epigenomic domains from ChIP-Seq data. In other field of application, Barndorff-Nielsen et al. (2012) used also this distribution for modeling discrete price changes in a financial context.

Another example of recurrence relation is that obtained for the difference of two independent binomial rv's $X_h \sim Bin(v_h, q_h)$, $h = 1, 2$, $v_h > 0$, $q_h \in (0, 1)$, with finite support in $\{-v_2, ..., v_1\}$. See Section 3.12.2 of the book by Kotz et al. (2005) for details and properties of this distribution.

As a last example also shown in Table 5.1, we propose a novel integer-valued distribution arising from the difference of two independent Hermite rv's $X_h \sim Her(a_h, b_h)$, $h = 1, 2$, with parameters $a_h, b_h > 0$. Interestingly, this distribution can be expressed as the distribution of $W_1 + 2W_2$, being W_1 and W_2 independent Skellam rv's with parameters a_1, a_2 and b_1, b_2 respectively. The proof is direct using the fact that a Hermite distribution with parameters a, b can be expressed as the distribution of $Y_1 + 2Y_2$, being Y_1 and Y_2 independent Poisson rv's with means a and b respectively. From here, it is straightforward to calculate the first moments of a Hermite difference distribution: $\mathbb{E}(X) = a_1 - a_2 + 2(b_1 - b_2)$ and $\mathbb{V}(X) = a_1 + a_2 + 4(b_1 + b_2)$.

Table 5.1: Examples of recurrence relations for $p_n = P(X = n)$, where $X = X_1 - X_2$, X_h, $h = 1, 2$, both independent.

X_h	pgf of X_h	$T(z)$, $G(z)$	$p_n := P(X = n)$
$X_h \sim Pois(\lambda_h)$	$f_{X_h}(z) = \exp[\lambda_h(z-1)]$	$T(z) = z^2$ $G(z) = -\lambda_2 + \lambda_1 z^2$	$\begin{cases} p_{n+1} = \frac{\lambda_1}{\lambda_2} p_{n-1} - \frac{n}{\lambda_2} p_n, & n = 0, 1, ... \\ p_{n-1} = \frac{n}{\lambda_1} p_n + \frac{\lambda_2}{\lambda_1} p_{n+1}, & n = 0, -1, ... \end{cases}$
$X_h \sim NB(v_h, q_h)$	$f_{X_h}(z) = \left(\frac{1-q_h}{1-q_h z}\right)^{v_h}$	$T(z) = -q_2 z + (1 + q_1 q_2) z^2 - q_1 z^3$ $G(z) = -v_2 q_2 + (v_2 - v_1) q_1 q_2 z + v_1 q_1 z^2$	$\begin{cases} p_{n+1} = \frac{(1-n-v_1)q_1}{(n-v_2+1)q_2} p_{n-1} + \frac{n+(n+v_1-v_2)q_1 q_2}{(n-v_2+1)q_2} p_n, & n = 0, 1, ... \\ p_{n-1} = \frac{n+(n+v_1-v_2)q_1 q_2}{(n+v_1-1)q_1} p_n + \frac{(v_2-n-1)q_2}{(n+v_1-1)q_1} p_{n+1}, & n = 0, -1, ... \end{cases}$
$X_h \sim Her(a_h, b_h)$	$f_{X_h}(z) = \exp[a_h(z-1) + b_h(z^2-1)]$	$T(z) = z^3$ $G(z) = -2b_2 - a_2 z + a_1 z^3 + 2b_1 z^4$	$\begin{cases} p_{n+1} = \frac{b_1}{b_2} p_{n-3} + \frac{a_1}{2b_2} p_{n-2} + \frac{1-n}{2b_2} p_{n-1} - \frac{a_2}{2b_2} p_n, & n = 0, 1, ... \\ p_{n-1} = -\frac{a_1}{2b_1} p_n + \frac{1+n}{2b_1} p_{n+1} + \frac{a_2}{2b_1} p_{n+2} + \frac{b_2}{b_1} p_{n+3}, & n = 0, -1, ... \end{cases}$
$X_h \sim Bin(v_h, q_h)$	$f_{X_h}(z) = (1 - q_h + q_h z)^{v_h}$	$T(z) = (1-q_1)q_2 z + (1 - q_1 - q_2 + 2q_1 q_2) z^2 + (1 - q_2) q_1 z^3$ $G(z) = (q_1 - 1) v_2 q_2 + (v_1 - v_2) q_1 q_2 z + (1 - q_2) v_1 q_1 z^2$	$\begin{cases} p_{n+1} = \frac{q_1(1-q_2)(v_1-n+1)}{q_2(1-q_1)(n+v_2+1)} p_{n-1} + \frac{(v_1-v_2-2n)q_1 q_2 + n(q_1+q_2-1)}{q_2(1-q_1)(n+v_2+1)} p_n, & n = 0, 1, ..., v_1 \\ p_{n-1} = \frac{(v_2-v_1+2n)q_1 q_2 + n(1-q_1-q_2)}{q_1(1-q_2)(v_1-n+1)} p_n + \frac{q_2(1-q_1)(n+v_2+1)}{q_1(1-q_2)(v_1-n+1)} p_{n+1}, & n = 0, -1, ..., -v_2 \end{cases}$

The initial values of the recurrence relations shown in Table 5.1 can be computed by evaluating numerically the integral (5.6). For example, for the Hermite difference distribution, substituting its pgf resulting from (5.10) into (5.6), one has that

$$\begin{aligned} p_n &= \frac{C}{2\pi}\int_0^{2\pi} \exp[a_1\exp(i\theta)+a_2\exp(-i\theta)+b_1\exp(2i\theta)+b_2\exp(-2i\theta)-i\theta n]\,\mathrm{d}\theta = \\ &= \frac{C}{2\pi}\int_0^{2\pi} \Gamma(\theta)\exp[(a_1+a_2)\cos(\theta)+(b_1+b_2)\cos(2\theta)]\,\mathrm{d}\theta, \end{aligned}$$

where $\Gamma(\theta)=\cos[(a_1-a_2)\sin(\theta)+(b_1-b_2)\sin(2\theta)-n\theta]$ and $C=\exp\left(-a_1-b_1-a_2-b_2\right)$.

5.3.2

Sum of independent count random variables

In some contexts, one may be interested in studying the distribution of the sum of count rv's. For instance, the total claims on a portfolio of insurance contracts is usually a rv of interest. Then, the total claims is modeled as the sum of all claims on the individual policies, which are assumed to be independent. Let X_h, $h=1,...,k$, be k independent rv's, and we define $X=X_1+...+X_k$. Then, the pgf of the rv X is given by, $f_X(z)=\prod_{h=1}^{k} f_{X_h}(z)$. Again, if each $f_{X_h}(z)$ satisfies (5.3), $f_X(z)$ also satisfies (5.3) and then, by Proposition 1, the pmf of X can be recursively obtained by (5.7). Table 5.2 shows some examples of recurrence relationships for the pmf of the sum of two independent count rv's ($k=2$). These examples could be directly generalized for $k>2$. Initial values for these recursive relations can be obtained directly by computing the convolution sum, or alternatively, by evaluating numerically the integral (5.6).

In the first example, we derive a recurrence relation for the distribution of the sum of two Binomial rv's. See Ong (1995) for a discussion of several physical models leading to this distribution. Butler & Stephens (2017) developed several methods to approximate the distribution of the sum of independent Binomial variables. It is interesting to remark that the recurrence relation for the difference of two independent binomial rv's (Table 5.1), can be deduced from that obtained for the sum of two binomial rv's (Table 5.2). The proof is straightforward using the properties of sums and differences of Binomial-type rv's (see Section 3.12.2 of the book by Kotz et al. (2005)). Specifically, let X_h, $h=1,2$,

be two Binomial rv's with parameters $v_h > 0$, $q_h \in (0, 1)$ and X_3 be a Binomial rv with parameters v_2 and $1 - q_2$, all them independent. Then, it holds that $P(X_1 - X_2 = n) = P(X_1 + X_3 = n + v_2)$.

Another example of recurrence relation also shown in Table 5.2 is that obtained for the distribution of the sum of two independent Negative Binomial rv's. See Furman (2007) for the exact, and cumbersome, derivation of the pmf of this distribution.

The next two examples of recurrence relations shown in Table 5.2 are related to the so-called INAR models (Al-Osh & Alzaid, 1987) and its extensions. The maximum likelihood estimation problem connected with INAR models is computationally intensive because it implies the convolution of the count distribution of the so-called innovations and that of a Binomial. The classical Poisson-INAR model (called INAR(1)) assumes that the innovations are Poisson distributed. The exact, but computationally inefficient, pmf of the sum of Poisson and Binomial rv's both independent can be found, for instance, in Freeland & McCabe (2004).

Some authors have proposed other extensions of the classical INAR model. For instance, Fernández-Fontelo et al. (2017) proposed a generalization whose innovations followed a Hermite distribution. We show in Table 5.2 that the probabilities needed for computing the likelihood function in these models can also be recursively obtained in an efficient way.

5.4 Conclusions

In this paper we present a way for constructing recurrence relations for the probabilities of integer-valued and count random variables belonging to a large class of distributions which includes the class of Sundt (1992). As far as we know, the recurrence relations provided in Tables 5.1 and 5.2 are new. They are useful for a practical and efficient computation of the probabilities and its related likelihood function. It is also remarkable the simple recursive relation found for the probabilities of the Skellam distribution which are usually computed using a modified Bessel function.

An anonymous referee has suggested, as an useful example of application of our

methodology, to calculate the recurrence relation for an INAR(2) model. The computation of the likelihood function for higher-order INAR(p) models are qualified numerically intractable in the literature, and various methods have been proposed (see e.g. Pedeli et al. (2015) and Lu (2018)). Our methodology is also suitable for this kind of problems. The conditional pmf of an INAR(2) model is the convolution of two Binomial distributions with parameters $v_h > 0$, $q_h \in (0,1)$, $h = 1, 2$, and the Poisson distribution of the innovations with parameter $\lambda > 0$. Direct calculations show that the pgf in this case fulfils the equation (5.3) and then, by Proposition 1, the pmf satisfies the following recurrence relation,

$$p_{n+1} = \frac{1}{(n+1)(1-q_1-q_2+q_1q_2)} \left[\lambda q_1 q_2 p_{n-2} + A(n)p_{n-1} + B(n)p_n\right], \quad n = 2, 3, ...$$

where $A(n) = q_1q_2(v_1+v_2) + (q_1+q_2-2q_1q_2)\lambda - (n-1)q_1q_2$
and $B(n) = v_1q_1 + v_2q_2 - q_1q_2(v_1+v_2) + (1-q_1-q_2+q_1q_2)\lambda - n(q_1+q_2-2q_1q_2)$.

The initial values needed for the recursive relations can be numerically computed from expression (5.6). This expression can also be useful for checking the computed values of p_n, for n large, in anticipation of possible round-off errors produced after many recurrences.

Table 5.2: Examples of recurrence relations for $p_n = P(X = n)$ where $X = X_1 + X_2$, X_h, $h = 1, 2$, both independent.

X_h	pgf of X_h	$T(z), G(z)$	$p_n := P(X = n)$
$X_h \sim Bin(v_h, q_h)$	$f_{X_h}(z) = (1 - q_h + q_h z)^{v_h}$	$T(z) = 1 - q_1 - q_2 + q_1q_2 + (q_1 + q_2 - 2q_1q_2)z + q_1q_2z^2$ $G(z) = v_1q_1 + v_2q_2 - (v_1 + v_2)q_1q_2 + (v_1 + v_2)q_1q_2z$	$p_{n+1} = \frac{(v_1+v_2+1-n)q_1q_2}{(n+1)(q_1-1)(q_2-1)}p_{n-1} + \frac{(n-v_1)(q_2-1)q_1+(n-v_2)(q_1-1)q_2}{(n+1)(q_1-1)(q_2-1)}p_n, \quad n = 1, 2, ..., v_1 + v_2$
$X_1 \sim Bin(v, q)$ $X_2 \sim Pois(\lambda)$	$f_{X_1}(z) = (1 - q + qz)^v$ $f_{X_2}(z) = \exp[\lambda(z - 1)]$	$T(z) = 1 - q + qz$ $G(z) = vq + (1 - q)\lambda + \lambda qz$	$p_{n+1} = \frac{\lambda q}{(n+1)(1-q)}p_{n-1} + \frac{vq+(1-q)\lambda-nq}{(n+1)(1-q)}p_n, \quad n = 1, 2, ...$
$X_1 \sim Bin(v, q)$ $X_2 \sim Her(a, b)$	$f_{X_1}(z) = (1 - q + qz)^v$ $f_{X_2}(z) = \exp[a(z - 1) + b(z^2 - 1)]$	$T(z) = 1 - q + qz$ $G(z) = vq + a(1 - q) + (2b(1 - q) + aq)z + 2bqz^2$	$p_{n+1} = \frac{2bq}{(n+1)(1-q)}p_{n-2} + \frac{aq+(1-q)2b}{(n+1)(1-q)}p_{n-1} + \frac{vq+(1-q)a-nq}{(n+1)(1-q)}p_n, \quad n = 2, 3, ...$
$X_h \sim NB(v_h, q_h)$	$f_{X_h}(z) = \left(\frac{1-q_h}{1-q_hz}\right)^{v_h}$	$T(z) = 1 - (q_1 + q_2)z + q_1q_2z^2$ $G(z) = v_1q_1 + v_2q_2 - (v_1 + v_2)q_1q_2z$	$p_{n+1} = \frac{(1-n-v_1-v_2)q_1q_2}{n+1}p_{n-1} + \frac{v_1q_1+v_2q_2+n(q_1+q_2)}{n+1}p_n, \quad n = 1, 2, ...$

5.5 Numerical behaviour (Online Supplement)

Recursive formulas can be very useful to compute the probabilities for a given model. Table 5.3 shows the computing execution times of the probabilities obtained using the recursive formulas, and those computed evaluating numerically the Cauchy's integral (equation (5.6)), for two of the examples previously considered in Section 5.3. Cauchy's integral has been computed using the function *integrate()* of the base package in R software (R Core Team, 2017). Both methods are compared with the so-called 'direct' calculation. For the Skellam distribution, direct calculation has been done with the *Skellam* package (Lewis et al., 2016) in R software (R Core Team, 2017), which uses the modified Bessel function of the first kind. For the sum of Poisson and Binomial rv's, direct calculation means to compute directly the convolution sum. The computing execution times reported in Table 5.3 have been obtained using the *Microbenchmark* package (Mersmann, 2019) in R software (R Core Team, 2017). Note that for the Skellam distribution the direct evaluation is around 13 times slower than the recursive evaluation. For the sum of Poisson and Binomial rv's the execution time of the direct calculation is nearly 2 times slower than the numerical integration and around 40 times slower than the recursive formula.

Table 5.3: Execution time (in seconds $\times 10^6$) for the calculation of the probabilities p_n for: (1) a Skellam distribution with parameters $\lambda_1 = 10$ and $\lambda_2 = 15$ with $n = -25, ..., 25$; (2) the sum of Poisson and Binomial rv's with parameters $\lambda = 20$ and $v = 60$, $q = 0.3$, respectively, with $n = 20, ..., 70$.

Distribution	Direct	Recurrence	Cauchy integral
Skellam	1575.13	123.73	7497.65
Poisson + Binomial rv's	14552.88	365.20	8454.59

$(*)$ Notes: these runs were carried out on a Toshiba Portege Z30-C-10W with two 2.30 GHz Intel Core i5-6200U and 8 GB of RAM.

As with any algorithm, round-off errors may occur with recursive evaluations since computers only represent a finite number of digits. The propagation of such round-off errors in a recursive evaluation may grow without bound and may lead to meaningless values. Theorems 7 and 9 by Panjer & Wang (1993) are remarkable because establish the

stability (in a strong sense) of those linear recurrences with non-negative coefficients. As illustration, consider the recursive relation for the probabilities of the Skellam distribution shown in Table 5.1. Note that the coefficients of the forward recursion are all non-negative, for $n < 0$, leading to a stable recurrence accordingly to Panjer & Wang (1993). Similarly, the coefficients of the backward recursion are all non-negative for $n > 0$.

Table 5.4 and Table 5.5 show some results of accuracy measured in terms of the relative error, calculated as $(p_n - \tilde{p}_n)/\tilde{p}_n$, where $\tilde{p}_n$ is the exact value obtained using the direct calculation mentioned above. In the case of the Skellam recurrence, one can observe from Table 5.4 that, for $n > 0$, round-off errors blow up rapidly in the forward direction; instead, the results in the backward direction confirm the strong stability with an overall small relative error of order 10^{-15}. Therefore, one can combine both directions to efficiently evaluate the probabilities of the Skellam distribution.

A combined usage of forward and backward directions can also be suitable for other recurrence relations studied in Section 5.3. As illustration, note that for the sum of Poisson and Binomial rv's (Table 5.2), the coefficients of the forward recursion are all non-negative over the range $0 \leq n \leq n^*$, with $n^* = \lfloor v + (1-q)\lambda/q \rfloor$. Likewise, the coefficients of the backward recursion are all non-negative for $n > n^*$. Table 5.5 shows that a combined evaluation in both directions (in this example $n^* = 5$) leads to very stable results with an overall small relative error of order 10^{-15}.

However, the numerical stability of recurrence relations cannot always be guaranteed. It happens, for instance, for the recursive relation of the difference of two independent Hermite rv's (Table 5.1). Note that for the forward recursion, the coefficients are all non-negative for $n \leq 1$, except the coefficient $-a_2/2b_2$. Analogously, the backward recursion has non-negative coefficients for $n \geq -1$, except the coefficient $-a_1/2b_1$. To explore in this case the numerical behaviour, we have computed the probabilities $\{p_n : n = -21, ..., 0\}$ of a Hermite difference distribution with parameters $(a_1, b_1, a_2, b_2) = (2, 1, 12, 2)$, by using forward recursion with initial values p_{-25} to p_{-22}. We have obtained that the relative error (e_n) was gradually increasing from $e_{-21} = -4.257337 \times 10^{-16}$ to $e_0 = -2.616686 \times 10^{-4}$. It may be a good practice for such unstable recurrences to check the computed values of p_n after many recurrences for assessing their accuracy.

Table 5.4: Recursive evaluation (in both directions) of the Skellam distribution with parameters $\lambda_1 = 2$ and $\lambda_2 = 0.2$ and initial values: $p_0 = 1.597585 \times 10^{-1}$, $p_1 = 2.689828 \times 10^{-1}$ (forward recursion) and $p_{11} = 5.877388 \times 10^{-6}$, $p_{10} = 3.242334 \times 10^{-5}$ (backward recursion).

n	p_n (forward)	relative error	n	p_n (backward)	relative error
2	2.526713×10^{-1}	$-4.393943 \times 10^{-16}$	9	1.627044×10^{-4}	2.332270×10^{-15}
3	1.631156×10^{-1}	1.020954×10^{-15}	8	7.354123×10^{-4}	1.326850×10^{-15}
4	7.997906×10^{-2}	$-4.737035 \times 10^{-14}$	7	2.957920×10^{-3}	2.785720×10^{-15}
5	3.157457×10^{-2}	2.437381×10^{-12}	6	1.042626×10^{-2}	2.662084×10^{-15}
6	1.042626×10^{-2}	$-1.882155 \times 10^{-10}$	5	3.157457×10^{-2}	2.417383×10^{-15}
7	2.957920×10^{-3}	2.016323×10^{-8}	4	7.997906×10^{-2}	3.123320×10^{-15}
8	7.354102×10^{-4}	-2.865149×10^{-6}	3	1.631156×10^{-1}	3.403179×10^{-15}
9	1.627893×10^{-4}	5.216764×10^{-4}	2	2.526713×10^{-1}	3.515154×10^{-15}
10	2.858271×10^{-5}	-1.184526×10^{-1}	1	2.689828×10^{-1}	3.921112×10^{-15}
11	1.987576×10^{-4}	3.281733×10	0	1.597585×10^{-1}	3.995894×10^{-15}

Table 5.5: Recursive evaluation (in both directions) of the pmf for the sum of Poisson and Binomial rv's with parameters $\lambda = 0.5$ and $v = 5$, $q = 0.7$, respectively, and initial values: $p_0 = 1.473870 \times 10^{-3}$, $p_1 = 1.793208 \times 10^{-2}$ (forward recursion) and $p_{11} = 2.569414 \times 10^{-6}$, $p_{10} = 3.158537 \times 10^{-5}$ (backward recursion).

n	p_n (forward)	relative error	n	p_n (backward)	relative error
2	8.902581×10^{-2}	7.794249×10^{-16}	9	3.265430×10^{-4}	$-1.660122 \times 10^{-15}$
3	2.295381×10^{-1}	7.255155×10^{-16}	8	2.743129×10^{-3}	$-1.580972 \times 10^{-15}$
4	3.224526×10^{-1}	1.205070×10^{-15}	7	1.780219×10^{-2}	$-1.753999 \times 10^{-15}$
5^*	2.362820×10^{-1}	1.292148×10^{-15}	6	8.238929×10^{-2}	$-1.852858 \times 10^{-15}$
6	8.238929×10^{-2}	1.010650×10^{-15}	5^*	2.362820×10^{-1}	$-2.114424 \times 10^{-15}$
7	1.780219×10^{-2}	2.338665×10^{-15}	4	3.224526×10^{-1}	$-1.377223 \times 10^{-15}$
8	2.743129×10^{-3}	$-2.529555 \times 10^{-15}$	3	2.295381×10^{-1}	$-5.562285 \times 10^{-15}$
9	3.265430×10^{-4}	3.287041×10^{-14}	2	8.902581×10^{-2}	4.224483×10^{-14}
10	3.158537×10^{-5}	$-3.252397 \times 10^{-13}$	1	1.793208×10^{-2}	$-1.526921 \times 10^{-12}$
11	2.569414×10^{-6}	4.528539×10^{-12}	0	1.473870×10^{-3}	1.608906×10^{-10}

Bibliografía

Acharya, V., Sanjiv, R., & K., R. (2002). Pricing credit derivatives with rating transitions. *Financial Analysts Journal*, 58(3):28–44.

Al-Osh, M. & Alzaid, A. (1987). First-order integer-valued autoregressive (INAR(1)) process. *Journal of Time Series Analysis*, 8(3):261–275.

Albert, P. (1991). A two-state markov mixture model for a time series of epileptic seizure counts. *Biometrics*, 47(4):1371–1381.

Altman, E. (1998). The importance and subtlety of credit rating migration. *Journal of Banking and Finance*, 22:1231–1247.

Altman, E. & Kao, D. (1992). The implications of corporate bond ratings drift. *Financial Analysts Journal*, 48(3):64–75.

Altman, R. (2007). Mixed hidden markov models: an extension of the hidden markov model to the longitudinal data setting. *Journal of the American Statistical Association*, 102(477):201–210.

Anderson, J. A. (1984). Regression and ordered categorical variables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(1):1–30.

Avery, J. P. & Henderson, A. D. (1999). Fitting markov chain models to discrete state

series such as dna sequences. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 48(1):53–61.

Baddeley, A. (1986). *Working memory*. Oxford: Oxford University Press.

Baddeley, A., Emslie, H., Kolodny, J., & Duncan, J. (1998). Random generation and the executive control of working memory. *Quarterly Journal of Experimental Psychology*, 51(4):819–852.

Baena-Mirabete, S., Espinal, A., & Puig, P. (2019). Exploring the randomness of mentally generated head–tail sequences. *Statistical Modelling*.

Baena-Mirabete, S. & Puig, P. (2018). Parsimonious higher-order markov models for rating transitions. *Journal of the Royal Statistical Society. Series A*, 181(1):107–131.

Baena-Mirabete, S. & Puig, P. (2020). Computing probabilities of integer-valued random variables by recurrence relations. *Statistics & Probability Letters*, 161.

Bangia, A., Diebold, F., Kronimus, A., Schagen, C., & Schuermann, T. (2002). Ratings migration and the business cycle, with application to credit portfolio stress testing. *Journal of Banking and Finance*, 26(2–3):445–474.

Bar-Hillel, M. & Wagenaar, W. A. (1991). The perception of randomness. *Advances in Applied Mathematics*, 12(4):428–454.

Barndorff-Nielsen, O. E., Pollard, D. G., & Shephard, N. (2012). Integer-valued lévy processes and low latency financial econometrics. *Quantitative Finance*, 12(4):587–605.

Bartolucci, F., Bacci, S., & Pennoni, F. (2014a). Longitudinal analysis of self-reported health status by mixture latent auto-regressive models. *Journal of the Royal Statistical Society. Series C*, 63(2):267–288.

Bartolucci, F. & Farcomeni, A. (2010). A note on the mixture transition distribution and hidden markov models. *Journal of Time Series Analysis*, 31(2):132–138.

Bartolucci, F., Farcomeni, A., & Pennoni, F. (2013). *Latent Markov Models for Longitudinal Data*. Chapman & Hall/CRC, Boca Raton.

Bartolucci, F., Farcomeni, A., & Pennoni, F. (2014b). Latent markov models: a review of a general framework for the analysis of longitudinal data with covariates. *TEST*, 23(3):433–465.

Bartolucci, F., Lupparelli, M., & Montanari, G. (2009). Latent markov model for longitudinal binary data: an application to the performance evaluation of nursing homes. *The Annals of Applied Statistics*, 3(2):611–636.

Bartolucci, F., Pandolfi, S., & Pennoni, F. (2017). Lmest: An r package for latent markov models for longitudinal categorical data. *Journal of Statistical Software*, 81(4):1–38.

Bartolucci, F. & Pennoni, F. (2007). A class of latent markov models for capture-recapture data allowing for time, heterogeneity, and behavior effects. *Biometrics*, 63(2):568–578.

Basel Committee on Banking Supervision (2011). *Basel III: A global regulatory framework for more resilient banks and banking systems*. Basel Committee on Banking Supervision, Basel, http://www.bis.org/bcbs/basel3.htm.

Bellotti, T. & Crook, J. (2014). Retail credit stress testing using a discrete hazard model with macroeconomic factors. *Journal of the Operational Research Society*, 65(3):340–350.

Bender, R. & Grouven, U. (1998). Using binary logistic regression models for ordinal data with non-proportional odds. *J Clin Epidemiol*, 51(10):809–816.

Bercedis, P. & Harrell, F. (1990). Partial proportional odds models for ordinal response variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 39(2):205–217.

Berchtold, A. & Raftery, A. (2002). The mixture transition distribution model for high-order markov chains and non-gaussian time series. *Statistical Science*, 17(3):328–356.

Boreiko, D. V., Kaniovski, S. Y., Kaniovski, Y. M., & Pflug, G. C. (2019). Identification of hidden markov chains governing dependent credit-rating migrations. *Communications in Statistics - Theory and Methods*, 48(1):75–87.

Breslow, N. E. & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25.

Brugger, P., Monsch, A. U., Salmon, D., & Butters, N. (1996). Random number generation in dementia of the alzheimer type: a test of frontal executive functions. *Neuropsychologia*, 34(2):97–103.

Budescu, D. V. (1985). Analysis of dichotomous variables in the presence of serial dependence. *Psychological Bulletin*, 97(3):547–561.

Budescu, D. V. (1987). A markov model for generation of random binary sequences. *Journal of Experimental Psychology: Human Perception & Performance*, 13(1):25–39.

Butler, K. & Stephens, M. A. (2017). The distribution of a sum of independent binomial random variables. *Methodology and Computing in Applied Probability*, 19(2):557–571.

Cornaggia, J., Cornaggia, K., & Hund, K. (2015). Credit ratings across asset classes: A long-term perspective. *Working paper*.

Cox, D. (1981). Statistical analysis of time series: some recent developments. *Scandinavian Journal of Statistics*, 8(2):93–115.

Crook, J. & Bellotti, T. (2010). Time varying and dynamic models for default risk in consumer loans. *Journal of the Royal Statistical Society: Series A*, 173(2):283–305.

Crouhy, M., Galai, D., & Mark, R. (2000). A comparative analysis of current credit risk models. *Journal of Banking and Finance*, 24:59–117.

Dayton, C. & Macready, G. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association*, 83(401):173–178.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood estimation from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

Dullmann, K., Scheicher, M., & Schmieder, C. (2007). Asset correlations and credit portfolio risk: an empirical analysis. *Discussion Paper Series 2: Banking and Financial Studies. Deutsche Bundesbank, Research Centre*, 13.

European Banking Authority (2014). *Methodological Note: EU-Wide Stress Test 2014*. EBA, http://www.eba.europa.eu.

European Central Bank (2007). *The use of portfolio credit risk models in central banks*. ECB, http://www.ecb.int.

Falk, R. & Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgment. *Psychological review*, 104(2):301–318.

Fernández-Fontelo, A., Fontdecaba, S., Alba, A., & Puig, P. (2017). Integer-valued AR processes with hermite innovations and time-varying parameters: An application to bovine fallen stock surveillance at a local scale. *Statistical Modelling*, 17(3):172–195.

Fienberg, S. E. (1980). The analysis of cross-classifed categorical data. *(2nd ed.), Cambridge, MA: MIT Press.*

Fitzmaurice, G., Davidian, M., Verbeke, G., & Molenberghs, G. (2008). *Longitudinal data analysis*. Chapman and Hall, CRC, London.

Fokianos, K. & Kedem, B. (2003). Regression theory for categorical time series. *Statistical Science*, 18(3):357–376.

Freeland, R. & McCabe, B. P. M. (2004). Analysis of low count time series data by poisson autoregression. *Journal of Time Series Analysis*, 25(5):701–722.

Fuertes, A., Fei, F., & Kalotychou, E. (2012). Credit rating migration risk and business cycles. *Journal of Business Finance and Accounting*, 39(1–2):229–263.

Fuertes, A. & Kalotychou, E. (2007). On sovereign credit migration: A study of alternative estimators and rating dynamics. *Computational Statistics and Data Analysis*, 51(7):3448–3469.

Furman, E. (2007). On the convolution of the negative binomial random variables. *Statistics & Probability Letters*, 77(2):169–172.

García-Viedma, M., Fernández-Guinea, S., Martos, R., & Ortega-Martínez, A. (2015). A random generation task with minimum memory requirements for mild alzheimer's disease. *Acción Psicológica*, 12(1):47–56.

Gaver, D. & O'Muircheartaigh, I. (1987). Robust empirical bayes analyses of event rates. *Technometrics*, 29(1):1–15.

Gibbons, R., Hedeker, D., Charles, S., & Frisch, P. (1994). A random-effects probit model for predicting medical malpractice claims. *Journal of the American Statistical Association*, 89(427):760–767.

Gómez-Déniz, E., Dávila-Cárdenes, N., & Pérez-Sánchez, J. (2019). A probabilistic model for explaining the points achieved by a team in football competition. forecasting and regression with applications to the spanish league. *SORT-Statistics and Operation Research Transactions*, 43(1):95–112.

Gupton, G., Finger, C., & Bhatia, M. (1997). Creditmetrics technical document. *J.P. Morgan*.

Guttler, A. & Raupach, P. (2008). The impact of downward rating momentum on credit portfolio risk. *Discussion Paper Series 2: Banking and Financial Studies. Deutsche Bundesbank, Research Centre*, 16.

Jackson, J., Albert, P., & Zhang, Z. (2015). A two-state mixed hidden markov model for risky teenage driving behaviour. *The Annals of Applied Statistics*, 9(2):849–865.

Jackson, J., Albert, P., Zhang, Z., & Simons-Morton, B. (2013). Ordinal latent variable models and their application in the study of newly licensed teenage drivers. *Journal of the Royal Statistical Society. Series C*, 62(3):435–450.

Jackson, P. & Perraudin, W. (2000). Regulatory implications of credit risk modelling. *Journal of Banking and Finance*, 24:1–14.

Kareev, Y. (1992). Not that bad after all. generation of random sequences. *Journal of Experimental Psychology: Human Perception & Performance*, 18(4):1189–1194.

Karlis, D. & Ntzoufras, I. (2003). Analysis of sports data by using bivariate poisson models. *The Statistician*, 52(3):381–393.

Katz, L. (1965). Unified treatment of abroad class of discrete probability distributions. *In: Classical and Contagious Discrete Distributions (Ed. G. P. Patil). Calcutta: Statistical Publishing Society; Oxford: Pergamon*, pages 175–182.

Katz, R. (1981). On some criteria for estimating the order of a markov chain. *Technometrics*, 23(3):243–249.

Kharin, Y. & Maltsau, M. (2014). Markov chain of conditional order: Properties and statistical analysis. *Austrian Journal of Statistics*, 43(3):205–216.

Kharin, Y. & Petlitskii, A. (2007). A markov chain of order s with r partial connections and statistical inference on its parameters. *Discrete Mathematics and Applications*, 17(3):295–317.

Kiesel, R., Perraudin, W., & Taylor, A. (2001). The structure of credit risk: spread volatility and rating transitions. *Working paper. Bank of England. Quarterly Bulletin 41.2*.

Kim, Y. & Young, S. (2008). Random effects model for credit rating transitions. *European Journal of Operational Research*, 184:561–573.

Kliger, D. & Sarig, O. (2000). The information value of bond ratings. *The Journal of Finance*, 55(6):2879–2902.

Koopman, S. J., Lit, R., & Lucas, A. (2017). Intraday stochastic volatility in discrete price changes: The dynamic skellam model. *Journal of the American Statistical Association*, 112(520):1490–1503.

Korolkiewicz, M. & Elliott, R. (2008). A hidden markov model of credit quality. *Journal of Economic Dynamics and Control*, 32(12):3807–3819.

Kotz, S., Johnson, N., & Kemp, A. (2005). *Univariate Discrete Distributions*. 3rd, John Wiley & Sons, Inc., New York.

Kuss, O. (2006). On the estimation of the stereotype regression model. *Computational Statistics and Data Analysis*, 50(8):1877–1890.

Lando, D. & Skϕdeberg, T. (2002). Analyzing rating transitions and rating drift with continuous observations. *Journal of Banking and Finance*, 26(2–3):423–444.

Lèbre, S. & Bourguignon, P. (2008). An em algorithm for estimation in the mixture transition distribution model. *Journal of Statistical Computation and Simulation*, 78(8):713–729.

Leroux, B. & Puterman, M. (1992). Maximum-penalized-likelihood estimation for independent and markov-dependent mixture models. *Biometrics*, 48(2):545–558.

Lewis, J. W., Brown, P. E., & Tsagris, M. (2016). *Densities and Sampling for the Skellam Distribution*. R package version 0.2.0.

Liu, G. (2012). Municipal bond insurance premium, credit rating and underlying credit risk. *Public Budgeting and Finance*, 32(1):128–156.

Lopes, L. L. & Oden, G. C. (1987). Distinguishing between random and nonrandom events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13(3):392–400.

Lu, Y. (2018). Exact likelihood estimation and probabilistic forecasting in higher-order INAR(p) models. *MPRA Paper 83682, Munich: University Library of Munich*.

Markov, A. A. (2006). An example of statistical investigation of the text eugene onegin concerning the connection of samples in chains. *Science in Context*, 19(4):591–600.

Masson, P. & Ruge-Murcia, F. (2005). Explaining the transition between exchange rate regimes. *The Scandinavian Journal of Economics*, 107(2):261–278.

McNeil, A. & Wendin, J. (2006). Dependent credit migrations. *Journal of Credit Risk*, 2(3):87–114.

Mersmann, O. (2019). *Microbenchmark: Accurate Timing Functions*. R package version 1.4-7.

Merton, R. (1974). On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance*, 29:449–470.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63(2):81–97.

Morris, C. (1983). Parametric empirical bayes inference: theory and applications. *Journal of the American Statistical Association*, 78(381):47–55.

Neuringer, A. (1986). Can people behave randomly?": The role of feedback. *Journal of Experimental Psychology: General*, 115(1):62–75.

Nickell, P., Perraudin, W., & Varotto, S. (2000). Stability of rating transitions. *Journal of Banking and Finance*, 24(1–2):203–227.

Nickerson, R. S. (2002). The production and perception of randomness. *Psychological review*, 109(2):330–357.

Ong, S. H. (1995). Some stochastic models leading to the convolution of two binomial variables. *Statistics & Probability Letters*, 22(2):161–166.

Palumbo, G. & Shick, R. (2006). Factors affecting a municipality's bond rating: an empirical study. *Journal of Business and Economics Research*, 4(11).

Panjer, H. (1981). Recursive evaluation of a family of compound distributions. *ASTIN Bulletin*, 12:22–26.

Panjer, H. & Wang, S. (1993). On the stability of recursive formulas. *ASTIN Bulletin*, 23:227–258.

Park, J. & Basawa, I. (2002). Estimation for mixtures of markov processes. *Statistics & Probability Letters*, 59(3):235–244.

Pedeli, X., Davison, A. C., & Fokianos, K. (2015). Likelihood estimation for the INAR(p) model by saddlepoint approximation. *Journal Of The American Statistical Association*, 110(511):1229–1238.

Petrie, T. (1969). Probabilistic functions of finite state markov chains. *The Annals of Mathematical Statistics*, 40(1):97–115.

Pinheiro, J. & Bates, D. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, 4(1):12–35.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raftery, A. (1985). A model for high-order markov chains. *Journal of the Royal Statistical Society, Series B*, 47(3):528–539.

Rao, J., Scott, A., & Skinner, C. (1998). Quasi-score tests with survey data. *Statistica Sinica*, 8:1059–1070.

Rapoport, A. & Budescu, D. V. (1997). Randomization in individual choice behavior. *Psychological Review*, 104(3):603–617.

Reichenbach, H. (1949). The theory of probability: An inquiry into the logical and mathematical foundations of the calculus of probability. *Berkeley: University of California Press*.

Richard, F. (2011). Measuring perfomance in us municipalities: Do personnel policies predict system level outcomes? *Journal of Comparative Policy Analysis*, 13(1):11–33.

Rosen, O., Jiang, W., & Tanner, M. (2000). Mixtures of marginal models. *Biometrika*, 87(2):391–404.

Saaty, T. L. & Ozdemir, M. S. (2003). Why the magic number seven plus or minus two. *Mathematical and Computer Modelling*, 38(3-4):233–244.

Salamé, P. & Danion, J. (2007). Inhibition of inappropriate responses is preserved in the think-no-think and impaired in the random number generation tasks in schizophrenia. *Journal of the International Neurospychological Society*, 13(2):277–287.

Salazar, J., Schmitt, F., Yu, L., Mendiondo, M., & Kryscio, R. (2007). Shared random effects analysis of multi-state markov models: application to a longitudinal study of transitions to dementia. *Statistics in Medicine*, 26(3):568–580.

Sarmiento, C. (2006). Statistical analysis of municipal bond ratings under spatial correlation. *Applied Financial Economics Letters*, 2(4):233–237.

SAS Institute (2011). *SAS/STAT Software, Version 9.3*. Cary: SAS Institute.

Skrondal, A. & Rabe-Hesketh, S. (2009). Prediction in multilevel generalized linear models. *Journal of the Royal Statistical Society. Series A*, 172(3):659–687.

Slud, E. & Kedem, B. (1994). Partial likelihood analysis of logistic regression and autoregression. *Statistica Sinica*, 4(1):89–106.

Song, Q. & Smith, A. D. (2011). Identifying dispersed epigenomic domains from chip-seq data. *Bioinformatics*, 27(6):870–871.

Spatt, J. & Goldenberg, G. (1993). Components of random generation by normal subjects and patients with dysexecutive syndrome. *Brain and Cognition*, 23(2):231–242.

Standard and Poor's (2013). U.S. Local Governments General Obligation Ratings: Methodology and Assumptions.

Standard and Poor's (2014). U.S. Public Finance Defaults and Rating Transition Data: 2014 Update.

Stefanescu, C., Tunaru, R., & Turnbull, S. (2009). The credit rating process and estimation of transition probabilities: a bayesian approach. *Journal of Empirical Finance*, 16(2):216–234.

Stuart, R., Garrett, M., Scott, E., Debajyoti, S., Joseph, G., & Atul, A. (2013). Bias correction for the proportional odds logistic regression model with application to a study of surgical complications. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 62(2):233–250.

Sundt, B. (1992). On some extensions of Panjer's class of counting distributions. *ASTIN Bulletin*, 22(1):61–80.

Thijs, H., Molenberghs, G., Michiels, B., Verbeke, G., & Curran, D. (2002). Strategies to fit pattern-mixture models. *Biostatistics*, 3(2):245–265.

Towse, J. (1998). On random generation and the central executive of working memory. *The British Journal of Psychology*, 89(1):77–101.

Van de Pol, F. & Langeheine, R. (1990). Mixed markov latent class models. *Sociological Methodology*, 20:213–247.

Van der Linden, M., Beerten, A., & Pesenti, M. (1998). Age-related differences in random generation. *Brain and Cognition*, 38(1):1–16.

Vasicek, O. (2002). The distribution of loan portfolio value. *Risk*, 15(December):160–162.

Vermunt, J. (2010). *Longitudinal Research Using Mixture Models. In: van Montfort K., Oud J., Satorra A. (eds). Longitudinal Research with Latent Variables*. Springer, Berlin, Heidelberg.

Vermunt, J., Langeheine, R., & Bockenholt, U. (1999). Discrete-time discrete-state latent markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics*, 24(2):179–207.

Wiggins, L. (1973). *Panel Analysis: Latent probability models for attitude and behavior processes*. Elsevier, Amsterdam.

Wozabal, D. & Hochreiter, R. (2012). A coupled markov chain approach to credit risk modeling. *Journal of Economic Dynamics and Control*, 36(3):403–415.

Yee, T. (2010). The vgam package for categorical data analysis. *Journal of Statistical Software*, 32(10):1–34.

Yee, T. & Hastie, T. (2003). Reduced-rank vector generalized linear models. *Statistical Modelling*, 3:15–41.

Zeger, S. & Qaqish, B. (1988). Markov regression models for time series: A quasi-likelihood approach. *Biometrics*, 44(4):1019–1031.

Zhu, R. & Joe, H. (2010). Negative binomial time series models based on expectation thinning operators. *Journal of Statistical Planning and Inference*, 140(7):1874–1888.