Capítulo 4

Análisis Bayesiano de sistemas de colas con múltiples servidores. Casos reales y problemas de diseño óptimo.

En este Capítulo, se desarrolla inferencia y predicción Bayesiana en sistemas de colas con varios servidores. Se proponen también métodos Bayesianos para abordar problemas de diseño óptimo en los que el número de servidores es la variable de control. El punto de partida en este Capítulo para ilustrar los modelos a estudiar y los problemas a resolver son dos situaciones reales concretas que se ubican en un hospital geriátrico de Londres y en un establecimiento bancario de Madrid. Para hacer más amena la lectura de este Capítulo, se plantean directamente los procedimientos típicos de la estimación Bayesiana así como los problemas clásicos en el diseño óptimo de sistemas de colas sobre los modelos a los que se pueden adaptar los dos casos reales anteriormente mencionados, resolviéndose directamente cuestiones de interés como por ejemplo, la estimación del número de clientes presentes en el sistema o la elección del número óptimo de servidores. No obstante, los métodos que se exponen en este Capítulo no están restringidos a estos ejemplos y pueden aplicarse en otras situaciones equivalentes.

En este Capítulo, se consideran dos modelos de colas con varios servidores que permiten analizar cada una de las dos situaciones reales consideradas. Concretamente, el sistema M/G/c/c y el sistema GI/M/c. Siguiendo la notación introducida en el Capítulo 1, el primer sistema de colas se caracteriza por un proceso de llegadas de Poisson, distribución general de servicio y c servidores. Además, este modelo es un caso particular de los denominados sistemas de pérdida (*loss systems*), que son aquellos sistemas con capacidad finita en la ocupación del sistema y que en este caso coincide con el número de servidores. Recuérdese que la capacidad de un sistema es el número máximo posible de clientes en el mismo, véase la Sección 1.1.1. Consecuentemente, un cliente que llega a un sistema M/G/c/c y encuentra todos los servidores ocupados abandona el mismo sin posibilidad de ser servido. El análisis Bayesiano desarrollado para este sistema se ilustra con un conjunto de datos reales sobre la duración de los ingresos en un hospital geriátrico de Londres. Con objeto de estudiar el comportamiento de la ocupación de camas en el hospital, se considera un sistema de colas M/G/c/c donde se identifican las camas con los servidores del sistema y la duración de la estancia de cada uno de los enfermos con el tiempo de servicio que recibe cada cliente en el sistema de colas. Además, cuando no hay camas disponibles, no se produce una línea de espera, ya que los pacientes abandonan el hospital siendo admitidos en otros centros hospitalarios. El otro modelo de colas a estudiar en este Capítulo es el sistema GI/M/c que tiene capacidad infinita y por tanto, los clientes que encuentran todos los servidores ocupados permanecen en el sistema formando una línea de espera y abandonan el mismo sólo cuando han sido servidos. Los tiempos entre las llegadas a este sistema constituyen una sucesión de variables aleatorias independientes

e idénticamente distribuidas según una distribución general desconocida. Los tiempos de servicio son también independientes y se distribuyen exponencialmente. La motivación para el análisis Bayesiano de este sistema es otro conjunto de datos reales recogidos en una oficina de una entidad bancaria en Madrid y que modelan el proceso de llegadas y servicios en dicha entidad.

Además, en este Capítulo, se formulan para cada sistema funciones de coste que permiten abordar el diseño del mismo y decidir, en cada caso, el número óptimo de servidores. Se consideran, en todos los casos, estructuras clásicas de coste lineal evaluadas en el estado estacionario. Es decir, las funciones de coste dependerán linealmente de los valores esperados de las distribuciones estacionarias de las características del sistema de colas, o equivalentemente, dependerán del valor que tomen estas medidas por unidad de tiempo (u.t.). Por ejemplo, las funciones de coste que se construyen en este Capítulo dependen del número medio de servidores ocupados por unidad de tiempo, entre otras cantidades. Se asume, por tanto, que el problema es de *horizonte infinito*, es decir, se evalúa el coste en el límite, en el estado estacionario y, consecuentemente, la función de objetivo representa el valor del coste medio total por unidad de tiempo. Alternativamente, se podrían establecer otros horizontes de planificación utilizando, por ejemplo, el criterio del ciclo, véase Lillo (2000c), en el que se evalúa el coste medio por ciclo de ocupación y que aunque en algunos casos, confluyen en la misma función de coste a optimizar, en otros varía tanto la filosofía subyacente en el control del sistema como la función de coste y por tanto, los óptimos.

Por otro lado, las funciones de coste que se construyen en este Capítulo pretenden establecer un equilibrio entre los intereses del diseñador del sistema o personal encargado de decidir el número apropiado de servidores, y los intereses de los clientes. Por este motivo, siempre se consideran dos clases diferentes de costes en el sistema de colas. Se distinguen, por una parte, los gastos originados por la actividad desarrollada por los servidores y que, obviamente, están asociados a los intereses del diseñador del sistema. En este grupo se engloban los costes relacionados con el número de servidores ocupados y desocupados y con la tasa de clientes que reciben su servicio. El otro grupo de costes representan los intereses de los clientes e incluyen los costes originados, fundamentalmente, por la espera de los clientes y en sistemas con capacidad finita, penalizan los clientes que no pueden recibir el servicio.

El diseño y control óptimo de sistemas de colas es un área de investigación con una literatura muy extensa, véase una recopilación, por ejemplo, en Kitaev y Rykov (1995). Sin embargo, como la mayoría de estos trabajos se enmarcan dentro del área de Investigación Operativa se supone, generalmente, que no hay incertidumbre en los procesos de llegadas y de servicio, asumiéndose conocidos los parámetros del sistema. En la práctica, si el diseñador del sistema no conoce el verdadero valor de los parámetros, se enfrenta con la dificultad de encontrar una estimación para los mismos antes de resolver el problema de la optimización. En este Capítulo, se muestra cómo la metodología Bayesiana permite incorporar de forma natural la incertidumbre de las estimaciones en las funciones de coste directamente.

Aunque, como ya se ha comentado, la inferencia Bayesiana en sistemas de colas es, actualmente, un área bastante desarrollada, muy pocos estudios se han dedicado al análisis estadístico del diseño de estos sistemas. En Bagchi y Cunningham (1972), que es uno de los primeros trabajos de estimación Bayesiana para modelos de colas, se desarrolla un procedimiento de diseño óptimo que permite seleccionar la mejor tasa de servicio y la capacidad máxima de un sistema con un único servidor. Armero y Bayarri (1996) examinan numerosos criterios cuando se aborda el problema de la decisión del número de servidores en un sistema M/M/c y Wiper (1998) analiza algunos de estos criterios en el modelo Er/M/c. Sin embargo, en ningún caso, se especifica una expresión cerrada para la estructura del coste que permita tomar decisiones englobando todos los objetivos que se hayan preestablecido y como la solución a un problema de optimización. Este hecho, motiva la construcción de una función de coste medio que dependa de las medidas estacionarias de interés, como el tiempo de espera en el sistema, número de clientes ocupados,...

Este Capítulo está dividido en tres Secciones. En la Sección 4.1, se desarrollan métodos Bayesianos para el análisis y el diseño del sistema M/G/c/c. Estos procedimientos se ilustran, a lo largo de toda la Sección, analizando el problema de la ocupación de camas en un hospital geriátrico. En la Sección 4.2, se proponen

métodos Bayesianos para estimar las características del modelo GI/M/c y decidir el número óptimo de servidores en este sistema. En este caso, los procedimientos desarrollados se aplican sobre el sistema de colas observado en una sucursal bancaria. Estas dos primeras Secciones están divididas a su vez en tres Subsecciones que incluyen, en primer lugar, la presentación de los datos y la estimación de las distribuciones de interés haciendo uso de las técnicas introducidas en el Capítulo 2. En segundo lugar, la estimación de las características del sistema de colas observado, y por último, el diseño del mismo. Finalmente, en la Sección 4.3, se incluyen algunos comentarios y extensiones.

4.1. Análisis del sistema de colas M/G/c/c. Aplicaciones sobre la ocupación de camas en hospitales.

En esta Sección, se proponen métodos Bayesianos para la inferencia y el diseño del sistema de colas M/G/c/c. El análisis de este sistema se ilustra y se motiva examinando la duración de los ingresos hospitalarios en pacientes geriátricos y resolviendo algunos problemas de gestión sanitaria. En la Subsección 4.1.1, se describe el conjunto de datos reales disponible, que consiste en los tiempos que permanecen un grupo de pacientes en un hospital. La distribución general de servicio se aproxima utilizando el modelo de mixtura MGE. En la Subsección 4.1.2, se describen y se estiman las características del sistema M/G/c/c. El procedimiento desarrollado sobre este sistema de colas se ilustra estimando algunas medidas de interés para el hospital, como la distribución del número de camas ocupadas. En la Subsección 4.1.3, se formula una función de coste medio por unidad de tiempo dependiendo del número de servidores del sistema M/G/c/c, que se aplica después para decidir el número óptimo de camas que debe ofertar el hospital.

4.1.1. Descripción de los datos y estimación de su distribución utilizando el modelo MGE.

La distribución del tiempo que permanecen los pacientes en hospitales geriátricos tiene aparentemente un comportamiento muy complejo. Inicialmente, los enfermos suelen recibir un cuidado intensivo que, en general, va seguido de varias etapas diferentes de tratamiento. La mayoría de los pacientes reciben el alta tras un breve espacio de tiempo. Sin embargo, algunos enfermos deben permanecer en el hospital durante muchos meses, incluso años, recibiendo una atención continuada. Este hecho provoca que la distribución de la duración de la estancia en este tipo de hospitales sea muy heterogénea.

La Figura 4.1 muestra diagramas de caja de la distribución de la duración de los tiempos de ingreso de 1092 pacientes geriátricos en el Hospital St. George de Londres en el transcurso de 1965 hasta 1984. Los datos son un subconjunto de los analizados en Taylor et al. (2000) y se pueden obtener en la siguiente dirección de Internet, `http://www.blackwellpublishers.co.uk/rss/`. Los diagramas de caja muestran que los datos son muy asimétricos a la derecha con un número muy elevado de atípicos.

Puesto que los pacientes atraviesan varias etapas de tratamiento, las distribuciones de tipo PH constituyen modelos apropiados para describir la estancia de enfermos en el hospital ya que, como se comentó en el Capítulo 3, estas distribuciones asumen, esencialmente, que cada observación se puede descomponer en diferentes fases exponenciales. Algunos autores han utilizado técnicas clásicas para ajustar distintos modelos de distribuciones PH a los datos del Hospital de St. George. Por ejemplo, en Harrison y Millard (1991) y Gorenescu et al. (1999), se consideran mixturas de exponenciales y en Faddy y McClean (1999), se utiliza el modelo de distribución MGE aunque, en este caso, se supone que el número de fases es un valor fijo. Una de las dificultades principales en estos métodos es que no están diseñados para enfrentarse a situaciones en las que la dimensión del espacio paramétrico sea desconocida, como sucede con el número de fases de la distribución MGE. Sin embargo, la inferencia Bayesiana permite abordar este tipo de problemas haciendo uso de los métodos MCMC de dimensión variable, como son las técnicas de salto reversible o los métodos

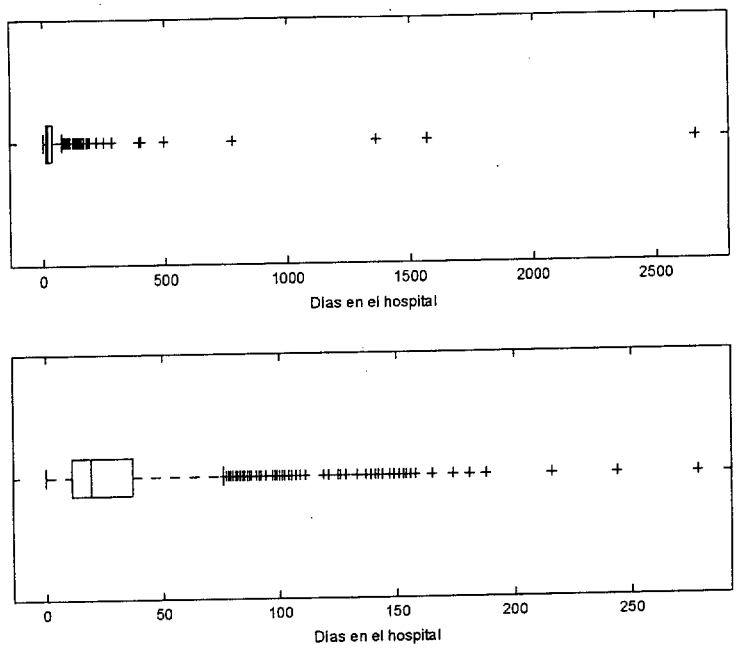

Figura 4.1: Diagramas de caja del número de días que permanecen los enfermos en el hospital (arriba) y el mismo diagrama para aquellos cuya estancia es inferior a 300 días (abajo).

BDMCMC, que se utilizaron para construir los algoritmos del Capítulo 2, y que se aplicarán también en este Capítulo.

En esta Sección, se opta por el modelo de distribución MGE entre los dos modelos de distribución PH que se han propuesto en esta tesis, para modelar la distribución subyacente en el conjunto de datos considerado. La distribución MGE resulta aparentemente bastante adecuada ya que supone que el tiempo total que permanece un enfermo en el hospital se descompone en una sucesión de fases exponenciales consecutivas, de modo que, cada paciente puede abandonar el hospital en cualquiera de estas fases con cierta probabilidad, véase la Figura 2.2. Además, en la Sección 2.5 del Capítulo 2, se introdujo este conjunto de datos y se observó que el modelo de mixtura HEr presenta, en este caso, algunos problemas de sobreajuste incluyendo un número elevado de componentes en la mixtura.

La Figura 4.2 ilustra la función de densidad predictiva del número de días que permanece un paciente en el hospital junto al histograma de los datos truncando en los valores inferiores a 300 días. La función de densidad estimada se obtiene utilizando el algoritmo `RJMGE` propuesto en el Capítulo 2, donde se describe detalladamente el procedimiento de estimación, véase la Figura 2.13 en la Sección 2.5, en la que se compara esta densidad estimada con la obtenida con el método de Faddy y McClean (1999). La estimación del número medio de días que permanece un paciente en el hospital resulta ser aproximadamente igual a 37.3 días.

La Tabla 4.1 muestra algunos de los cuantiles de la distribución estimada que permiten examinar su comportamiento por encima de los 300 días. Se puede observar que la distribución presenta una cola asimétrica pesada a la derecha. Por ejemplo, nótese que, aunque la mitad de los pacientes reciben el alta antes de 20 días, el 3% de los pacientes permanece más de 5 meses en el hospital. Además, aunque la probabilidad de que el tiempo de ingreso sea de uno o más años es pequeña, es importante considerarla a la hora de planificar el diseño del hospital, como se verá en la Sección 4.1.3.

Con el fin de estimar algunas medidas de interés en el hospital, en la siguiente Subsección, se asocia el conjunto de observaciones sobre los ingresos hospitalarios que acaba de analizarse con una muestra de datos

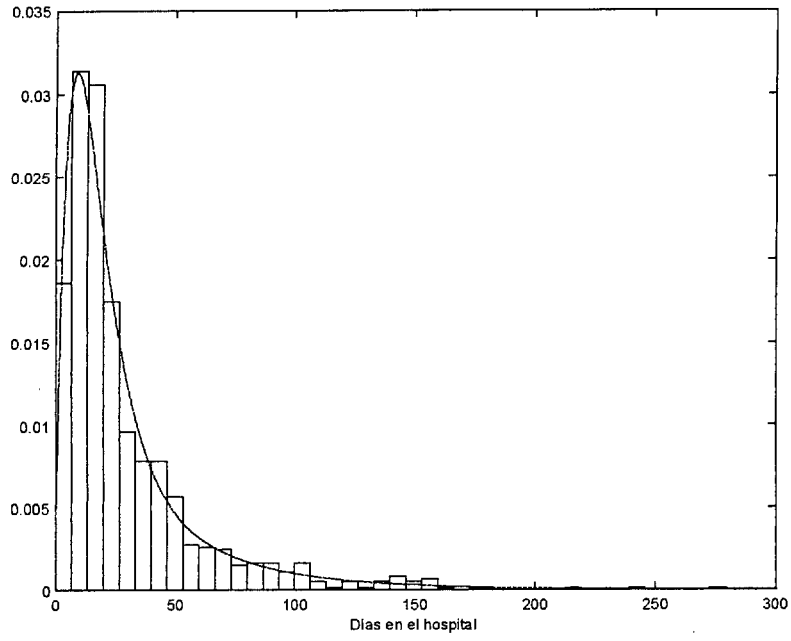

Figura 4.2: Función de densidad estimada del tiempo de permanencia de los pacientes en el hospital e histograma de los datos truncando en valores inferiores a 300 días.

Orden	0.25	0.50	0.75	0.8	0.9	0.95	0.97	0.98	0.99	0.995	.999
Cuantil	10.27	19.55	37.38	44.55	72.75	110.55	151.35	221.40	536.22	1200.47	1900.21

Tabla 4.1: Algunos cuantiles de la distribución del número de días que permanecen los enfermos en el hospital.

de tiempos de servicio en un sistema de colas cuyos servidores se corresponden con las camas del hospital. Lamentablemente, no se dispone de información real sobre las llegadas de los pacientes al Hospital de St. George, pero se describirá como desarrollar inferencia Bayesiana suponiendo que se tiene no sólo una muestra de la distribución a posteriori de los parámetros del tiempo de servicio, sino también información sobre la distribución a posteriori de los parámetros que definen el proceso de llegadas al sistema.

4.1.2. Sistema M/G/c/c y estimación de sus características.

En esta Subsección, se desarrolla inferencia y predicción Bayesiana en el modelo de colas M/G/c/c. A continuación, se utilizan los procedimientos propuestos, junto con los resultados de la Subsección anterior, para obtener estimaciones de diferentes medidas de interés en el centro hospitalario. Para ello, se considera, en particular, el sistema de colas M/MGE/c/c, en el que el tiempo de servicio se asocia con el número de días que permanece cada paciente en el hospital y los c servidores del sistema con las c camas que oferta el hospital.

4.1.2.1. Propiedades e inferencia Bayesiana para el sistema M/G/c/c.

Se considera un sistema de colas M/G/c/c con parámetros conocidos, $\theta = \{\lambda, \theta_\mu\}$, donde λ es la tasa del proceso de llegadas Poisson y θ_μ representa los parámetros de la distribución general de servicio. Tal como

se indicó al principio de este Capítulo, este sistema tiene capacidad finita e igual al número de servidores, c. Con estas condiciones, como se indicó en (1.18), la distribución estacionaria del número de servidores ocupados, N_b, viene dada por, véase, por ejemplo, Tijms (1990),

$$P(N_b = n) = \frac{\rho^n/n!}{\sum_{k=0}^{c} \rho^k/k!}, \qquad n = 0, ..., c, \tag{4.1}$$

donde ρ es la intensidad de tráfico, véase (1.5),

$$\rho = \lambda E[S \mid \theta_\mu] \tag{4.2}$$

y donde S es la variable aleatoria que representa el tiempo de servicio en el sistema. Es importante puntualizar que, en sistemas con capacidad finita, la ergodicidad está siempre asegurada, véase, por ejemplo, Kleinrock (1975) y Lillo (2000e) con condiciones muy suaves, sobre los soportes de las distribuciones entre llegadas y de servicio. Por tanto, no es necesario que los parámetros del sistema verifiquen la condición de equilibrio propia de los sistemas con capacidad infinita, $\rho < c$, para que existan las distribuciones estacionarias. Consecuentemente, se puede calcular la distribución de N_b, (4.1), sea cual sea el valor de la intensidad de tráfico, ρ, (4.2), aunque, evidentemente, si ρ es mucho mayor que c, la probabilidad de que todos los servidores estén ocupados será muy próxima a uno. Por otro lado, se puede comprobar fácilmente, que la intensidad de tráfico, ρ, coincide con el número medio de llegadas durante un periodo de servicio. Esta característica se verifica en cualquier modelo de colas con proceso de llegadas Poisson y por eso, en estos casos, la intensidad de tráfico, ρ, recibe el nombre de carga ofrecida por el sistema (*offered load*). Obsérvese también que (4.1) es una distribución de Poisson de parámetro ρ truncada en el intervalo $[0, c]$. En particular, la probabilidad de que un cliente encuentre todos los servidores ocupados (*blocking probability*) es,

$$B(c, \rho) = \frac{\rho^c/c!}{\sum_{k=0}^{c} \rho^k/k!}. \tag{4.3}$$

Esta expresión se conoce como la fórmula de Erlang B (*Erlang's loss formula*) y equivale a la proporción de clientes que no pueden incorporarse al sistema porque todos los servidores están ocupados. Además, el número medio de servidores ocupados resulta ser igual a, véase, por ejemplo, Tijms (1990),

$$E[N_b] = \rho[1 - B(c, \rho)]. \tag{4.4}$$

Para obtener expresiones explícitas de estas cantidades, se puede aproximar la distribución general de servicio del sistema M/G/c/c con el modelo de mixtura MGE de parámetros, véase (2.4),

$$\theta_\mu = \{L, \mathbf{P} = (P_1, ..., P_L), \boldsymbol{\mu} = (\mu_1, ..., \mu_L)\},$$

y considerar el sistema M/MGE/c/c. En este caso, la carga ofrecida dada en (4.2) es, véase (3.15),

$$\rho = \lambda \sum_{r=1}^{L} \frac{1}{\mu_r} \left(1 - \textstyle\sum_{s=1}^{r-1} P_s\right). \tag{4.5}$$

Alternativamente, se puede aproximar la distribución general de servicio con una mixtura HEr de parámetros, véase (2.1),

$$\theta_\mu = \{k, \mathbf{w} = (w_1, ..., w_k), \boldsymbol{\mu} = (\mu_1, ..., \mu_k), \boldsymbol{\nu} = (\nu_1, ..., \nu_k)\},$$

y considerar entonces el sistema M/HEr/c/c cuya carga ofrecida es, véase (3.13),

$$\rho = \lambda \sum_{r=1}^{k} \frac{w_r}{\mu_r}. \tag{4.6}$$

Teniendo como datos las observaciones del proceso de servicio, $\mathbf{s}$, y de llegadas, $\mathbf{t}$, a un sistema M/G/c/c, se pueden estimar las características que se acaban de describir utilizando una muestra MCMC de la distribución a posteriori de los parámetros del sistema, $\{\lambda, \theta_\mu\}$. Se puede seguir un procedimiento análogo al utilizado en la Sección 3.3.2 para estimar las propiedades del modelo de colas M/G/1. Por ejemplo, si se ha aproximado la distribución general de servicio con una mixtura MGE, se obtiene una estimación de la carga ofrecida por el sistema calculando la media de los valores de (4.5) para el conjunto de valores de la muestra MCMC, es decir,

$$E[\rho \mid \mathbf{t}, \mathbf{s}] \approx \frac{1}{J} \sum_{j=1}^{J} \lambda^{(j)} \sum_{r=1}^{L^{(j)}} \frac{1}{\mu_r^{(j)}} \left(1 - \textstyle\sum_{s=1}^{r-1} P_s^{(j)}\right), \tag{4.7}$$

donde J es el tamaño de la muestra MCMC. Análogamente, se puede obtener una estimación para ρ si se aproxima la distribución general de servicio con una mixtura HEr.

Como se ha comentado anteriormente, en el sistema M/G/c/c, la ergodicidad está siempre asegurada y no es necesario asumir una condición de equilibrio para estimar las distribuciones estacionarias. Consecuentemente, en este caso, no es tan relevante estimar la probabilidad, $P(\rho < c \mid \mathbf{t}, \mathbf{s})$, aunque puede ser una medida interesante de la congestión del sistema. Una aproximación de esta probabilidad se obtiene, análogamente a (3.31), evaluando, para cada valor del número de servidores, c, la proporción de valores de la muestra MCMC que verifican $\rho < c$.

Para un determinado valor del número de servidores, c, la probabilidad predictiva de que existan n clientes ocupados se puede estimar con,

$$P(N_b = n \mid \mathbf{t}, \mathbf{s}) \approx \frac{1}{J} \sum_{j=1}^{J} \frac{\rho_{(j)}^n / n!}{\sum_{k=0}^{c} \rho_{(j)}^k / k!}, \tag{4.8}$$

donde $\rho_{(j)}$ es la carga ofrecida por el sistema para cada valor de los parámetros de la muestra MCMC. Obsérvese que, en este caso, se considera la muestra completa de tamaño J de la distribución a posteriori de los parámetros del sistema y no se restringe la muestra a los valores que verifiquen una condición de equilibrio, como era necesario, por ejemplo, en (4.8).

4.1.2.2. Modelo de colas M/MGE/c/c para la ocupación de camas en el hospital.

En este apartado, se considera un sistema de colas M/MGE/c/c para analizar la ocupación de camas en el Hospital de St. George. Se utilizan los resultados que se acaban de describir en la Sección anterior y la muestra MCMC obtenida en la Sección 4.1.1, donde se consideró la mixtura MGE para aproximar la distribución general del tiempo de ingreso en el hospital. Lamentablemente, no se dispone de información real sobre las llegadas de los pacientes al Hospital de St. George. Por tanto, no es posible generar una muestra de la distribución a posteriori de la tasa λ del proceso de Poisson con el que se ha asumido que se producen las llegadas al centro hospitalario. Así que, por simplicidad, se supone, en esta situación, que λ es conocida.

Si se supone, por ejemplo, que $\lambda = 1.5$, lo que equivale aproximadamente a los valores utilizados en Gorenescu et al. (1999), la estimación del número medio de pacientes que llegan durante un ingreso medio, obtenida a partir de (4.7), es $E[\rho \mid \mathbf{t}, \mathbf{s}] = 55.95$ pacientes. Obsérvese que esta estimación se obtiene multiplicando el valor de λ por el número medio de días en el hospital, obtenido en la Sección 4.1.1, que es aproximadamente igual a 37.3 días.

La Figura 4.3 muestra la estimación de la distribución del número de camas ocupadas, N_b, utilizando (4.8), para los datos del Hospital de St. George, considerando varios valores para el número de servidores o camas, c. Obsérvese que a medida que aumenta el número de camas en el hospital, la distribución estimada es más simétrica ya que, cuando c es mucho más grande que ρ, la distribución de N_b, dada en (4.1), se aproxima a una Poisson de parámetro ρ.

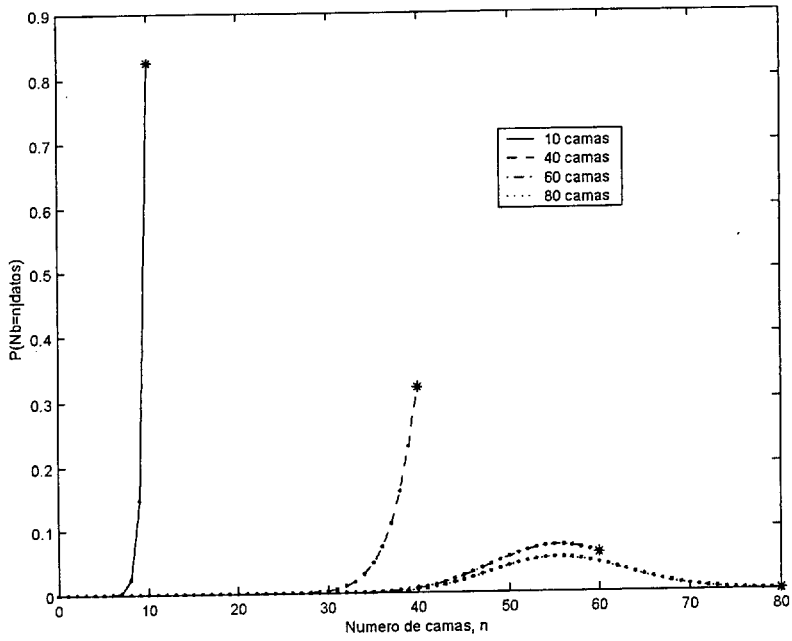

Figura 4.3: Estimación de la probabilidad de que haya n camas ocupadas, $P\left(N_b = n \mid \text{datos}\right)$, para distintos valores del número de camas, c, en el hospital. Se representa con un asterisco la probabilidad de que todas las camas estén ocupadas.

La Tabla 4.2 muestra la estimación de la probabilidad de que todas las camas estén ocupadas, $B(c,\rho)$, para varios valores del número de camas en el hospital. Estas probabilidades se han indicado también en la Figura 4.3 con un asterisco. Como era de esperar, esta probabilidad disminuye a medida que aumenta el número de camas, c. Obsérvese además que las estimaciones de $B\left(c,\rho\right)$ equivalen a la proporción de pacientes que no pueden ingresar en el hospital debido a la insuficiencia de camas. La Tabla 4.2 muestra también las estimaciones del número medio de camas ocupadas en cada caso.

c	10	40	60	80
$B\left(c,\rho \mid \mathbf{s},\mathbf{t}\right)$	0.8249	0.3188	0.06079	4.6873×10^{-4}
$E\left[N_b \mid \mathbf{s},\mathbf{t}\right]$	9.7654	37.8452	51.443	55.0144

Tabla 4.2: Estimaciones de la probabilidad de bloqueo, $B\left(c,\rho\right)$, que representa la probabilidad de que todas las camas estén ocupadas.

4.1.3. Función de coste y diseño óptimo del modelo.

El objetivo, en esta Subsección, es resolver el problema de la decisión del número óptimo servidores en el sistema M/G/c/c, que se aplicará más adelante en el diseño hospital. Para ello, se formula una función de coste con las características indicadas al principio de este Capítulo. Es decir, una estructura de coste lineal, definido en el estado estacionario y que depende de las características del sistema evaluadas por término medio y por unidad de tiempo. Una vez que se formula la función de coste y dadas las observaciones del proceso de llegadas y de servicio, se puede estimar el coste medio por unidad de tiempo utilizando las muestras MCMC de la distribución a posteriori de los parámetros del sistema. Muchas de las medidas introducidas

en la Sección 4.1.2.1 intervienen en las funciones de coste que se formulan a continuación y por tanto, los procedimientos de estimación allí descritos se utilizarán en esta Sección para estimar el coste medio por unidad de tiempo.

En concreto, se consideran cuatro costes posibles. En primer lugar, se supone un coste, r_b, por servidor ocupado por u.t. La distribución estacionaria del número de servidores ocupados, N_b, viene dada por (4.1) y su valor medio por (4.4). Por tanto, el coste medio por u.t. causado por los clientes ocupados es,

$$r_b E[N_b] = r_b E[N_b] = r_b \{\rho [1 - B(c, \rho)]\}. \tag{4.9}$$

En segundo lugar, se supone un coste, r_e, por cada servidor que permanece vacío por u.t. El número de servidores vacíos, N_e, es igual al número total de servidores, c, menos el número de servidores ocupados, N_b. Entonces, el coste medio por u.t. originado por los servidores desocupados es,

$$r_e E[N_e] = r_e \{c - E[N_b]\} = r_e \{c - \rho [1 - B(c, \rho)]\} \tag{4.10}$$

En tercer lugar, se asume un coste, r_s, por cliente servido. Nótese que, generalmente, el valor de r_s será negativo para reflejar un beneficio. El número de clientes servidos por u.t., N_s, es igual al número de servidores ocupados por el número clientes que atiende cada uno de estos servidores por u.t. Puesto que la tasa de clientes que atiende un servidor ocupado es la inversa del tiempo medio de servicio, $E[S \mid \boldsymbol{\theta}_\mu]^{-1}$, el coste o beneficio por cliente servido es igual a,

$$r_s E[N_s] = \frac{r_s E[N_b]}{E[S \mid \boldsymbol{\theta}_\mu]} = r_s \left\{ \frac{\rho [1 - B(c, \rho)]}{E[S \mid \boldsymbol{\theta}_\mu]} \right\} = r_s \{\lambda [1 - B(c, \rho)]\}. \tag{4.11}$$

Obsérvese que el número medio de clientes servidos por u.t., $E[N_s]$, es igual a la tasa de llegadas, λ, por la proporción de clientes que pueden ser atendidos, $[1 - B(c, \rho)]$.

Por último, se considera un coste, r_l, por cada cliente que no puede ingresar en el sistema porque todos los servidores están ocupados. El número de clientes que no pueden ser atendidos por u.t., N_l, es igual al número de clientes que llegan en una unidad de tiempo por la probabilidad de no ser admitido, que viene dada por la Fórmula de Erlang, $B(c, \rho)$. Por tanto, la demanda media perdida es, $E[N_l] = \lambda B(c, \rho)$ y el coste medio por u.t. originado por la demanda pérdida viene dado por,

$$r_l E[N_l] = r_l \{\lambda B(c, \rho)\}. \tag{4.12}$$

Obsérvese que de los λ clientes que llegan por término medio por u.t., $\lambda [1 - B(c, \rho)]$ clientes son atendidos, véase (4.11), y $\lambda B(c, \rho)$ se pierden, véase (4.12).

Combinando estos costes, se puede formular una función de coste medio por unidad de tiempo dependiente del número de servidores, c,

$$\begin{aligned} g(c) &= r_b E[N_b] + r_e E[N_e] + r_s E[N_s] + r_l E[N_l] \\ &= r_b \{\rho [1 - B(c, \rho)]\} + r_e \{c - \rho [1 - B(c, \rho)]\} + r_s \{\lambda [1 - B(c, \rho)]\} + r_l \{\lambda B(c, \rho)\} \\ &= r_s \lambda + (r_b - r_e) \rho + r_e c + \{(r_e - r_b) \rho + (r_l - r_s) \lambda\} B(c, \rho) \end{aligned} \tag{4.13}$$

Cuando se formula una función de coste es importante analizar si existe un único mínimo, o si al menos, se puede averiguar cuántos mínimos existen y su localización, lo cual es posible, en el caso discreto, mediante un procedimiento de optimización monótono, véase Lillo (2000a) y Bell (1971). Este procedimiento consiste en examinar la función discreta de interés, $g(c)$, y encontrar un valor c_0, donde $g(c_0 + 1) - g(c_0) > 0$, y tal que, $g(c + 1) - g(c) > 0$ para todo $c > c_0$. Esta propiedad permite asegurar que existe un punto, c_0, a partir del cual no es posible encontrar un mínimo y, en ese caso, los valores anteriores a $g(c_0)$ pueden ser examinados puesto que constituyen un conjunto finito de elementos. Este procedimiento se puede aplicar sobre la función

de coste dada en (4.13) si $r_e > 0$ ya que, cuando c tiende a infinito, la fórmula de Erlang, $B(c,\rho)$, (4.3), se aproxima a cero y por tanto, la función dada en (4.13) tenderá a una recta de pendiente $r_e > 0$, en la que claramente se puede determinar un punto que verifique las condiciones mencionadas anteriormente.

Además, si $r_e > 0$, una condición suficiente para que la función de coste dada en (4.13) tenga un único mínimo (o dos mínimos iguales y consecutivos) es que el coste por servidor vacío sea mayor que el coste por servidor ocupado, $r_e > r_b$, y que el coste por cliente perdido sea mayor que el coste por cliente servido, $r_l > r_s$, lo cual resulta bastante natural. Para comprobarlo, se puede construir una función de variable real que sea continua, convexa, monótona y que tome los mismos valores que (4.13) en los valores positivos enteros. Para ello, basta con reemplazar en (4.13) la Fórmula de Erlang, $B(c,\rho)$, por su versión continua que es,

$$B(x,\rho) = \left\{ \rho \int_0^\infty e^{-\rho t} (1+t)^x \, dt \right\}^{-1}$$

puesto que esta función es también convexa, véase Jagers y Van~Doorn (1986).

Una vez que se ha formulado una función de coste, el paso siguiente es estimar esta función utilizando los datos que se tienen del proceso de llegadas y de servicio, $\{\mathbf{t}, \mathbf{s}\}$. Dados unos costes fijos, se puede aproximar la función de coste medio, (4.13), como se hace habitualmente, utilizando la muestra MCMC de la distribución a posteriori de los parámetros del sistema,

$$g(c \mid \mathbf{t}, \mathbf{s}) \simeq \frac{1}{J} \sum_{j=1}^{J} g^{(j)}(c),$$

donde $g^{(j)}(c)$ es el coste medio, (4.13), evaluado para cada valor de los parámetros de la muestra. Una medida del error de estas estimaciones es la varianza del coste predictivo, que se obtiene calculando, para cada valor de c, la varianza de los valores $\left\{ g^{(j)}(c);\ j = 1, ..., J \right\}$. Análogamente, se pueden construir intervalos predictivos para el coste calculando los cuantiles de $\left\{ g^{(j)}(c);\ j = 1, ..., J \right\}$ para cada c.

4.1.3.1. Optimización del número de camas en el hospital.

Se ilustra a continuación el análisis del coste descrito anteriormente examinando el número óptimo de camas en el Hospital de St. George. Se tiene en cuenta los costes originados por las camas desocupadas, por la demanda perdida y por los distintos tipos de pacientes en el hospital. No se consideran costes ni beneficios por los pacientes atendidos, es decir, $r_s = 0$. Las unidades de tiempo que se consideran son días puesto que las duraciones de los tiempos de ingreso se han medido en número de días en el hospital geriátrico.

De este modo, se asume un coste, r_e, por cada cama que permanece vacía al día y un coste, $r_l > 0$, por cada paciente que no puede ingresar en el hospital porque no hay camas suficientes. Para cuantificar el coste, r_b, por cada paciente en el hospital al día, se requiere un poco más de cuidado puesto que, como se observó en la Sección 4.1.1, la distribución de los tiempos de ingreso no es homogénea. Se asumen entonces diferentes costes según la duración del periodo de ingreso de cada paciente. Concretamente, se suponen tres tipos de costes diferentes, r_S, r_M y r_L, para estancias cortas en el hospital (S), estancias de duración media (M) y estancias largas (L), respectivamente. En la práctica, el primer caso corresponde, por ejemplo, a pacientes que ingresan por motivos urgentes necesitando tratamientos costosos, como, por ejemplo, operaciones quirúrjicas, y abandonan el hospital en un periodo de tiempo relativamente corto, mientras que, el último caso, puede corresponder a pacientes de salud frágil que permanecen largos periodos de tiempo en el hospital pero sin necesidad de un tratamiento intensivo. Entonces, utilizando la estimación de la densidad del tiempo en el hospital, obtenida en la Sección 4.1.1, se puede calcular fácilmente la proporción de pacientes que pertenecen a cada clase, p_S, p_M y p_L. Teniendo en cuenta (4.4), se tiene que el número medio de camas ocupadas por cada tipo de paciente viene dada por $p_k E[N_b] = p_k \rho [1 - B(c,\rho)]$; con $k = S, M$ o L. Y por tanto, el coste

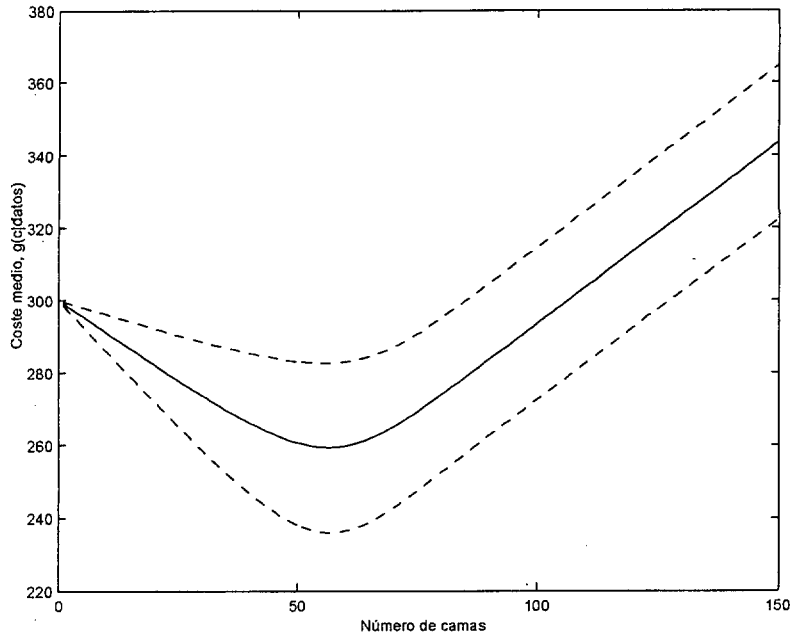

Figura 4.4: Coste medio en función del número de camas y su intervalo predictivo al 80 %.

	c óptimo	$g(c)$	*d.t.* $(g(c))$
$\lambda = 1$	38	174.26	15.22
$\lambda = 1.5$	57	259.31	23.28
$\lambda = 2$	75	344.08	31.37

Tabla 4.3: Número óptimo de camas para diferentes tasas de llegadas al sistema.

medio al día por camas ocupadas es,

$$r_b E[N_b] = r_b \{\rho [1 - B(c, \rho)]\} \tag{4.14}$$

donde $r_b = r_S p_S + r_M p_M + r_L p_L$.

La Figura 4.4 ilustra con línea continua la estimación de la función de coste medio y, con línea discontinua, un intervalo predictivo al 80 %, donde se han asumido los costes siguientes: $r_e = 1, r_l = 200, r_S = 4, r_M = 2$ y $r_L = 5$ unidades. Obsérvese que los valores para r_e y r_l son iguales a los considerados en Gorenescu et al. (1999). Como se puede observar, la estimación del número óptimo de camas resulta igual a 57 con un coste mínimo de 259.31 unidades. Sin embargo, la variación máxima de la estimación se obtiene cerca del mínimo, en concreto, cuando el número de camas es $c = 58$, y la amplitud del intervalo predictivo es mayor en el rango $[55, 60]$.

Es importante también analizar la sensibilidad de estos resultados con respecto a diferentes tasa de llegadas de los pacientes. La Tabla 4.3 muestra el número óptimo de camas para distintos valores de la tasa de llegadas. Como se puede observar, el valor óptimo de c aumenta cuando crece λ así como el valor del coste medio estimado y su varianza. Se puede afirmar que existe un grado de sensibilidad considerable con respecto a la elección de λ lo cual es predecible porque al aumentar λ automáticamente aumenta la intensidad de tráfico, ρ, que influye directamente en todas las medidas implicadas en el coste.

Como se comentó en la Sección 2.5.2.1 del Capítulo 2, la distribución predictiva de la duración de los ingresos hospitalarios no es sensible a la elección de la distribución a priori del número de fases, L, del modelo

MGE. Este mismo fenómeno se ha observado en la práctica en la función de coste predictiva y en el valor del número óptimo de camas que no se ven influenciados por la distribución a priori de L. Obsérvese que una extensión de este análisis podría consistir en examinar la sensibilidad a las elecciones de los valores de los costes, r_e, r_l, etc.

4.2. Análisis del sistema GI/M/c. Aplicaciones sobre el diseño de establecimientos bancarios.

A partir de esta Sección el interés se centra en el modelo de colas GI/M/c, en el cual, a diferencia del anterior, se permite un número ilimitado de clientes en la cola de espera. Se describe un procedimiento Bayesiano para la estimación de las características de este sistema y el diseño óptimo del mismo. El análisis Bayesiano desarrollado en esta Sección se motiva con un conjunto de observaciones reales tomadas en una oficina de una determinada entidad bancaria en Madrid, donde se generan líneas de espera cuyo comportamiento puede analizarse mediante el modelo GI/M/c. En la Subsección 4.2.1, se describe el conjunto de datos reales procedentes del proceso de llegadas y de servicio en el citado establecimiento bancario y se utilizan los métodos de estimación de densidades propuestos en el Capítulo 2 para estimar las distribuciones de interés. En la Sección 4.2.1, se describen las propiedades del sistema GI/M/c y se proponen métodos Bayesianos para su estimación que se aplican sobre el conjunto de datos introducidos previamente. De este modo, se estiman algunas medidas como la distribución del tiempo de espera en la cola de la sucursal bancaria según el número de servidores. Finalmente, en la Sección 4.2.3, se proponen distintas funciones de coste que permiten abordar el problema del diseño de un sistema GI/M/c así como introducir problemas de optimización. Se obtienen también estimaciones de estas funciones que se utilizan para decidir el número óptimo de servidores en el banco local referido anteriormente.

4.2.1. Descripción de los datos y estimación de su distribución utilizando el modelo HEr.

Un grupo de estudiantes de Ingeniería Informática de la Universidad Carlos III de Madrid elaboró, en las prácticas de la asignatura de Investigación Operativa, un proyecto en el que se analizaba el comportamiento de un banco local en Madrid. Para ello, se utilizaron modelos de colas Markovianos y métodos de estimación clásica de densidades. En esta Sección, se considera este mismo conjunto de datos y se aplican los métodos de estimación propuestos en el Capítulo 2 para aproximar las distribuciones del tiempo de servicio y del tiempo entre llegadas a la sucursal bancaria.

Se observan los tiempos entre las llegadas de 98 clientes así como sus tiempos de servicio en el periodo de tiempo transcurrido entre las 10:00 y las 11:30 de la mañana durante tres días. El tiempo medio de servicio es aproximadamente igual a 275.16 segundos. Con cualquiera de los métodos de estimación Bayesiana de densidades propuestos en el Capítulo 2, se predice una distribución exponencial para el tiempo de servicio del sistema. Concretamente, utilizando el algoritmo `RJHEr`, se obtiene que la probabilidad a posteriori de que los datos de servicio, $\mathbf{s}$, se distribuyan exponencialmente es, $P(\nu = 1, k = 1 \mid \mathbf{s}) = 0.998$, véase (2.31), y si se utiliza el algoritmo `RJMGE`, la probabilidad a posteriori de que el tiempo de servicio sea exponencial es, $P(L = 1 \mid \mathbf{s}) = 0.989$, véase (2.34). Por tanto, se asume que el tiempo de servicio en el establecimiento bancario es exponencial y se utilizan los procedimientos de inferencia de la Sección 1.4.2.3 para distribuciones exponenciales. Procediendo de este modo, se considera una distribución a priori no informativa para la tasa de servicio, μ, dada en (1.21) con a y b iguales a cero y entonces, la distribución a posteriori es $\mu \mid \mathbf{s} \sim G(98, 26965.6)$.

A continuación, se selecciona una clase de distribuciones para ajustar una distribución adecuada al conjunto de observaciones de tiempos entre llegadas al establecimiento bancario, $\mathbf{t}$. La Figura 4.5 muestra un

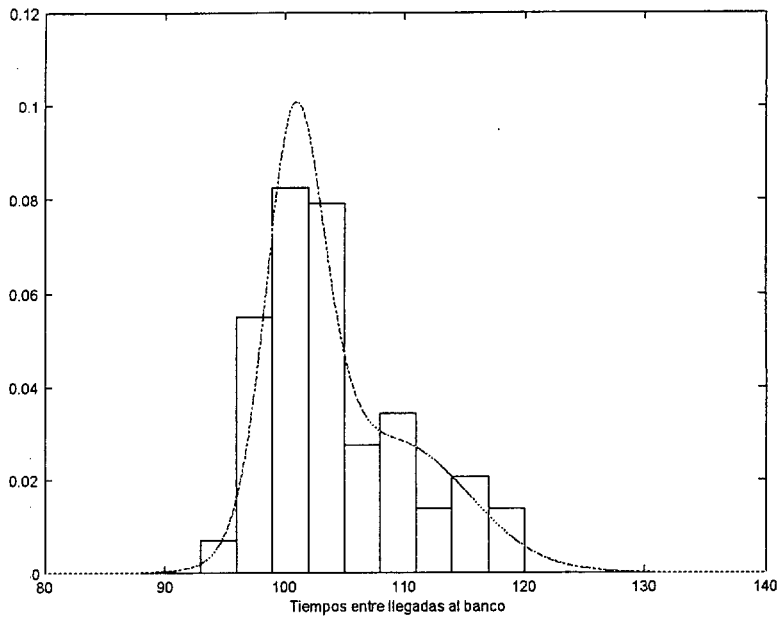

Figura 4.5: Histograma de los datos de tiempos entre llegas al establecimiento bancario y función de densidad estimada.

histograma de las 98 observaciones. Como su distribución no parece unimodal, siguiendo las indicaciones de la Sección 2.2.3, se escoge el modelo de mixtura HEr que permite captar este comportamiento sin necesidad de utilizar un número elevado de parámetros. En la Figura 4.5, se muestra también la función de densidad estimada obtenida mediante (2.29) a partir del algoritmo `RJHEr`, desarrollado en la Sección 2.4.1.1. Ninguno de los tiempos entre las llegadas al sistema es superior a 2 minutos y la función de densidad predictiva tiene, como el histograma, un aspecto bimodal. De hecho, la distribución a posteriori de que haya 2 componentes en la mixtura de Erlangs es muy elevada, $P(k = 2 \mid \mathbf{t}) \approx 0.958$.

4.2.2. Sistema GI/M/c y estimación de sus características.

Teniendo en cuenta los resultados de la Subsección anterior, un modelo de colas apropiado para describir el comportamiento observado en el establecimiento bancario es el sistema HEr/M/c, con c servidores, capacidad infinita, distribución de servicio exponencial y distribución HEr para el tiempo entre las llegadas al sistema. En esta Subsección, se describe cómo estimar las características de un sistema GI/M/c aproximando la distribución general del tiempo entre llegadas con las mixturas HEr y MGE, es decir, considerando los sistemas HEr/M/c y MGE/M/c. En primer lugar, se describen las propiedades del sistema GI/M/c y a continuación, se proponen métodos Bayesianos para su estimación. Por último, se aplica el procedimiento descrito sobre el conjunto de datos observados en la sucursal bancaria.

4.2.2.1. Características del sistema GI/M/c.

Se supone, en este apartado, un sistema de colas GI/M/c con parámetros conocidos, $\theta = \{\theta_\lambda, \mu\}$, donde θ_λ representa el conjunto de parámetros de la distribución general del tiempo entre llegadas y μ es la tasa de servicio. Se asume también que se verifica la condición de equilibrio, es decir, $\rho < c$, donde ρ es la intensidad

de tráfico dada por, véase (1.5),

$$\rho = \left(\mu E\left[A \mid \theta_\lambda\right]\right)^{-1}$$

y donde A es la variable aleatoria que representa el tiempo entre las llegadas al sistema.

Un resultado muy conocido, véase Kleinrock (1975), es que en sistemas de colas con proceso de llegadas que no es Markoviano la distribución del número de clientes, N^*, que encuentra un cliente que llega al sistema es distinta de la distribución del número de clientes, N, que encuentra un cliente que llega en un instante aleatorio, véanse los comentarios y resultados relativos al sistema GI/M/1 en la Subsección 1.3.1 del Capítulo 1. En particular, para el sistema de colas GI/M/c, la distribución estacionaria del número de clientes en el sistema en los instantes de llegada tiene la expresión siguiente, véase Allen (1990),

$$P\left(N^* = n \mid \theta\right) = \begin{cases} \Sigma_{m=1}^{c-1}(-1)^{m-n}\binom{m}{n}U_m & \text{para } n = 0,1,...,c-2, \\ D\sigma^{n-c} & n \geq c-1, \end{cases} \tag{4.15}$$

donde σ es la única raíz en el intervalo $(0,1)$ de la ecuación,

$$\sigma = f_A^*\left(c\mu\left(1-\sigma\right)\right), \tag{4.16}$$

y f_A^* es la transformada de Laplace-Stieljes, definida en (1.16), de la distribución del tiempo entre llegadas y

$$\begin{aligned} U_n &= DC_n \sum_{m=n+1}^{c} \frac{\binom{c}{m}}{C_m(1-g_m)} \frac{c(1-g_m)-m}{c(1-\sigma)-m}, \qquad \text{para } n = 0,1,...,c-1, \\ g_m &= f_A^*\left(m\mu\right), \qquad \text{para } m = 1,...,c, \\ C_m &= \begin{cases} 1 & \text{si } m = 0, \\ \prod_{n=1}^{m}\left(\frac{g_n}{1-g_n}\right) & \text{si } m = 1,2,...,c, \end{cases} \\ D &= \left[\frac{1}{1-\sigma} + \sum_{m=1}^{c} \frac{\binom{c}{m}}{C_m(1-g_m)} \frac{c(1-g_m)-m}{c(1-\sigma)-m}\right]^{-1}. \end{aligned} \tag{4.17}$$

Obsérvese que en el caso particular en el que haya un único servidor, la distribución de N^* es geométrica con parámetro σ.

La distribución estacionaria del número de clientes que encuentra una llegada aleatoria, N, depende de la distribución de N^* y viene dada por, véase Allen (1990),

$$P\left(N = n \mid \theta\right) = \begin{cases} 1-\rho-\rho c\Sigma_{j=1}^{c-1}P\left(N^* = j-1 \mid \theta\right)\left(\frac{1}{j}-\frac{1}{c}\right) & \text{para } n = 0, \\ \frac{\rho c}{n}P\left(N^* = n-1 \mid \theta\right) & \text{para } n = 1,...,c-1, \\ \rho P\left(N^* = n-1 \mid \theta\right) & \text{para } n \geq c. \end{cases} \tag{4.18}$$

Para aproximar la distribución general del tiempo entre llegadas, se puede considerar el modelo de mixtura HEr con parámetros $\theta_\lambda = \{k, \mathbf{w}, \boldsymbol{\lambda}, \boldsymbol{\nu}\}$, véase (2.1). Obsérvese que las tasas de las fases de la distribución HEr se han denotado con λ_r en lugar de μ_r por tratarse del tiempo entre llegadas. Entonces, se pueden calcular las distribuciones de N^* y de N para el sistema HEr/M/c teniendo en cuenta que la transformada de Laplace de la distribución HEr viene dada por,

$$f_{HEr}^*\left(s\right) = \sum_{r=1}^{k} w_r\left(\frac{\nu_r\lambda_r}{s+\nu_r\lambda_r}\right)^{\nu_r}.$$

Alternativamente, se puede considerar la mixtura MGE de modo que se aproxime el sistema GI/M/c con el sistema MGE/M/c. En ese caso, para obtener las distribuciones estacionarias, hay que tener en cuenta que la transformada de Laplace de la distribución MGE viene dada por,

$$f_{MGE}^*\left(s\right) = \sum_{r=1}^{L} P_r \prod_{i=1}^{r}\frac{\lambda_i}{\left(\lambda_i+s\right)},$$

donde $\theta_\lambda = \{L, \mathbf{P}, \boldsymbol{\lambda}\}$ son los parámetros de la mixtura MGE, véase (2.4). Tanto en el sistema HEr/M/c como en el sistema MGE/M/c, es fácil aproximar la raíz σ, dada en (4.16), utilizando el método de Newton-Raphson o un procedimiento similar.

Otra medida importante que interesa, fundamentalmente, a los clientes que llegan al sistema, es el tiempo que tienen que permanecer esperando en cola. Dados los parámetros del sistema, θ, la distribución estacionaria del tiempo de espera en cola, W, es proporcional a una distribución exponencial y toma un valor no negativo, $P(W = 0)$, en el origen. La función de distribución viene dada por, véase Allen (1990),

$$F_W(x \mid \theta) = 1 - P(W > 0) \exp\{-c\mu(1-\sigma)x\}, \qquad x \geq 0, \tag{4.19}$$

donde,

$$P(W > 0 \mid \theta) = \frac{D}{1-\sigma} \tag{4.20}$$

y donde σ y D vienen dados en (4.16) and (4.17), respectivamente.

4.2.2.2. Inferencia y predicción Bayesiana en el sistema GI/M/c.

Se describe a continuación cómo desarrollar inferencia Bayesiana en el sistemas de colas GI/M/c a partir de una muestra de datos de tiempos de servicio, $\mathbf{s}$, y de tiempos entre llegadas, $\mathbf{t}$. Se supone que se han utilizado una de las distribuciones HEr ó MGE para aproximar la distribución general del tiempo entre llegadas, A, y que se ha obtenido una muestra MCMC de tamaño J de la distribución a posteriori de θ_λ a partir de los algoritmos introducidos en el Capítulo 2 y otra muestra de la distribución a posteriori de μ generando valores de la distribución gamma dada en 1.21. Puesto que el modelo de colas considerado tiene capacidad infinita, es importante examinar si en el sistema observado se verifica la condición de equilibrio para cerciorarse de que existen las distribuciones estacionarias de las cantidades de interés. Para un determinado valor del número de servidores, c, se puede estimar la probabilidad de que haya equilibrio en el sistema análogamente a como se describió en el Capítulo 3, véase (3.31), es decir,

$$P(\rho < c \mid \mathbf{t}, \mathbf{s}) \approx \frac{1}{J} \# \left\{ \rho^{(j)} < c \right\}, \tag{4.21}$$

donde,

$$\rho^{(j)} = \left(\mu^{(j)} E\left[A \mid \theta_\lambda^{(j)} \right] \right)^{-1}, \tag{4.22}$$

y $\{\theta_\lambda^{(j)}, \mu^{(j)}\}_{j=1}^{J}$ son una colección de pares de las muestras MCMC obtenidas. El valor de la media a posteriori del tiempo entre llegadas, $E[A \mid \theta_\lambda^{(r)}]$, se obtiene utilizando (3.28) y (3.29), para las distribuciones HEr y MGE, respectivamente. Si, para un número determinado de servidores, c, la probabilidad dada en (4.21) es lo suficientemente grande (normalmente, mayor que 0.8), resulta razonable considerar que el sistema es estable y se pueden estimar las distribuciones estacionarias, siguiendo el mismo procedimiento del Capítulo 3 para sistemas con un único servidor. Por ejemplo, si se pretende decidir un número apropiado de servidores para el sistema, es recomendable considerar valores de c tales que la probabilidad dada en (4.21) sea lo suficientemente elevada como para asegurar el equilibrio del sistema. Concretamente, en la Sección siguiente, las funciones de coste que se construyen toman valores a partir de un mínimo que garantiza el equilibrio con alta probabilidad.

Suponiendo que existe equilibrio, se pueden aproximar las distribuciones predictivas de N^* y de N, calculando las medias de las probabilidades dadas en (4.15) y en (4.18), respectivamente, para los valores de la muestra MCMC que verifican la condición de equilibrio. Por ejemplo, la distribución predictiva de N se puede aproximar mediante,

$$P(N = n \mid \mathbf{t}, \mathbf{s}, \rho < c) \approx \frac{1}{R} \sum_{j:\rho^{(j)} < c} P\left(N = n \mid \theta_\lambda^{(j)}, \mu^{(j)} \right) \tag{4.23}$$

donde $R = \#\left\{\rho^{(j)} < c\right\}$. Obsérvese que, para obtener esta estimación, la ecuación dada en (4.16) tiene que resolverse para cada valor de los parámetros $(\theta_\lambda^{(j)}, \mu^{(j)})$ de la colección de pares de la muestra MCMC. Además, se pueden aproximar otras distribuciones estacionarias en el sistema GI/M/c. Por ejemplo, la distribución del número de servidores ocupados se puede estimar fácilmente teniendo en cuenta que depende del número de clientes en el sistema. Nuevamente, es necesario distinguir entre el número de servidores ocupados que encuentra un cliente que llega, N_b^*, y el número de servidores ocupados que encuentra una llegada aleatoria, N_b. Obsérvese que el número de servidores ocupados es igual al número de clientes en el sistema si el número de clientes es menor o igual que el de servidores. Y es igual a c si hay más clientes que servidores. Consecuentemente, las distribuciones predictivas de N_b^* y de N_b se pueden estimar a partir de las distribuciones predictivas de N^* y de N, respectivamente. Por ejemplo, la distribución de N_b se aproxima a partir de (4.23) del modo siguiente,

$$P\left(N_b = n \mid \mathbf{t}, \mathbf{s}, \rho < c\right) = \begin{cases} P\left(N = n \mid \mathbf{t}, \mathbf{s}, \rho < c\right) & \text{si } n < c, \\ P\left(N \geq c \mid \mathbf{t}, \mathbf{s}, \rho < c\right) & \text{si } n = c. \end{cases} \tag{4.24}$$

Obsérvese que no es necesario que el sistema sea estable para que existan las distribuciones estacionarias de N_b^* y de N_b ya que no pueden tomar valores por encima de c. Consecuentemente, no es necesario asumir equilibrio para estimar sus distribuciones predictivas. Por ejemplo, se puede estimar la distribución predictiva de N_b sin condicionar a $\rho < c$ teniendo en cuenta que,

$$P\left(N_b = n \mid \mathbf{t}, \mathbf{s}\right) = P\left(\rho < c \mid \mathbf{t}, \mathbf{s}\right) P\left(N_b = n \mid \mathbf{t}, \mathbf{s}, \rho < c\right) + P\left(\rho \geq c \mid \mathbf{t}, \mathbf{s}\right) P\left(N_b = n \mid \mathbf{t}, \mathbf{s}, \rho \geq c\right)$$

y que, si no hay equilibrio, el número de servidores ocupados es c con probabilidad uno.

Del mismo modo, se puede estimar la distribución predictiva del número de clientes que encuentra esperando en cola un cliente que llega, N_q^*, y también, el número de clientes en cola que encuentra una llegada aleatoria, N_q, utilizando la distribución predictiva del número de clientes presentes en el sistema dada en (4.23),

$$P\left(N_q = n \mid \mathbf{t}, \mathbf{s}, \rho < c\right) = \begin{cases} P\left(N \leq c \mid \mathbf{t}, \mathbf{s}, \rho < c\right) & \text{si } n = 0, \\ P\left(N = c + n \mid \mathbf{t}, \mathbf{s}, \rho < c\right) & \text{si } n \geq 1. \end{cases} \tag{4.25}$$

Por último, se puede estimar la distribución predictiva del tiempo de espera en cola utilizando la siguiente aproximación,

$$F_W\left(x \mid \mathbf{t}, \mathbf{s}, \rho < c\right) \approx \frac{1}{R} \sum_{j:\rho^{(j)} < c} F_W\left(x \mid \theta_\lambda^{(j)}, \mu^{(j)}\right) \tag{4.26}$$

donde, al igual que antes, $R = \#\left\{\rho^{(j)} < c\right\}$.

El mismo problema con la no existencia de momentos que se indicó en los Capítulos 1 y 2 aparece también en este contexto. Con las condiciones a priori que se han considerado, los momentos de las distribuciones predictivas de N^*, de N y de W no existen para el sistema GI/M/c. Para comprobarlo basta con generalizar los argumentos de Wiper (1998), desarrollados para el sistema GI/M/1, y observar que si se asumen distribuciones a priori independientes para los parámetros de la distribución del tiempo de servicio y del tiempo entre llegadas a un sistema GI/M/c y su función de densidad conjunta es continua y positiva en $\rho = c$, las distribuciones dadas en (4.23) y en (4.26) no tienen momentos finitos. Sin embargo, es posible evaluar los valores esperados de estas distribuciones predictivas si se asume que $\rho < c - \varepsilon$ en lugar de $\rho < c$, véase Lehoczky (1990), pero, en la práctica, se ha comprobado que este procedimiento es muy sensible a la elección de ε. Obsérvese, por otra parte, que la distribución predictiva del número de servidores ocupados, N_b, dada en (4.24), sí que tiene momentos finitos. Concretamente,

$$E\left[N_b \mid \mathbf{t}, \mathbf{s}, \rho < c\right] = E\left[\rho \mid \mathbf{t}, \mathbf{s}, \rho < c\right] \approx \frac{1}{R} \sum_{j:\rho^{(j)} < c} \rho^{(j)} \tag{4.27}$$

donde $\rho^{(j)}$ viene dado en (4.22). Se puede estimar también el número medio de servidores ocupados sin asumir equilibrio,

$$E[N_b \mid \mathbf{t},\mathbf{s}] = P(\rho < c \mid \mathbf{t},\mathbf{s})\, E[\rho \mid \mathbf{t},\mathbf{s},\rho < c] + P(\rho \geq c \mid \mathbf{t},\mathbf{s})\, E[N_b \mid \mathbf{t},\mathbf{s},\rho \geq c] \approx \frac{1}{J}\left\{\sum_{R} \rho^{(j)} + (J-R)\, c\right\}$$

donde J es el tamaño de la muestra MCMC completa y R es el número de conjuntos de parámetros que verifican la condición de equilibrio.

4.2.2.3. Modelo de colas HEr/M/c para el establecimiento bancario.

Se aplican ahora los procedimientos que se acaban de describir para desarrollar inferencia y predicción Bayesiana en el sistema de colas observado en el establecimiento bancario. El objetivo es predecir el comportamiento del sistema en función del número de servidores y como consecuencia, todas las características anteriores como por ejemplo, la distribución del tiempo de espera, se deben estimar como funciones de c. Se pretende, por ejemplo, estimar la distribución del tiempo de espera en cola según el número de servidores. En la Subsección 4.2.1, se comprobó que el modelo de colas HEr/M/c era adecuado para el sistema observado y se obtuvieron muestras MCMC de las distribuciones a posteriori de los parámetros de la distribución del tiempo de servicio y del tiempo entre llegadas al sistema. Con los métodos descritos anteriormente y las muestras MCMC obtenidas se estiman a continuación las características del sistema de colas del banco para cada valor del número de servidores, c.

Dados los tiempos de servicio y de llegadas a la sucursal bancaria, se estiman, para cada valor de c, la probabilidad a posteriori de que el sistema sea estable, véase (4.21), como se muestra en la Tabla 4.4. Se observa que, al menos, se necesitan tres servidores para asumir que se verifica la condición de ergodicidad, $\rho < c$. Sin embargo, tres servidores pueden no ser suficientes para satisfacer ciertas condiciones de optimalidad que son el resultado de un balanceo adecuado entre todos los costes implicados en el sistema. Por tanto, el siguiente apartado se dedica a estudiar la elección óptima de c para estructuras de coste prefijadas.

La Tabla 4.4 muestra también las estimaciones de la intensidad de tráfico para cada número de servidores asumiendo que existe equilibrio en el sistema, obtenidas a partir de (4.27). Teniendo en cuenta el resultado general dado en (1.8), estas cantidades coinciden con el número medio de servidores ocupados para cada valor de c. Obsérvese que, puesto que con 1 y 2 servidores el sistema no está probablemente en equilibrio, todos los servidores están casi siempre ocupados por término medio. Sin embargo, a partir de 3 servidores el sistema es estable y se observa que, por término medio, están ocupados 2.66 servidores, aproximadamente.

c	1	2	3	4	5	6
$P(\rho < c \mid \mathbf{s},\mathbf{t})$	.00001	.00181	.89194	.99996	1.00	1.00
$E[\rho \mid \mathbf{s},\mathbf{t},\rho < c]$	0.999	1.976	2.661	2.660	2.660	2.6580

Tabla 4.4: Estimaciones de la probabilidad a posteriori de que exista equilibrio en el sistema y valores esperados de la intensidad de tráfico, según el número de servidores, c.

La Figura 4.6 ilustra las estimaciones de la distribución del número de clientes presentes en el sistema, N, que encuentra un observador que llega en un instante aleatorio, obtenidas mediante (4.23), para 3, 4 y 5 servidores. Obsérvese que la probabilidad de encontrar 2 ó 3 clientes en el sistema es muy similar para cualquier valor del número de servidores. Se ha observado que este fenómeno no aparece en la distribución predictiva de N^* donde se observa que la moda de N^* es igual a 2 para cualquier valor del número de servidores. Este hecho implica que, aunque el número medio de clientes en el sistema que se observa en un instante aleatorio es 2.66 (supóngase, por ejemplo, observado por el director de la oficina), el número de clientes que encuentran los clientes que llegan al banco por término medio es inferior a 2.66. Por otro lado, nótese que la distribución de N en un sistema con 3 servidores tiene una cola más pesada que el resto de

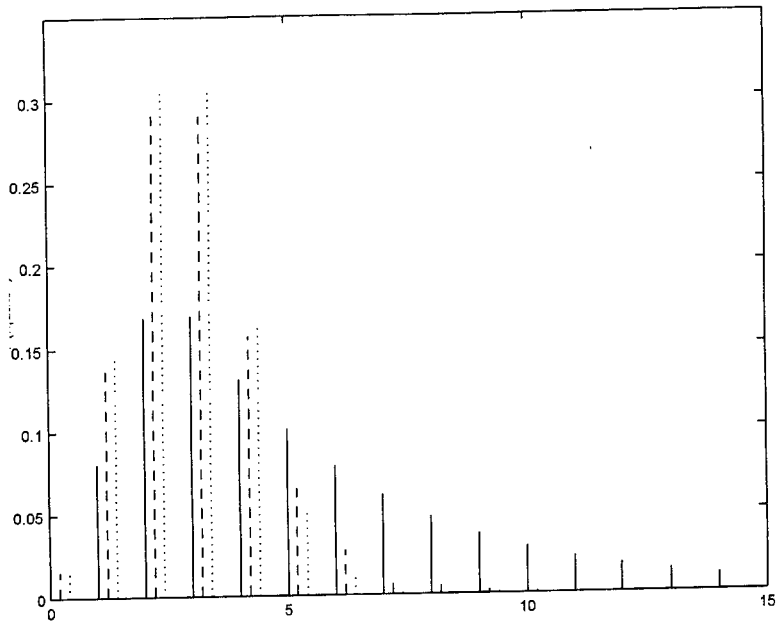

Figura 4.6: Probabilidades predictivas del número de clientes presentes en el banco para 3 servidores (−) , 4 servidores (−−) y 5 servidores (· · ·).

sistemas. En particular, se ha observado que simplemente incrementando el número de servidores de 3 a 4, la estimación de la probabilidad de que la cola esté vacía, $P(N_q = 0 \mid \mathbf{t}, \mathbf{s})$, obtenida mediante (4.25), aumenta de 0.42 a 0.89.

La Figura 4.7 muestra la distribución del tiempo de espera en cola, W, obtenida mediante (4.26), para 3, 4 y 5 servidores. Se puede observar en este caso que 3 servidores pueden ser suficientes, en términos generales, para satisfacer las necesidades de un establecimiento bancario habitual. Por ejemplo, con tres servidores, la probabilidad estimada de esperar en cola menos de 10 minutos es bastante elevada, aproximadamente 0.85. De hecho, el número real de servidores que se encontraban atendiendo en la sucursal observada era de 3 servidores. Sin embargo, nuevamente, si se aumenta el número de servidores de 3 a 4 la probabilidad de no tener que esperar en cola, $P(W = 0 \mid \mathbf{t}, \mathbf{s})$, obtenida mediante (4.26), se incrementa de 0.35 a 0.83.

4.2.3. Función de coste y diseño óptimo del modelo.

En esta Subsección, se considera el problema de la determinación del número óptimo de servidores en el sistema GI/M/c. La estructura de coste que se considera es similar a la desarrollada para el sistema M/G/c/c. Sin embargo, como ya se ha comentado, el sistema GI/M/c tiene capacidad infinita y por tanto, los clientes que llegan y encuentran todos los servidores ocupados permanecen en el sistema formando una línea de espera. Consecuentemente, se incluyen, en este caso, costes asociados a la espera de los clientes que reemplazan a los costes por demanda perdida considerados en el sistema M/G/c/c. En concreto, se consideran los costes siguientes por unidad de tiempo:

r_b = Coste por u.t. por cada servidor ocupado.
r_e = Coste por u.t. por cada servidor vacío.
r_s = Coste por cada cliente servido.
r_q = Coste por u.t. por cada cliente esperando en cola.

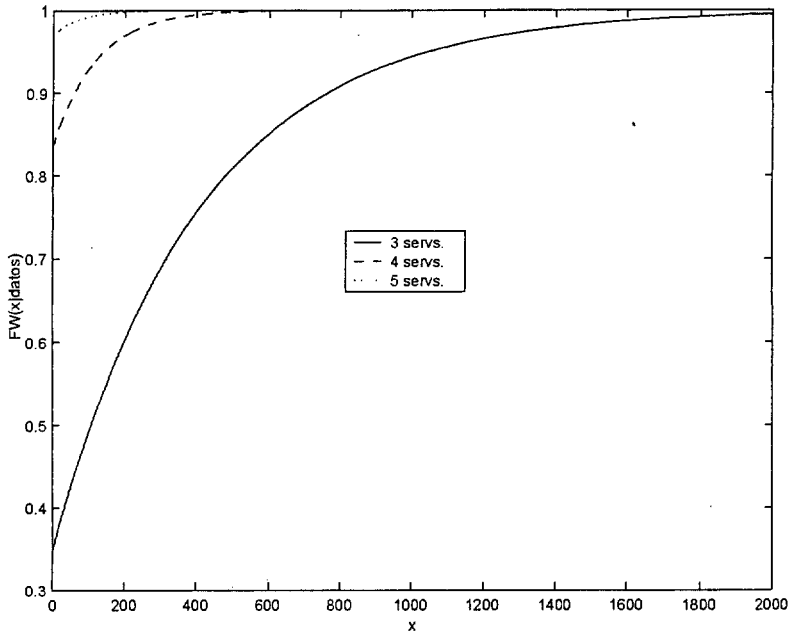

Figura 4.7: Funciones de distribución predictivas del tiempo de espera en la cola del banco para 3, 4 y 5 servidores.

r_W = Coste por u.t. de espera en cola de un cliente.

Todos estos costes tomarán valores positivos o negativos dependiendo de si indican beneficio o pérdida, como se considerará en la aplicación al establecimiento bancario. Una vez definidos los costes, se puede formular una función de coste evaluada cuando el sistema se encuentra en el estado estacionario. Puesto que el diseño del sistema no es, en general, una labor propia de los clientes sino de un personal independiente, se consideran en la función de coste medidas del sistema observadas en un instante aleatorio y no en los instantes de llegadas, que se introdujeron en la Sección anterior. Con esta construcción, dados los parámetros del sistema, el coste total por unidad de tiempo es,

$$Coste = r_b N_b + r_e \{c - N_b\} + r_s N_s + r_q L(N_q) + r_W L(W) \tag{4.28}$$

donde N_s es el número de clientes servidos por u.t. y, N_b, N_q y W son el número de servidores ocupados, el número de clientes y el tiempo de espera en cola, respectivamente, definidos en la Sección anterior. $L(N_q)$ representa la pérdida originada por el número de personas esperando en cola para recibir su servicio y $L(W)$ es la pérdida originada por el tiempo de espera en cola. Por ejemplo, se puede considerar una pérdida con la siguiente estructura,

$$L_1(N_q) = \begin{cases} 0 & \text{si } N_q \leq n_0, \\ 1 & \text{si } N_q > n_0. \end{cases} \tag{4.29}$$

De este modo, se considera un coste, r_q, por u.t. si el número de clientes esperando en cola supera una cota fijada previamente, n_0. Alternativamente, se puede considerar otra función de pérdida más realista considerando,

$$L_2(N_q) = \begin{cases} N_q & \text{si } N_q \leq n_0, \\ n_0 & \text{si } N_q > n_0. \end{cases} \tag{4.30}$$

En este caso, se asume un coste, r_q, por u.t. por cliente esperando en cola si el número de clientes en cola no excede del límite, $n_0 < \infty$. Se pueden formular funciones de pérdida análogas, $L_1(W)$ y $L_2(W)$, para

el tiempo de espera en cola. En ese caso, se debe fijar una cota para el tiempo de espera en cola, $w_0 < \infty$. Obsérvese que el motivo por el que se han incluido funciones de pérdida en el coste, (4.28), es que, como se puntualizó en la Sección anterior, las distribuciones predictivas de W y N_q no tienen momentos finitos y consecuentemente, si se asume, por ejemplo, una cota infinita, $n_0 = \infty$, en (4.30), se obtendrán valores infinitos para la estimación del coste medio esperado, $E[Coste \mid \mathbf{t}, \mathbf{s}]$, y no tendría sentido el problema de optimización.

Conocidos los parámetros, $\{\boldsymbol{\theta}_\lambda, \mu\}$, del sistema GI/M/c y suponiendo que verifica la condición de equilibrio, $\rho < c$, se puede calcular la esperanza del coste por u.t., dado en (4.28), para cada valor del número de servidores, c,

$$g(c) = E[Coste \mid \boldsymbol{\theta}_\lambda, \mu] = r_e c + (r_b - r_e + r_s \mu)\rho + r_q P(N_q > n_0 \mid \boldsymbol{\theta}_\lambda, \mu) + r_W P(W > w_0 \mid \boldsymbol{\theta}_\lambda, \mu). \quad (4.31)$$

Para comprobar esta expresión basta con tener en cuenta que, en media, cada servidor atiende a μ clientes por unidad de tiempo y por tanto, se tiene que,

$$E[N_s \mid \boldsymbol{\theta}_\lambda, \mu] = \mu E[N_b \mid \boldsymbol{\theta}_\lambda, \mu],$$

y además, como se puntualizó en (1.8), el número medio de servidores ocupados en cualquier sistema GI/G/c en equilibrio es igual a la intensidad de tráfico, luego,

$$E[N_b \mid \boldsymbol{\theta}_\lambda, \mu] = \rho.$$

Por último, nótese que en (4.31) se han considerado funciones de pérdida con la estructura dada en (4.29). Análogamente, se pueden obtener expresiones para el coste medio utilizando funciones de pérdida del tipo dado en (4.30). Las cantidades necesarias para calcular las probabilidades referentes a N_q y a W en (4.31) se obtienen utilizando las distribuciones introducidas en la Sección 4.2.2.2, véase (4.18) y (4.19).

Es fácil comprobar que la función, $g(c)$, dada en (4.31), admite un procedimiento monótono de optimización, como el descrito en la Sección 4.1.3, si el coste por u.t. por servidor vacío, r_e, es positivo. Obsérvese que las probabilidades de que el número de clientes en cola, N_q, y el tiempo de espera en cola, W, superen unas cotas fijas, n_0 y w_0, tienden a cero a medida que se incrementa el número de servidores, c. Por tanto, la función de coste, $g(c)$, será cada vez más parecida a una recta de pendiente, r_e, a medida que aumente c. Lo cual implica que existe un valor de c a partir del cual la función $g(c)$ es monótona creciente. El mismo argumento permite comprobar que la función de coste es monótona si se utilizan funciones de pérdida con la estructura dada en (4.30).

Si no se conocen los parámetros del sistema pero se tiene una muestra de tiempos entre las llegadas y de servicio, $\{\mathbf{t}, \mathbf{s}\}$, se puede estimar el valor del coste medio por u.t. de la manera habitual,

$$E[Coste \mid \mathbf{t}, \mathbf{s}, \rho < c] \approx \frac{1}{R} \sum_{j:\rho^{(j)}<c} g^{(j)}(c), \quad (4.32)$$

donde $g_j(c)$ es el valor del coste medio por u.t. dado en (4.31) para cada valor de los parámetros de la muestra MCMC que verifican la condición de equilibrio,

$$g^{(j)}(c) = E\left[Coste \mid \boldsymbol{\theta}_\lambda^{(j)}, \mu^{(j)}\right].$$

Obsérvese que el coste predictivo, (4.32), está definido para cualquier valor del número de servidores, c, siempre que la probabilidad de que se verifique la condición de equilibrio, $\rho < c$, sea mayor que cero. Sin embargo, en la práctica, si dicha probabilidad es muy pequeña, no resulta razonable evaluar el coste puesto que, en ese caso, se está asumiendo un hecho que es muy poco probable. Siguiendo esta argumentación, las funciones de coste que se estiman en la Sección siguiente para la sucursal bancaria, están evaluadas a partir del mínimo valor del número de servidores que origina una probabilidad de equilibrio mayor de 0.8.

4.2.3.1. Optimización del número del servidores en el banco.

Se pretende a continuación analizar el diseño del sistema de colas observado en la sucursal bancaria. Para ello, se utilizan los resultados de las estimaciones obtenidas para este sistema en la Sección 4.2.2.3 y los procedimientos desarrollados en la Sección anterior.

Se formulan varias funciones de coste definidas a partir del mínimo valor del número de servidores para el cual se ha asumido equilibrio, es decir, tres servidores. En la práctica, resulta complicado establecer los costes asociados a la espera de los clientes mientras que, en general, se conocen, los costes asociados a los gastos en servidores. Por tanto, en las funciones que se formulan a continuación, se suponen valores fijos para los costes por servidor ocupado, $r_b = 1.5$, por servidor vacío, $r_e = 1$, y por cliente servido, $r_s = -0.05$, y se consideran diferentes valores para los costes relacionados con el número de clientes en cola, N_q, y el tiempo de espera en cola, W, dados por, r_q y r_W, respectivamente. Se suponen además funciones de pérdida con la estructura dada en (4.30) para N_q y con la estructura de (4.29) para W. Con esta formulación, el valor del coste medio por u.t. para cada valor de c es,

$$\begin{aligned} g(c) &= (r_b - r_e + r_s\mu)\,\rho + r_e c + r_q E\left[L_2\left(N_q\right)\right] + r_W E\left[L_1\left(W\right)\right] \\ &= cte. + r_e c + r_q \left\{E\left[N_q \mid N_q \leq n_0\right] P\left(N_q \leq n_0\right) + n_0 P\left(N_q > n_0\right)\right\} + r_W P\left(W > w_0\right). \end{aligned} \quad (4.33)$$

Por último, se asume que $n_0 = 20$ y $w_0 = 2000$, puesto que la probabilidad de superar estas cotas es muy pequeña, como se puede comprobar el las Figuras 4.6 y 4.7.

La Tabla 4.5 muestra las estimaciones del coste medio por unidad de tiempo dado en (4.33) obtenidas mediante (4.32) para varios valores de r_q y r_W. Cada columna origina una función de coste dependiente de c. Se indican en negrita los valores óptimos del coste en cada caso. Como se comentó anteriormente, a medida que aumenta el valor de c, todas las funciones tienden a una recta de pendiente $r_e = 1$. Se observa además que las dos primeras funciones toman valores prácticamente iguales a partir de $c \geq 5$ y lo mismo sucede para las dos últimas funciones. El motivo es que, si el sistema dispone de más de cuatro servidores, el incremento de r_W no influye en el valor del coste, (4.33), puesto que, en ese caso, la probabilidad de que el tiempo de espera sea superior a $w_0 = 2000$ segundos es prácticamente nula. Sin embargo, si $c \geq 5$, el incremento del coste por cliente en cola, r_q, sí que afecta algo en la función de coste, (4.33), porque el valor de $E\left[N_q \mid N_q \leq 20\right]$ es un poco mayor en ese caso. En cualquier caso, para valores elevados de c, todas las funciones de coste tenderán a la misma recta de pendiente $r_e = 1$ y no afectarán los valores de r_q ni de r_W. Este fenómeno puede apreciarse en la Tabla 4.5 observando que para $c = 7$ todas las funciones toman un valor muy parecido.

	$E\left[Coste \mid \mathbf{s}, \mathbf{t}, \rho < c\right]$			
c	$r_W = .01$ $r_q = .01$	$r_W = 1$ $r_q = .01$	$r_W = .01$ $r_q = 1$	$r_W = 1$ $r_q = 1$
3	**4.3861**	**4.4536**	9.9632	10.0307
4	5.3364	5.3365	**6.0739**	**6.0740**
5	6.3309	6.3309	6.5111	6.5111
6	7.3295	7.3295	7.3778	7.3778
7	8.3292	8.3292	8.3419	8.3419

Tabla 4.5: Estimaciones del coste medio por u.t. para diferentes valores de r_q y r_W . Se indican en negrita los valores óptimos.

4.3. Comentarios y extensiones.

En este Capítulo, se han desarrollado procedimientos Bayesianos para la inferencia, predicción y el diseño de los modelos de colas M/G/c/c y GI/M/c. El análisis de estos sistemas se ha motivado e ilustrado con información real procedente de dos situaciones concretas ubicadas en un hospital geriátrico de Londres y un establecimiento bancario de Madrid.

Como se puede observar, los sistemas de colas que se han considerado son casos particulares de los modelos de colas más generales, GI/G/c/K, con K$\leq \infty$, sobre los que se pueden extender los métodos desarrollados en este Capítulo, aunque, no es siempre fácil calcular las distribuciones estacionarias de las características de interés de estos sistemas dados sus parámetros. Lo que sí resulta, generalmente, posible es formular y estimar una función de coste para cualquier sistema GI/G/c/K, con K$\leq \infty$. Obsérvese que el caso general, el número de clientes en cola no puede ser superior a $K-c$, y por tanto, se deben considerar simultáneamente los costes originados por la espera en cola y por la demanda perdida. En un sistema GI/G/c/K, con K$\leq \infty$, se tiene que,

$$\begin{aligned} E[N_b] &= \rho[1-P(N=K)], \\ E[N_s] &= \mu E[N_b] = \lambda[1-P(N=K)], \\ E[N_e] &= c - E[N_b] = c - \rho[1-P(N=K)], \\ E[N_l] &= \lambda[1-P(N=K)], \end{aligned} \tag{4.34}$$

donde $P(N=K)$ es la probabilidad de que haya K clientes presentes en el sistema, es decir, que el sistema esté bloqueado. Si $K=\infty$, se supone que $P(N=K)=0$. Las tasas de servicio y de llegadas, λ y μ, no tienen que ser necesariamente exponenciales. Con las expresiones dadas en (4.34) y los valores medios de la pérdidas que se consideren, $E[L(N_q)]$ y $E[L(W)]$, se puede estimar la función de coste medio por u.t. en cualquier sistema GI/G/c/K utilizando procedimientos análogos a los desarrollados en este Capítulo para los sistemas M/G/c/c y GI/M/c.

En este Capítulo, para caracterizar los intereses de los clientes, se han considerado por separado el tamaño de la línea de espera, N_q, y el tiempo de espera en cola, W. Sin embargo, se pueden incluir simultáneamente ambas medidas en la función de coste medio. Para ello, una posibilidad es tener en cuenta lo que se denomina, generalmente, tiempo de espera acumulado en cola, $\tilde{W}$, que es la suma total de los tiempos de espera de los clientes en cola, véase Lillo (2000c), donde se considera el valor de $\tilde{W}$ durante un periodo de ocupación. Por ejemplo, obsérvese que, en un sistema GI/M/c, si hay N_q clientes esperando en cola, el tiempo de espera del cliente n-ésimo, W_n, se distribuye como la suma de n exponenciales de tasa $c\mu$. Por tanto, el valor esperado del tiempo de espera acumulado en cola es,

$$E\left[\tilde{W}\right] = E\left[W_1 + ... + W_{N_q}\right] = E\left[\sum_{n=1}^{N_q} E\left[W_n \mid N_q\right]\right] = E\left[\sum_{n=1}^{N_q} \frac{n}{c\mu}\right] = \frac{1}{c\mu} E\left[N_q^2 + N_q\right] \tag{4.35}$$

Esta cantidad es una alternativa posible para cuantificar simultáneamente el número de personas en cola y el tiempo que estas personas permanecen en espera de su servicio. Sin embargo, en la práctica, puede resultar más complicado establecer su coste. Además, en este caso, es necesario también incluir funciones de pérdida para $\tilde{W}$ puesto que los momentos de su distribución predictiva no son finitos porque tampoco lo son los momentos de la distribución predictiva de N_q.

Como se indicó en la introducción, existen también muchas alternativas a la estructura de coste considerada en este Capítulo. Se podría, por ejemplo, utilizar el criterio del ciclo, véase Lillo (2000c), o considerar horizontes de planificación finito suponiendo sistemas dinámicos en los que varía el número de servidores en función del tiempo. En este último caso, sería necesaria la formulación de funciones de coste dependientes de distribuciones transitorias, que son generalmente difíciles de estimar, como se mostrará en el Capítulo 5, para sistemas con un único servidor.

Se pueden utilizar también otras herramientas de decisión, diferentes de las funciones de coste, para determinar el número de servidores apropiado en el sistema. Por ejemplo, una posibilidad es hacer uso

de las técnicas de decisión multicriterio, véase una recopilación, por ejemplo, en Ríos et al. (1989). Estas estructuras permiten construir funciones objetivo no lineales, de manera que, por ejemplo, no se fije un coste por cliente esperando en cola, sino que el diseñador del sistema pueda establecer una función de utilidad, no necesariamente lineal, según el número de clientes en cola.

Por último, se podrían incorporar procedimientos de optimización sobre otros aspectos del sistema de colas diferente al número de servidores, como por ejemplo, políticas exhaustivas de comportamiento del sistema que implican la inclusión de otro tipo de costes, véase Lillo (2000b) y Lillo (2000d) o considerar clientes impacientes que abandonan el sistema si su tiempo de espera supera un valor determinado, véase Lillo (2001).

Capítulo 5

Estimación del comportamiento transitorio y del periodo de ocupación del sistemas de colas GI/G/1.

En este Capítulo, se describen procedimientos para desarrollar inferencia Bayesiana sobre el comportamiento probabilístico del modelo de colas general, GI/G/1, que dispone de un único servidor y cuyo proceso de llegadas y de servicios están constituidos por sucesiones de variables aleatorias i.i.d. con distribuciones generales y desconocidas definidas en la semirecta real positiva. En este Capítulo, se describe un procedimiento para la estimación de la distribución de la longitud del periodo de ocupación de este modelo de colas general y además se puede realizar inferencia no sólo sobre el comportamiento asintótico, sino sobre el transitorio.

En los Capítulos anteriores a éste, se han propuesto métodos Bayesianos en diferentes modelos de colas que han permitido la estimación de diversas características de interés, tales como el tiempo de espera en cola. Sin embargo, en todos los casos, se ha considerado el valor que toman estas medidas cuando el sistema se encuentra en equilibrio, es decir, evaluadas en el estado estacionario. En este Capítulo, la principal novedad es que se puede estimar el comportamiento del sistema en sus estados transitorios, concretamente, el objetivo es estimar la distribución del número de clientes en el sistema en un instante, τ, así como la distribución del tiempo de espera de un cliente que llega en ese instante τ. Como es de esperar, cuando τ tiende a infinito, estas distribuciones se aproximan a sus distribuciones estacionarias, en caso de ergodicidad. En la práctica, la estimación de las características del sistema cuando éste no se encuentra en equilibrio es a menudo muy importante, como por ejemplo, en situaciones en las que el funcionamiento del sistema se detiene y se reanuda constantemente, o en sistemas en los que la convergencia al estado estacionario es muy lenta.

En este Capítulo, se considera también la longitud del periodo de ocupación que, como se comentó en el Capítulo 1, es otra de las cantidades de mayor interés en los modelos de colas GI/G/1 con aplicaciones en áreas muy diferentes como los sistemas de inventariado o la Teoría General del Seguro. La distribución del periodo de ocupación se analizó también en el Capítulo 3 para el caso particular en el que las llegadas al sistema se producen según un proceso de Poisson. Para ello, se consideraron distribuciones de tipo PH para el tiempo de servicio y se utilizaron las propiedades de los sistemas de colas M/PH/1. En este Capítulo, se pretende estimar la distribución del periodo de ocupación en sistemas GI/G/1 basándose en los resultados introducidos por Bertsimas y Nakazato (1992).

En la Teoría de Colas clásica, tanto el análisis del comportamiento transitorio como el estudio del periodo de ocupación, se han considerado, tradicionalmente, problemas de complejidad elevada. En particular, para

el modelo de colas M/M/1, la distribución transitoria del tamaño del sistema y la distribución del periodo de ocupación se formulan en términos de funciones de Bessel modificadas cuando los parámetros del sistema son conocidos, véase Gross y Harris (1985). Otros resultados sobre el comportamiento transitorio de este sistema de colas se pueden encontrar en Abate y Whitt (1987) y Abate y Whitt (1988). En Bertsimas y Nakazato (1992), se han obtenido resultados importantes sobre el comportamiento transitorio y el periodo de ocupación del sistema MGE/MGE/1. Como ya se ha comentado a lo largo de esta tesis, el modelo de mixtura MGE es muy flexible y permite aproximar arbitrariamente cualquier distribución general definida en la semirecta real positiva, aunque, en algunos casos particulares, se requiera para ello un número elevado de fases, véase la Sección 2.2. Debido a esta versatilidad, Bertsimas y Nakazato (1992) consideran el sistema MGE/MGE/1 para aproximar el modelo de colas GI/G/1 con la particularidad de que obtienen expresiones cerradas para las trasformadas de Laplace de la distribución transitoria del número de clientes en el sistema y del tiempo de espera en cola cuando éste se encuentra inicialmente vacío, y de la distribución del periodo de ocupación. En este Capítulo, se proponen procedimientos Bayesianos basados en estos resultados que permiten estimar estas medidas cuando los parámetros del sistema no son conocidos y se tiene un conjunto de datos procedentes del proceso de llegadas y de servicio al sistema.

Las expresiones de las transformadas obtenidas por Bertsimas y Nakazato (1992) requieren, entre otros cálculos algebraicos, la obtención de las raíces de una ecuación polinómica que depende de los parámetros del proceso de llegadas y de servicio en el sistema de colas. Cuando estos parámetros son conocidos, las raíces mencionadas no se calculan, generalmente, de forma analítica, sino que se obtienen utilizando paquetes simbólicos, tales como Mathematica. Sin embargo, el cálculo simbólico no es factible computacionalmente, cuando se emplean métodos MCMC de dimensión variable y además, ni siquiera es fijo el número de raíces de la ecuación que hay que resolver. Alternativamente, en este Capítulo, se propone una técnica para poder calcular numéricamente y con un coste computacional muy bajo, las raíces de la ecuación en cada iteración MCMC, lo que permite estimar las transformadas de Laplace de las distribuciones de interés. Se requieren también métodos de inversión numérica para aproximar las inversas de las transformadas de Laplace de estas distribuciones.

Este Capítulo se divide en cuatro Secciones y un Apéndice. En la Sección 5.1, se introduce al modelo de colas MGE/MGE/1 y se exponen los resultados relativos a la Teoría de Colas clásica obtenidos por Bertsimas y Nakazato (1992) para este sistema. En la Sección 5.2, se desarrolla inferencia Bayesiana en el modelo de colas considerado. En concreto, se analiza, en primer lugar, el equilibrio del sistema y, a continuación, se proponen diferentes procedimientos de estimación para las distribuciones transitorias del tamaño del sistema y del tiempo de espera en cola, así como la distribución de la longitud del periodo de ocupación, dados los datos observados del proceso de llegadas y de servicio. En la Sección 5.3, se ilustra la metodología desarrollada con diferentes sistemas de colas simulados. Y, por último, en la Sección 5.4, se concluye con unos comentarios y extensiones. Para completar este Capítulo, se incluye, un Apéndice, con una breve explicación de un algoritmo para la inversión numérica de transformadas de Laplace, propuesto por Hosono (1981), que es el que se utilizará a lo largo de este Capítulo.

5.1. Comportamiento transitorio y periodo de ocupación en el sistema de colas MGE/MGE/1.

En esta Sección, se describe el sistema de colas MGE/MGE/1 y se exponen los resultados obtenidos por Bertsimas y Nakazato (1992) sobre el comportamiento transitorio y el periodo de ocupación de este sistema suponiendo conocidos sus parámetros. En la Subsección 5.1.1, se introduce al modelo de colas considerado y la notación empleada. También, se describen algunas de las propiedades del sistema y las transformadas de Laplace de varias cantidades de las que dependerán los resultados de Bertsimas y Nakazato (1992) expuestos en las tres Subsecciones siguientes. En la Subsección 5.1.2, se presenta una expresión cerrada para la transformada de Laplace de la distribución transitoria del número de clientes presentes en el sistema MGE/MGE/1.

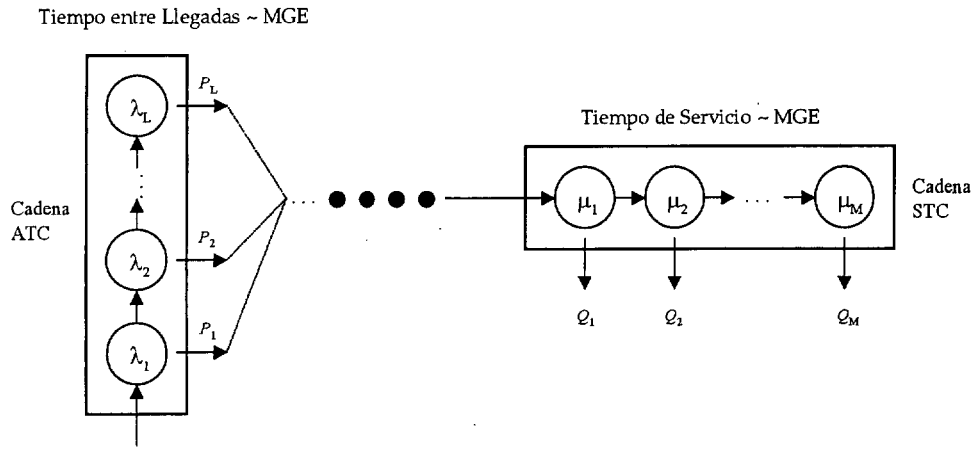

Figura 5.1: Ilustración esquemática del sistema de colas MGE/MGE/1.

En la Subsección 5.1.2, se incluye la expresión para la transformada de Laplace de la distribución transitoria del tiempo de espera en cola en este sistema y, por último, en la Subsección 5.1.3, la transformada de Laplace de la distribución de la longitud del periodo de ocupación.

5.1.1. Descripción y notación del sistema MGE/MGE/1

Se introduce en esta Sección el modelo de colas MGE/MGE/1 que dispone de un único servidor y tiene capacidad infinita. Los tiempos entre las llegadas a este sistema constituyen una sucesión de variables aleatorias independientes e idénticamente distribuidas según una mixtura MGE de parámetros $\theta_\lambda = \{L, \mathbf{P}, \boldsymbol{\lambda}\}$ cuya densidad se detalla en (2.4). Los tiempos de servicio son también independientes e idénticamente distribuidos según el mismo modelo de mixtura MGE, pero con parámetros $\theta_\mu = \{M, \mathbf{Q}, \boldsymbol{\mu}\}$.

La Figura 5.1 ilustra el comportamiento de este sistema de colas en el que el proceso de llegadas se determina con una cadena denominada ATC que está constituida por L fases exponenciales consecutivas de tasas $\lambda_1, ..., \lambda_L$, de modo que, antes de acceder a la línea de espera o, en su caso, al servicio, cada cliente debe atravesar $1, 2, ..$ ó L fases con probabilidad $P_1, P_2, ...$ ó P_L, respectivamente. En el momento en el que un cliente se incorpora al sistema abandonando la cadena ATC, otro cliente llega a la primera de las L fases que constituyen la cadena ATC. El proceso de servicio se determina también con otra una cadena (STC) de M fases exponenciales consecutivas de tasas $\mu_1, ..., \mu_M$, de modo que, el tiempo de servicio que recibe cada cliente se compone de $1, 2, ...$ ó M fases de servicio con probabilidad $Q_1, Q_2, ...$ ó Q_M, respectivamente.

Teniendo en cuenta el valor de la esperanza de una distribución MGE dada en (3.15), se obtiene que la intensidad de tráfico, ρ, del sistema MGE/MGE/1 viene dado por,

$$\rho = \frac{E[S_1]}{E[A_1]} = \frac{\Sigma_{r=1}^{M}\left(1 - \Sigma_{s=1}^{r-1} Q_s\right)\frac{1}{\mu_r}}{\Sigma_{r=1}^{L}\left(1 - \Sigma_{s=1}^{r-1} P_s\right)\frac{1}{\lambda_r}}, \tag{5.1}$$

donde A_1 y S_1 representan las variables aleatorias que definen el tiempo entre las llegadas y el tiempo de servicio en el sistema, respectivamente. La condición de ergodicidad que permite que este modelo de colas sea estable es $\rho < 1$.

Tal como se mostró en la Sección 3.2.1, el modelo de mixtura MGE es una distribución continua de tipo PH con soporte en $[0, \infty)$. Concretamente, la distribución del tiempo entre llegadas admite la siguiente

representación de tipo PH que es de orden L, véase (3.14),

$$\alpha_L = (1,0,...,0)_{(1\times L)}, \qquad T_L = \begin{bmatrix} -\lambda_1 & (1-P_1)\lambda_1 & & \\ & -\lambda_2 & \frac{1-P_1-P_2}{1-P_1}\lambda_2 & \\ & & \ddots & \\ & & & -\lambda_L \end{bmatrix}_{(L\times L)},$$

y la distribución del tiempo de servicio admite otra representación análoga de tipo PH, (α_M, T_M), que es de orden M. Por tanto, la función de densidad del tiempo entre llegadas viene dada por,

$$f_{A_1}(x) = \alpha_L \exp(T_L x)\mathbf{T}_L^0$$

y la transformada de Laplace-Stieljes, definida en (1.16), de la distribución del tiempo entre llegadas viene dada por, véase (3.4),

$$f_{A_1}^*(s) = \alpha_L (sI - T_L)^{-1}\mathbf{T}_L^0 = \sum_{r=1}^{L} P_r \prod_{i=1}^{r}\left(\frac{\lambda_i}{\lambda_i + s}\right). \tag{5.2}$$

Análogamente, se obtiene la función de densidad del tiempo de servicio, $f_{S_1}(x)$, y su transformada de Laplace-Stieljes, $f_{S_1}^*(s)$. Obsérvese que para obtener la expresión de la transformada dada en (5.2) no es necesario calcular la matriz inversa de $(sI - T_L)$, sino que basta con utilizar que la distribución MGE es una mixtura de sumas de exponenciales,

$$f_{A_1}^*(x) = E\left[e^{-sA_1}\right] = \sum_{r=1}^{L} P_r E\left[e^{-s(X_1+...+X_r)}\right] = \sum_{r=1}^{L} P_r \prod_{i=1}^{r}\left(\frac{\lambda_i}{\lambda_i + s}\right). \tag{5.3}$$

Se denota por A_k, con $k = 1,...,L$, a la variable aleatoria que representa el tiempo que le queda a un cliente para acceder a la línea de espera suponiendo que está en la fase k de la cadena ATC. Obsérvese que A_k se distribuye también según el modelo MGE y su densidad viene dada por,

$$f_{A_k}(x) = \mathbf{e}_k \exp(T_L x)\mathbf{T}_L^0$$

donde $\mathbf{e}_k = (0,...,1,...,0)$, un vector unidad en el que todos los elementos son ceros a excepción del uno en la posición k. Además, la transformada de Laplace-Stieljes de A_k viene dada por,

$$f_{A_k}^*(s) = E\left[e^{-sA_k}\right] = \mathbf{e}_k (sI - T_L)^{-1}\mathbf{T}_L^0 = \sum_{r=k}^{L} \frac{P_r}{1-\Sigma_{s=1}^{k-1}P_s} \prod_{i=k}^{r}\left(\frac{\lambda_i}{\lambda_i + s}\right), \qquad \text{para } k = 1,...,L. \tag{5.4}$$

Se pueden obtener expresiones equivalentes para $f_{S_k}(x)$ y $f_{S_k}^*(s)$, donde S_k representa el tiempo de servicio restante de un cliente que se encuentra en la fase k de servicio, con $k = 1,...,M$.

Por último, se denota por $A_{k,r}(x)$ a la probabilidad de que un cliente de la fase ATC pase de la fase k a la fase r durante el tiempo x sin que se produzca una nueva llegada. Obsérvese que la probabilidad $A_{k,r}(x)$ es cero si $k > r$. El vector que contiene todas las probabilidades, $A_{k,r}(x)$, para $r = k,..,L$, viene dado por,

$$\mathbf{A}_k(x) = (0,...,0,A_{k,k}(x),...,A_{k,L}(x)) = \mathbf{e}_k \exp(T_L x), \qquad \text{para } k = 1,...,L,$$

y el vector constituido por sus transformadas de Laplace, véase (1.17), es,

$$\mathbf{A}_k^*(s) = \left(0,...,0,A_{k,k}^*(s),...,A_{k,L}^*(s)\right) = \mathbf{e}_k (sI - T_L)^{-1}, \qquad \text{para } k = 1,...,L, \tag{5.5}$$

donde,

$$A_{k,r}^*(s) = \int_0^\infty e^{-sx} A_{k,r}(x)\,dx = \frac{1-\Sigma_{s=1}^{r-1}P_s}{1-\Sigma_{s=1}^{k-1}P_s} \frac{\prod_{i=k}^{r-1}\lambda_i}{\prod_{i=k}^{r}(\lambda_i + s)}, \qquad \text{para } r = k,...,L. \tag{5.6}$$

Se pueden obtener expresiones análogas para $\mathbf{S}_k^*(s)$ y $S_{k,r}^*(s)$ referentes a la distribución MGE de servicio.

Los resultados obtenidos por Bertsimas y Nakazato (1992) que se incluyen en las tres Subsecciones siguientes dependen directamente de las transformadas dadas en (5.2), (5.4), (5.5), y (5.6), así como de las expresiones análogas para las transformadas referentes al proceso de servicio.

5.1.2. Distribución transitoria del número de clientes presentes en el sistema.

En esta Subsección, se incluye una expresión cerrada, obtenida por Bertsimas y Nakazato (1992), para la transformada de Laplace de la distribución del número de clientes, $N(\tau)$, presentes en el sistema MGE/MGE/1 en el instante τ, asumiendo que el sistema está vacío en el momento inicial, $\tau = 0$.

Se denota por $R_a(\tau)$ a la fase de la cadena ATC en la que se encuentra un cliente que está llegando al sistema en el instante τ y por $R_s(\tau)$ a la fase de la cadena STC en la que se encuentra un cliente que está siendo servido en el instante τ. Con estas definiciones, se puede formular el sistema como una cadena de Markov en tiempo continuo con espacio de estados infinito dado por,

$$\{(N(\tau), R_a(\tau), R_s(\tau)), \quad N(\tau) = 0,1,...; \quad R_a(\tau) = 1,..,L; \quad R_s(\tau) = 1,...,M.\}.$$

Se denota por,

$$\Pi_n(\tau) = \Pr(N(\tau) = n), \qquad \text{para } n = 0,1,..., \tag{5.7}$$

que es la suma de los elementos del vector, $\boldsymbol{\pi}_n(\tau)$, dados por las probabilidades,

$$\pi_{n,i,j}(\tau) = \Pr(N(\tau) = n,\ R_a(\tau) = i,\ R_s(\tau) = j), \qquad \text{si } n \geq 1,$$

ordenados del modo siguiente,

$$\boldsymbol{\pi}_n(\tau) = (\pi_{n,1,1}(\tau), ..., \pi_{n,L,1}(\tau), ..., \pi_{n,1,M}(\tau), ..., \pi_{n,L,M}(\tau)),$$

o (5.7) es la suma de los elementos del vector, $\boldsymbol{\pi}_0(\tau)$, dados por,

$$\pi_{0,i}(\tau) = \Pr(N(\tau) = 0,\ R_a(\tau) = i), \qquad \text{si } n = 0. \tag{5.8}$$

Suponiendo que el sistema está inicialmente vacío, es decir, que la únicas probabilidades distintas de cero son $\pi_{0,k}(0)$ para todo k, y que $\rho < 1$, Bertsimas y Nakazato (1992) demuestran que las transformadas de Laplace $\boldsymbol{\pi}_n^*(s)$ y $\boldsymbol{\pi}_0^*(s)$ de $\boldsymbol{\pi}_n(\tau)$ y $\boldsymbol{\pi}_0(\tau)$, respectivamente, vienen dadas por,

$$\boldsymbol{\pi}_n^*(s) = \sum_{r=1}^{M} D_r \mathbf{S}_1^*(x_r(s)) \otimes \mathbf{A}_1^*(s - x_r(s)) \left(f_{A_1}^*(s - x_r(s))\right)^{n-1}, \qquad \text{para } n \geq 1, \tag{5.9}$$

$$\boldsymbol{\pi}_0^*(s) = \sum_{k=1}^{L} \pi_{0,k}(0)\, \mathbf{A}_k^*(x_r(s)) + \sum_{r=1}^{M} D_r \frac{1}{x_r(s)} f_{S_1}^*(x_r(s)) \left(\mathbf{A}_1^*(s - x_r(s)) - \mathbf{A}_1^*(s)\right),$$

donde,

$$D_r = \frac{\sum_{k=1}^{L} \pi_{0,k}(0) f_{A_k}^*(x_r(s))}{1 - f_{A_1}^*(s)} \frac{(-1)^M S_{1,M}^*(0)}{S_{1,M}^*(x_r(s))} x_r(s) \prod_{\substack{k=1 \\ k \neq r}}^{M} \frac{x_r(s)}{x_r(s) - x_k(s)} \tag{5.10}$$

y $x_r(s) \equiv x$, con $r = 1,..., M$, son las M raíces de la ecuación,

$$\left\{ \begin{array}{c} f_{A_1}^*(s - x) f_{S_1}^*(x) = 1, \\ \mathrm{Re}(x) > 0 \text{ para } \mathrm{Re}(s) > 0, \end{array} \right\} \tag{5.11}$$

y donde el producto tensorial de dos vectores, $\mathbf{u}$ y $\mathbf{v}$, se define del modo siguiente,

$$\mathbf{u} \otimes \mathbf{v} = (u_1, ..., u_n) \otimes (v_1, ..., v_m) = (u_1 v_1, ..., u_n v_1, ..., u_1 v_m, ..., u_n v_m).$$

Obsérvese que $\boldsymbol{\pi}_n^*(s)$ es un vector cuyos elementos son las transformadas de las probabilidades, $\pi_{n,i,j}(\tau)$, dependiendo del instante τ en el que son evaluadas, y que vienen dadas por,

$$\pi_{n,i,j}^*(s) = \int_0^\infty e^{-s\tau} \pi_{n,i,j}(\tau)\, d\tau = \sum_{r=1}^{M} D_r S_{1,j}^*(x_r(s))\, A_{1,i}^*(s - x_r(s)) \left(f_{A_1}^*(s - x_r(s))\right)^{n-1}, \tag{5.12}$$

para $i = 1, ..., L$, y $j = 1, ..., M$. Análogamente, $\pi_0^*(s)$ es un vector cuyos L elementos son las transformadas de las probabilidades, $\pi_{0,i}(\tau)$, dadas por,

$$\pi_{0,i}^*(s) = \int_0^\infty e^{-s\tau}\pi_{0,i}(\tau)\,d\tau = \sum_{k=1}^{L}\pi_{0,k}(0)\,A_{k,i}^*(x_r(s)) + \sum_{r=1}^{M}\frac{D_r f_{S_1}^*(x_r(s))}{x_r(s)}\left(A_{1,i}^*(s - x_r(s)) - A_{1,i}^*(s)\right), \tag{5.13}$$

para $i = 1, ..., L$. Conocidos los parámetros de la distribución del tiempo entre llegadas, $\boldsymbol{\theta}_\lambda = \{L, \mathbf{P}, \boldsymbol{\lambda}\}$, y de la distribución de servicio, $\boldsymbol{\theta}_\mu = \{M, \mathbf{Q}, \boldsymbol{\mu}\}$, del sistema MGE/MGE/1, se pueden obtener los valores exactos de D_r, $\pi_{n,i,j}^*(s)$ y $\pi_{0,i}^*(s)$ utilizando las expresiones dadas en (5.4) y (5.6).

Por último, teniendo en cuenta que la probabilidad, $\Pr(N(\tau) = n)$, véase (5.7), viene dada por,

$$\Pi_n(\tau) = \begin{cases} \sum_{i=1}^{L}\pi_{0,i}(\tau), & \text{si } n = 0, \\ \sum_{i=1}^{L}\sum_{j=1}^{M}\pi_{n,i,j}(\tau), & \text{si } n \geq 1, \end{cases} \tag{5.14}$$

y utilizando las propiedades de las transformadas de Laplace, véase, por ejemplo, Nelson (1995), se tiene que,

$$\Pi_n^*(s) = \int_0^\infty e^{-s\tau}\Pi_n(\tau)\,d\tau = \begin{cases} \sum_{i=1}^{L}\pi_{0,i}^*(s), & \text{si } n = 0, \\ \sum_{i=1}^{L}\sum_{j=1}^{M}\pi_{n,i,j}^*(s), & \text{si } n \geq 1, \end{cases} \tag{5.15}$$

Para calcular la distribución del número de clientes, $N(\tau)$, presentes en el sistema MGE/MGE/1 en un instante, τ, conocidos los parámetros, es necesario utilizar un algoritmo de inversión numérica de transformadas de Laplace que permita invertir la transformada $\Pi_n^*(s)$, dada en (5.15). En este Capítulo, se ha utilizado, al igual que Bertsimas y Nakazato (1992), el algoritmo de inversión propuesto por Hosono (1981), que se explica brevemente en el Apéndice 5.5 y es esencialmente equivalente al algoritmo EULER introducido por Dubner y Abate (1968). Básicamente, los pasos a seguir para calcular las probabilidades, $\Pr(N(\tau) = n)$, son los siguientes.

1. Para cada valor de τ, fijar varios valores de s según se indica en el Apéndice 5.5.
2. Obtener las M raíces de la ecuación (5.11) para los valores fijados de s.
3. Evaluar la cantidad $\Pi_n^*(s)$ a partir de $\pi_{0,i}^*(s)$ y $\pi_{n,i,j}^*(s)$, como se muestra en (5.15).
4. Por último, se aproxima el valor de $\Pi_n(\tau)$ a partir de los valores de $\Pi_n^*(s)$, tal como se detalla en el Apéndice 5.5.

Nótese que estos son los pasos a seguir cuando se conocen los parámetros del sistema. Si este no es el caso, como se asume más adelante, y se utilizan procedimientos de aproximación Monte Carlo para estimar $\Pr(N(\tau) = n)$, es importante desarrollar estos pasos de la manera más eficiente posible para evitar que se incremente demasiado el coste computacional. En particular, en el paso 3, se puede reducir considerablemente el tiempo de computación para la obtención de las raíces de interés, como se describirá más adelante, en la Sección 5.2.2.1.

Por otro lado, obsérvese que, aunque la distribución transitoria de $N(\tau)$ existe sin que se verifique la condición de ergodicidad, $\rho < 1$, en Bertsimas y Nakazato (1992), se asume esta condición cuando se obtienen las transformadas dadas en (5.9) puesto que es condición suficiente para asegurar que el número de raíces de la ecuación (5.11) sea igual a M, que es el orden de la distribución del tiempo de servicio.

5.1.3. Distribución transitoria del tiempo de espera en cola.

En esta Subsección, se considera la distribución del tiempo de espera en cola, $W(\tau)$, de un cliente que llega al sistema MGE/MGE/1 en el instante τ. Se supone, al igual que antes, que no hay clientes presentes en el sistema en el instante inicial, $\tau = 0$. En Bertsimas y Nakazato (1992), se obtiene una expresión para la transformada de la función de distribución de $W(\tau)$ asumiendo que en el instante inicial el vector de probabilidades, $\boldsymbol{\pi}_0(0)$, véase (5.8), que indica la fase en la que se encuentra el cliente que llega, verifica que,

$$\boldsymbol{\pi}_0(0) = \frac{\mathbf{A}_1^*(0)}{E[A_1]}. \tag{5.16}$$

En Bertsimas y Nakazato (1992), se demuestra que esta condición implica que el proceso de llegadas está en el estado estacionario desde el instante inicial, es decir,

$$\Pr\{R_a(0) = i\} = \Pr\{R_a(\tau) = i\},$$

para ello, se comprueba que el valor de $\boldsymbol{\pi}_0(0)$ verificando la condición (5.16) es la solución estacionaria de la ecuación de Kolmogorov que describe el proceso de llegadas.

Con estas condiciones y, suponiendo que $\rho < 1$, Bertsimas y Nakazato (1992) demuestran que, para cada valor de w, la transformada de Laplace de la probabilidad,

$$\Gamma_w(\tau) = \Pr(W(\tau) \leq w), \tag{5.17}$$

viene dada por,

$$\begin{aligned}\Gamma_w^*(s) &= \int_0^\infty e^{-s\tau} \Pr(W(\tau) \leq w)\, d\tau, \\ &= \frac{1}{s} + \sum_{r=1}^{M} \frac{(-1)^M}{s} \frac{S_{1,M}(0)}{S_{1,M}(x_r(s))} \left(\prod_{\substack{k=1 \\ k \neq r}}^{M} \frac{x_r(s)}{x_r(s) - x_k(s)} \right) e^{x_r(s)w},\end{aligned} \tag{5.18}$$

donde $S_{1,M}(s)$ se obtiene a partir de la transformada dada en (5.6), pero expresada en términos de los parámetros de servicio, es decir,

$$S_{1,M}^*(s) = \left(1 - \Sigma_{s=1}^{M-1} Q_s\right) \frac{\prod_{i=1}^{r-1} \mu_i}{\prod_{i=1}^{r} (\mu_i + s)}$$

y donde $x_r(s)$, con $r = 1, ..., M$, son las M raíces de la ecuación dada en (5.11) que intervienen también en la transformada del tamaño del sistema introducida en la Subsección anterior.

Obsérvese además que, si se verifica la condición (5.16), los elementos del vector de probabilidades inicial, $\boldsymbol{\pi}_0(0)$, definidos en (5.8), vienen dados por,

$$\pi_{0,i}(0) = \frac{\left(1 - \Sigma_{s=1}^{i-1} P_s\right) \frac{1}{\lambda_i}}{\Sigma_{r=1}^{L} \left(1 - \Sigma_{s=1}^{r-1} P_s\right) \frac{1}{\lambda_r}}, \qquad \text{para } i = 1, ..., L. \tag{5.19}$$

5.1.4. Distribución de la longitud del periodo de ocupación.

Bertsimas y Nakazato (1992) obtienen una expresión muy sencilla para la transformada de Laplace-Stieljes, $f_B^*(s)$, de la función de densidad de la longitud del periodo de ocupación, B, en el sistema MGE/MGE/1. Concretamente, se demuestra que,

$$f_B^*(s) = E\left[e^{-sB}\right] = 1 - \left(1 - f_{S_1}^*(s)\right) \frac{\prod_{k=1}^{M} (s + \mu_k)}{\prod_{r=1}^{M} (s - x_r(s))},$$

donde, nuevamente, $x_r(s)$, con $r = 1, ..., M$, son las M raíces de la ecuación dada en (5.11).

Nótese que este resultado simplifica considerablemente la caracterización obtenida por Ramaswami (1982) para la distribución del periodo de ocupación del sistema G/PH/1 utilizando los métodos matriciales. No obstante, en Ramaswami (1982) se derivan resultados que no requieren asumir que los tiempos entre las llegadas al sistema sean independientes.

Obsérvese también que, utilizando las propiedades de las transformadas de Laplace, véase, por ejemplo, Nelson (1995), se tiene que la transformada de Laplace de la función de distribución del periodo de ocupación viene dada por,

$$F_B^*(s) = \int_0^\infty e^{-sx} \Pr(B \leq x)\, dx = \frac{f_B^*(s)}{s} = \frac{1}{s} - \frac{1 - f_{S_1}^*(s)}{s} \frac{\prod_{k=1}^{M}(s+\mu_k)}{\prod_{r=1}^{M}(s - x_r(s))} \tag{5.20}$$

5.2. Inferencia Bayesiana para el sistema MGE/MGE/1.

En esta Sección, se desarrolla inferencia Bayesiana para estimar la distribución transitoria del número de clientes en el sistema y del tiempo de espera en cola, así como la distribución del periodo de ocupación en un modelo de colas MGE/MGE/1, teniendo como datos las observaciones del proceso de llegadas y de servicio, $\{\mathbf{t}, \mathbf{s}\}$. En esta Sección, se supone que se tiene una muestra MCMC de la distribución a posteriori de los parámetros de la distribución del tiempo entre llegadas y del tiempo de servicio en el sistema, obtenida a partir de los algoritmos `RJMGE` ó `BDMGE` propuestos en el Capítulo 2.

Se distinguen dos problemas fundamentales cuando se pretende incorporar los resultados de la Sección anterior dentro de un algoritmo MCMC.

1. Cálculo numérico de las raíces de la ecuación (5.11).

   Para cada valor de los parámetros de la muestra Monte Carlo, se necesita calcular el valor de las transformadas, $\Pi_n^*(s)$, $\Gamma_w^*(s)$ y $F_B^*(s)$, y para ello, es necesario obtener las raíces de la ecuación (5.11) en cada iteración MCMC. Es evidente que, en un algoritmo MCMC, no es factible computacionalmente calcular las raíces de modo simbólico, como se sugiere en Bertsimas y Nakazato (1992), utilizando, por ejemplo, el paquete Mathematica. Además, obsérvese que, puesto que se están utilizando algoritmos de dimensión paramétrica variable, el número de raíces de la ecuación (5.11) varía en cada iteración MCMC.

2. Inversión numérica de las transformadas de Laplace.

   Cuando los parámetros son conocidos, no se dispone de una expresión explícita para las distribuciones de interés, sino de sus transformadas de Laplace. Por tanto, es importante la elección de un procedimiento adecuado en el que se integren los métodos de inversión numérica de transformadas de Laplace que permita estimar las distribuciones predictivas de $N(\tau)$, $W(\tau)$ y B, sin que se incremente demasiado el coste computacional.

En la Subsección 5.2.1, se describe brevemente cómo desarrollar inferencia Bayesiana sobre la intensidad de tráfico, ρ, dada en (5.1), de un modo equivalente al considerado en Capítulos anteriores. En la Subsección 5.2.2, se considera la estimación de las distribuciones predictivas $N(\tau)$, $W(\tau)$ y B. En primer lugar, en el apartado 5.2.2.1, se introduce un procedimiento para obtener numéricamente las raíces de la ecuación (5.11) en cada iteración MCMC a partir de los coeficientes de un polinomio que tiene las mismas raíces que la ecuación de interés. Este procedimiento se utiliza en el apartado 5.2.2.2, donde se describen y se comparan dos métodos diferentes para estimar las distribuciones predictivas de $N(\tau)$, $W(\tau)$ y B.

5.2.1. Análisis del equilibrio del sistema.

Se ilustra a continuación el procedimiento de estimación de la probabilidad a posteriori de que el sistema de colas observado sea ergódico, es decir, la probabilidad de que ρ sea menor que 1. Como se ha comentado en Capítulos anteriores, si $\rho < 1$, el sistema observado es estable y por tanto, existen las distribuciones estacionarias de las medidas asociadas al sistema, que son las distribuciones a las que convergen las distribuciones transitorias cuando τ tiende a infinito. En caso contrario, si $\rho \geq 1$, las probabilidades transitorias de las cantidades de interés, evaluadas en cualquier instante τ, convergen a cero cuando τ tiende a infinito y por tanto, no existen sus distribuciones estacionarias.

Se supone que utilizando los datos, $\{\mathbf{t}, \mathbf{s}\}$, observados del proceso de llegadas y de servicio, se ha obtenido una muestra MCMC de la distribución a posteriori de los parámetros, $\{\theta_\lambda, \theta_\mu\}$, del sistema MGE/MGE/1, introducidos en la Sección 5.1. Es decir, se tiene una colección de valores,

$$\theta^{(j)} = \left\{\theta_\lambda^{(j)}, \theta_\mu^{(j)}\right\} = \left\{\left(L^{(j)}, \mathbf{P}^{(j)}, \boldsymbol{\lambda}^{(j)}\right), \left(M^{(j)}, \mathbf{Q}^{(j)}, \boldsymbol{\mu}^{(j)}\right)\right\}, \qquad \text{para } j = 1, ..., J, \tag{5.21}$$

donde J es el tamaño de la muestra MCMC. La probabilidad a posteriori de que el sistema de colas sea ergódico se puede estimar con,

$$P(\rho < 1 \mid \mathbf{t}, \mathbf{s}) \approx \frac{1}{J} \# \left\{\rho^{(j)} < 1\right\}, \tag{5.22}$$

donde $\rho^{(j)}$ es el valor de la intensidad de tráfico dado en (5.1) evaluado para cada valor de la muestra MCMC que viene dado por,

$$\rho^{(j)} = \frac{\Sigma_{r=1}^{M^{(j)}} \left(1 - \Sigma_{s=1}^{r-1} Q_s^{(j)}\right) \frac{1}{\mu_r^{(j)}}}{\Sigma_{r=1}^{L^{(j)}} \left(1 - \Sigma_{s=1}^{r-1} P_i^{(j)}\right) \frac{1}{\lambda_r^{(j)}}}. \tag{5.23}$$

Como siempre, si $P(\rho < 1 \mid \mathbf{t}, \mathbf{s})$ es lo suficientemente grande (usualmente, mayor que 0.8), se puede asumir que el sistema es ergódico. Sin embargo, aunque el sistema sea estable, generalmente, es conveniente la estimación de las distribuciones transitorias del sistema, ya que la convergencia al estado estacionario puede ser muy lenta, o también, para predecir el comportamiento transitorio desde el instante inicial en otro sistema equivalente al observado.

Procediendo de un modo análogo al considerado en otros sistemas de colas, una estimación de la intensidad de tráfico viene dada por,

$$E[\rho \mid \mathbf{t}, \mathbf{s}] \approx \frac{1}{J} \sum_{j=1}^{J} \frac{\Sigma_{r=1}^{M^{(j)}} \left(1 - \Sigma_{s=1}^{r-1} Q_s^{(j)}\right) \frac{1}{\mu_r^{(j)}}}{\Sigma_{r=1}^{L^{(j)}} \left(1 - \Sigma_{s=1}^{r-1} P_i^{(j)}\right) \frac{1}{\lambda_r^{(j)}}}. \tag{5.24}$$

5.2.2. Estimación de las distribuciones de interés.

Una vez analizado el equilibrio del sistema, el objetivo es ahora estimar las distribuciones predictivas de $N(\tau)$ y de $W(\tau)$, así como la distribución predictiva de B, teniendo como datos las observaciones, $\{\mathbf{t}, \mathbf{s}\}$. Concretamente, se pretende obtener estimaciones de,

$$\Pi_n(\tau \mid \mathbf{t}, \mathbf{s}) = \Pr(N(\tau) = n \mid \mathbf{t}, \mathbf{s}), \quad \Gamma_w(\tau \mid \mathbf{t}, \mathbf{s}) = \Pr(W(\tau) \leq w \mid \mathbf{t}, \mathbf{s}), \quad F_B(x \mid \mathbf{t}, \mathbf{s}) = \Pr(B \leq x \mid \mathbf{t}, \mathbf{s}).$$

Se proponen en esta Sección, dos procedimientos de estimación MCMC basados en los resultados de Bertsimas y Nakazato (1992), expuestos en la Sección 5.1. Como se verá más adelante, en ambos procedimientos se requiere calcular la expresión de las siguientes transformadas, cuyo valor viene dado en (5.15), (5.18) y (5.20),

$$\Pi_n^*\left(s \mid \theta^{(j)}\right), \qquad \Gamma_w^*\left(s \mid \theta^{(j)}\right), \qquad F_B^*\left(s \mid \theta^{(j)}\right),$$

para cada valor de los parámetros, $\theta^{(j)}$, dados en (5.21), de la muestra MCMC. Puesto que, como se mostró en la Sección 5.1, para evaluar estas transformadas, es necesario previamente obtener las raíces de la ecuación (5.11), se propone en primer lugar, en el apartado siguiente, un procedimiento que permite resolver numéricamente esta ecuación para cada valor de los parámetros $\theta^{(j)}$.

5.2.2.1. Cálculo numérico de las raíces de la ecuación (5.11).

El objetivo en este apartado es describir un procedimiento para obtener las M raíces de la ecuación (5.11) para cada valor de los parámetros del sistema $\theta^{(j)}$, dados en (5.21); es decir, dados unos parámetros,

$$\theta = \{L, \mathbf{P} = (P_1, ..., P_L), \boldsymbol{\lambda} = (\lambda_1, ..., \lambda_L), M, \mathbf{Q} = (Q_1, ..., Q_M), \boldsymbol{\mu} = (\mu_1, ..., \mu_M)\},$$

el problema a resolver es calcular numéricamente las raíces de la ecuación,

$$f^*_{A_1}(s-x) f^*_{S_1}(x) = 1, \tag{5.25}$$

que verifican que su parte real es positiva. Bertsimas y Nakazato (1992) demuestran que el número de raíces que verifican esta condición es M, que es a su vez el orden de la distribución MGE de servicio.

Se podría pensar en utilizar un método numérico para calcular directamente las raíces de la ecuación (5.25), como por ejemplo, el método de Newton-Raphson. Sin embargo, como se prueba en el siguiente teorema, se trata de una ecuación polinómica cuyas raíces coinciden con las de un polinomio con una estructura determinada y por tanto, resulta más razonable utilizar un método especialmente diseñado para calcular los ceros de una función polinómica, como por ejemplo, el método de Laguerre, véase, por ejemplo, Ralston y Rabinowitz (1978). Este es el método que se ha utilizado en este Capítulo haciendo uso de una rutina implementada en FORTRAN, véase Press et al. (1989).

Teorema 1 *El problema de calcular las raíces de la ecuación (5.25) es equivalente a extraer las M raíces con parte real positiva del polinomio de orden $L+M$ dado por,*

$$\mathcal{P}(x) = \left[\sum_{r=1}^{L}\sum_{t=1}^{M} P_r Q_t \prod_{i=1}^{r}\lambda_i \prod_{i=r+1}^{L}(\lambda_i + s - x)\prod_{j=1}^{t}\mu_j \prod_{j=t+1}^{M}(\mu_j + x)\right] - \left[\prod_{i=1}^{L}(\lambda_i + s - x)\right] \times \left[\prod_{j=1}^{M}(x + \mu_j)\right], \tag{5.26}$$

cuyo coeficiente de orden n, para $n = 0, ..., L+M$,

$$a_n = \sum_{k=0}^{n}(-1)^{n-k}\left[\sum_{r=1}^{L}\sum_{t=1}^{M} P_r Q_t \left(\prod_{i=1}^{r}\lambda_i \prod_{j=1}^{t}\mu_j\right) \sum_{\substack{\mathbf{i}(r,n-k)\\ \mathbf{j}(t,k)}} \prod_{\substack{i=i_s\\ j=j_s}} (\lambda_i + s)(\mu_j + x) - \sum_{\substack{\mathbf{i}(0,n-k)\\ \mathbf{j}(0,k)}} \prod_{\substack{i=i_s\\ j=j_s}} (\lambda_i + s)(\mu_j + x)\right], \tag{5.27}$$

donde $\mathbf{i}(r,k)$ son las $\binom{L-r}{L-r-k}$ combinaciones de los $(L-r)$ elementos del conjunto $\{(\lambda_i + s)\}_{i=r}^{L}$ tomados de $(L-r-k)$ en $(L-r-k)$ y, análogamente, $\mathbf{j}(t,k)$ son las $\binom{M-t}{M-t-k}$ combinaciones de los $(M-t)$ elementos del conjunto $\{(\mu_i + x)\}_{j=t}^{M}$ tomados de $(M-t-k)$ en $(M-t-k)$.

Demostración. Puesto que $f^*_{A_1}$ y $f^*_{S_1}$ son las transformadas de Laplace-Stieljes de la distribución del tiempo entre llegadas y de servicio, respectivamente, que se obtienen a partir de (5.2), la ecuación (5.25) resulta ser equivalente a,

$$\sum_{r=1}^{L}\sum_{t=1}^{M} P_r Q_t \prod_{i=1}^{r}\left(\frac{\lambda_i}{\lambda_i + s - x}\right) \times \prod_{j=1}^{t}\left(\frac{\mu_j}{\mu_j + x}\right) = 1, \tag{5.28}$$

y multiplicando ambos lados de esta ecuación por el término,

$$\left[\prod_{i=1}^{L}(\lambda_i+s-x)\right]\times\left[\prod_{j=1}^{M}(x+\mu_j)\right],$$

se tiene que, la ecuación (5.28) viene dada por,

$$\left[\sum_{r=1}^{L}\sum_{t=1}^{M}P_rQ_t\prod_{i=1}^{r}\lambda_i\prod_{i=r+1}^{L}(\lambda_i+s-x)\prod_{j=1}^{t}\mu_j\prod_{j=t+1}^{M}(\mu_j+x)\right]=\left[\prod_{i=1}^{L}(\lambda_i+s-x)\right]\times\left[\prod_{j=1}^{M}(x+\mu_j)\right],$$

cuyos ceros coinciden con los del polinomio $\mathcal{P}(x)$, dado en (5.26). Por tanto, la ecuación (5.25) tiene $L+M$ raíces, coincidentes con las del polinomio $\mathcal{P}(x)$, (5.26), de las cuales M tienen parte real positiva, que son las M raíces de interés.

Se pretende ahora encontrar los coeficientes del polinomio $\mathcal{P}(x)$, (5.26), lo que permitirá hacer uso del método de Laguerre para extraer sus raíces. Para ello, una posibilidad es considerar el desarrollo de Taylor de $\mathcal{P}(x)$ en $x=0$,

$$\mathcal{P}(x)=\mathcal{P}(0)+\mathcal{P}'(0)x+\frac{\mathcal{P}''(0)}{2!}x^2+...+\frac{\mathcal{P}^{(L+M)}(0)}{n!}x^{L+M},$$

y observar que el coeficiente de orden n es igual a $\mathcal{P}^{(n)}(0)/n!$, para $n=0,...,L+M$, donde $\mathcal{P}^{(n)}(0)$ representa la derivada n–ésima de $\mathcal{P}(x)$ evaluada en $x=0$. Obsérvese que,

$$\frac{\mathcal{P}^{(n)}(0)}{n!}=\left[\sum_{r=1}^{L}\sum_{t=1}^{M}P_rQ_t\left(\prod_{i=1}^{r}\lambda_i\prod_{j=1}^{t}\mu_j\right)\times\frac{f_{r,t}^{(n)}(0)}{n!}\right]-\frac{f_{0,0}^{(n)}(0)}{n!},\tag{5.29}$$

donde,

$$f_{r,t}(x)=g_r(s-x;\boldsymbol{\lambda})\times g_t(x;\boldsymbol{\mu})\tag{5.30}$$

y donde,

$$g_r(x;\boldsymbol{\lambda})=\prod_{i=r+1}^{L}(\lambda_i+x),\tag{5.31}$$

y análogamente para $g_t(x;\boldsymbol{\mu})$. Además, la derivada n–ésima de (5.30) viene dada por,

$$f_{r,t}^{(n)}(x)=\sum_{k=0}^{n}(-1)^{n-k}\binom{n}{k}g_r^{(n-k)}(s-x;\boldsymbol{\lambda})\,g_t^{(k)}(x;\boldsymbol{\mu}),$$

y por tanto, las derivadas, $\mathcal{P}^{(n)}(0)/n!$, dadas en (5.29), dependen de,

$$\frac{f_{r,t}^{(n)}(0)}{n!}=\sum_{k=0}^{n}(-1)^{n-k}\frac{g_r^{(n-k)}(s;\boldsymbol{\lambda})}{(n-k)!}\frac{g_t^{(k)}(0;\boldsymbol{\mu})}{k!},\tag{5.32}$$

que se puede calcular teniendo en cuenta que,

$$g_r^{(n)}(x;\boldsymbol{\lambda})=\frac{\partial^n}{\partial x^n}\prod_{i=r+1}^{L}(\lambda_i+x)=n!\sum_{\mathbf{i}(r,n)}\left(\prod_{i=i_s}(\lambda_i+x)\right),\tag{5.33}$$

donde $\mathbf{i}(r,n)=(i_1,i_2,...,i_{L-r-n})$ son las $\binom{L-r}{L-r-n}$ combinaciones de los $(L-r)$ elementos del conjunto $\{(\lambda_i+x)\}_{i=r}^{L}$ tomados de $(L-r-n)$ en $(L-r-n)$.

Sustituyendo (5.32) en (5.29), utilizando (5.33), se tiene,

$$\frac{\mathcal{P}^{(n)}(0)}{n!} = \left[\sum_{r=1}^{L}\sum_{t=1}^{M} P_r Q_t \left(\prod_{i=1}^{r}\lambda_i \prod_{j=1}^{t}\mu_j\right) \sum_{k=0}^{n}(-1)^{n-k} \sum_{\mathrm{i}(r,n-k)} \left(\prod_{i=i_s}(\lambda_i+x)\right) \sum_{\mathrm{i}(t,k)} \left(\prod_{i=i_s}(\mu_i+x)\right)\right] - \sum_{k=0}^{n}(-1)^{n-k} \sum_{\mathrm{i}(0,n-k)} \left(\prod_{i=i_s}(\lambda_i+x)\right) \sum_{\mathrm{i}(0,k)} \left(\prod_{i=i_s}(\mu_i+x)\right),$$

que equivale a la expresión dada en (5.27). ■

5.2.2.2. Predicción del comportamiento transitorio y del periodo de ocupación.

En este apartado, se proponen procedimientos de estimación Monte Carlo para las distribuciones predictivas de las cantidades de interés, $N(\tau)$, $W(\tau)$ y B, en los que se utiliza el cálculo numérico de raíces desarrollado en el apartado anterior. Con esta finalidad, se distinguen dos modos diferentes de actuación basados en los dos procedimientos introducidos en Conesa (2000) y en Armero y Conesa (2000) para estimar, entre otras cantidades, la distribución estacionaria del tiempo de espera en cola en un sistema $\mathrm{M}^X/\mathrm{M}/1$ a partir de su transformada de Laplace-Stieljes, conocida para cada valor de los parámetros del sistema

Si se supone, por ejemplo, que se desea estimar la distribución predictiva de $N(\tau)$, es decir, las probabilidades a posteriori,

$$\Pi_n(\tau \mid \mathbf{t},\mathbf{s}) = \Pr(N(\tau)=n \mid \mathbf{t},\mathbf{s}), \qquad \text{para } n=0,1,...,$$

se plantean dos procedimientos para la estimación de estas cantidades.

1. Primer procedimiento.

   Consiste en actuar de forma análoga a como se ha considerado en Capítulos anteriores, utilizando una aproximación Monte Carlo de la probabilidad de interés, es decir,

   $$\Pi_n(\tau \mid \mathbf{t},\mathbf{s},\rho<1) \approx \frac{1}{R}\sum_{j;\rho^{(j)}<1} \Pi_n\left(\tau \mid \theta^{(j)}\right), \tag{5.34}$$

   donde cada valor de $\Pi_n(\tau \mid \theta^{(j)})$ se obtiene invirtiendo numéricamente su transformada $\Pi_n^*(s \mid \theta^{(j)})$, dada en (5.15), para cada valor de los parámetros, $\theta^{(j)}$, de la muestra MCMC, dada en (5.21), utilizando, por ejemplo, el algoritmo de Hosono, véase el Apéndice 5.5.

   Obsérvese que, aunque la distribución transitoria predictiva, $\Pi_n(\tau \mid \mathbf{t},\mathbf{s})$, existe sin necesidad de asumir equilibrio, se ha condicionado sobre $\rho<1$, en (5.34), puesto que, como se comentó en la Sección 5.1, Bertsimas y Nakazato (1992) demuestran que esta condición es suficiente para asegurar que el número de raíces de la ecuación (5.11) sea M. Nótese que en (5.34), el valor de $\rho^{(j)}$ viene dado por (5.23) y R representa el tamaño de la submuestra MCMC formada por el conjunto de parámetros que verifican la condición de equilibrio.

2. Segundo procedimiento.

   Consiste en aproximar el valor de la transformada de la distribución predictiva, $\Pi_n^*(s \mid \mathbf{t},\mathbf{s})$, e invertirla numéricamente. Para ello, se puede tener en cuenta que, considerando la definición de transformada

de Laplace, resulta que,

$$\begin{aligned}\Pi_n^*(s \mid \mathbf{t},\mathbf{s}) &= \int_0^\infty e^{-s\tau}\Pi_n(\tau \mid \mathbf{t},\mathbf{s})\,d\tau,\\ &= \int_0^\infty e^{-s\tau}\left[\int_\Theta \Pi_n(\tau \mid \theta)\,f(\theta \mid \mathbf{t},\mathbf{s})\,d\theta\right]d\tau,\\ &= \int_\Theta \left[\int_0^\infty e^{-s\tau}\Pi_n(\tau \mid \theta)\,d\tau\right] f(\theta \mid \mathbf{t},\mathbf{s})\,d\theta,\\ &= \int_\Theta \Pi_n^*(s \mid \theta)\,f(\theta \mid \mathbf{t},\mathbf{s})\,d\theta,\end{aligned}$$

donde Θ es el espacio de estados del conjunto de parámetros $\theta = (\theta_\lambda, \theta_\mu)$, introducidos en la Subsección 5.1.1. Entonces, se puede obtener, la siguiente aproximación Monte Carlo de la transformada condicionada en $\rho < 1$,

$$\Pi_n^*(s \mid \mathbf{t},\mathbf{s},\rho<1) \approx \frac{1}{R}\sum_{j:\rho^{(j)}<1} \Pi_n^*\left(s \mid \theta^{(j)}\right). \tag{5.35}$$

En este segundo procedimiento, se obtiene una estimación de $\Pi_n(\tau \mid \mathbf{t},\mathbf{s},\rho<1)$ invirtiendo numéricamente la aproximación de su transformada, (5.35), mediante el algoritmo de Hosono (1981).

Es importante puntualizar que existen diferencias substanciales entre ambos procedimientos tanto en la precisión de las estimaciones como en el coste computacional. El primer método, es más costoso computacionalmente puesto que requiere la inversión numérica de la transformada $\Pi_n^*(s \mid \theta^{(j)})$ en cada iteración MCMC, mientras que con el segundo procedimiento, se requiere utilizar una sola vez el algoritmo de inversión numérica, para invertir la aproximación de la transformada de la distribución predictiva, $\Pi_n^*(s \mid \mathbf{t},\mathbf{s},\rho<1)$, dada en (5.35). Sin embargo, como se argumenta en Conesa (2000) para una situación equivalente, el segundo procedimiento es muy sensible a la precisión en la evaluación de la transformada de la distribución predictiva, dada en (5.35), y más aún, si, como es el caso, no se conoce su expresión explícita, sino que es el resultado de una aproximación. En la práctica, se ha observado que existen diferencias notables en el resultado de las estimaciones con ambos procedimientos, que se incrementan a medida que aumenta la complejidad del modelo de colas MGE/MGE/1. Por esta razón, en este Capítulo, se sugiere el uso del primer procedimiento a pesar del incremento en el coste computacional.

La estimación de la distribución predictiva de $W(\tau)$, determinada por las probabilidades a posteriori,

$$\Gamma_w(\tau \mid \mathbf{t},\mathbf{s}) = \Pr(W(\tau) \le w \mid \mathbf{t},\mathbf{s}), \qquad \text{para } w \le 0, \tag{5.36}$$

se puede abordar de forma análoga a como se ha procedido con la distribución predictiva de $N(\tau)$. Para evitar repetir el desarrollo anterior y con objeto de sintetizarlo, se incluyen a continuación, de modo esquemático, los dos procedimientos de estimación introducidos anteriormente, aplicados ahora sobre las probabilidades dadas en (5.36).

PROCEDIMIENTO 1 DE ESTIMACIÓN DE $\Gamma_w(\tau \mid \mathbf{t},\mathbf{s},\rho<1)$.

1. Para cada $\theta^{(j)}$ que verifica $\rho^{(j)}<1$, calcular $\Gamma_w\left(\tau \mid \theta^{(j)}\right)$ invirtiendo $\Gamma_w^*\left(s \mid \theta^{(j)}\right)$
2. Estimar,

$$\Gamma_w(\tau \mid \mathbf{t},\mathbf{s},\rho<1) \approx \frac{1}{R}\sum_{j:\rho^{(j)}<1} \Gamma_w\left(\tau \mid \theta^{(j)}\right), \tag{5.37}$$

PROCEDIMIENTO 2 DE ESTIMACIÓN DE $\Gamma_w(\tau \mid \mathbf{t},\mathbf{s},\rho<1)$.

1. Para cada $\theta^{(j)}$ que verifica $\rho^{(j)} < 1$, evaluar $\Gamma_w^*\left(s \mid \theta^{(j)}\right)$ y aproximar,

$$\Gamma_w^*\left(s \mid \mathbf{t}, \mathbf{s}, \rho < 1\right) \approx \frac{1}{R} \sum_{j;\rho^{(j)}<1} \Gamma_n^*\left(s \mid \theta^{(j)}\right).$$

2. Invertir numéricamente la transformada $\Gamma_w^*\left(s \mid \mathbf{t}, \mathbf{s}, \rho < 1\right)$.

Por último, la distribución predictiva de la longitud del periodo de ocupación, B, se puede estimar utilizando los dos procedimientos equivalentes a los que se acaban de exponer. En particular, haciendo uso del primer método y análogamente a la estimación de distribuciones predictivas en Capítulos anteriores, la función de distribución de la longitud del periodo de ocupación se puede estimar mediante,

$$F_B\left(x \mid \mathbf{t}, \mathbf{s}, \rho < 1\right) \approx \frac{1}{R} \sum_{j;\rho^{(j)}<1} F_B\left(x \mid \theta^{(j)}\right), \tag{5.38}$$

donde $F_B(x \mid \theta^{(j)})$ es la función de distribución de B obtenida invirtiendo numéricamente $F_B^*(s \mid \theta^{(j)})$, dada en (5.20).

Obsérvese que, en este caso, sí es necesario asumir que se verifica la condición de ergodicidad, $\rho < 1$, para que exista la distribución predictiva de B, puesto que la longitud del periodo de ocupación, B, sólo está definida bajo esta condición.

5.3. Ilustraciones.

En esta Sección, se ilustran los procedimientos de estimación desarrollados en este Capítulo sobre dos sistemas de colas simulados a partir de dos conjuntos de observaciones generadas, para cada sistema, de los tiempos entre llegadas y de servicio en el mismo. En primer lugar, se generan observaciones correspondientes a un sistema M/M/1. El motivo por el que se considera el modelo de colas más sencillo es el de comparar las distribuciones estimadas con su valor verdadero cuando los parámetros del sistema son conocidos. En segundo lugar, se consideran observaciones generadas de un sistema Weib/MGE/1, con el objeto de ilustrar los procedimientos de estimación expuestos sobre sistemas de colas GI/G/1 que no pertenezcan a la familia de modelos MGE/MGE/1.

Para simular estos sistemas se utilizan algunos de los conjuntos de datos generados en las Sección 2.5 del Capítulo 2 y un nuevo conjunto de datos. Concretamente, para el sistema M/M/1, se consideran:

- 100 tiempos entre llegadas generados en la Sección 2.5 de una distribución exponencial de media 1.0.
- 100 tiempos de servicio generados de una distribución exponencial de media 0.6.

Y, para el sistema Weib/MGE/1, se consideran:

- 100 tiempos entre llegadas al sistema generados en la Sección 2.5 de una distribución Weibull, $Weib\,(1.5, 1.5)$.
- 100 tiempos de servicio generados en la Sección 2.5 de una distribución de una distribución MGE con $\mathbf{Q} = (0.1, 0.9)$ y $\boldsymbol{\mu} = (5, 30)$

Dadas estas observaciones, el objetivo es estimar las distribuciones transitorias del número de clientes presentes en el sistema y del tiempo de espera en cola, así como la distribución de la longitud del periodo de

ocupación. Para aplicar los métodos propuestos en este Capítulo, se utilizan las muestras MCMC obtenidas en la Sección 2.5 a partir del algoritmo RJMGE y una nueva muestra obtenida mediante este mismo algoritmo con el nuevo conjunto de datos. De este modo, se tiene, para cada sistema de colas, una muestra del tipo dado en (5.21), con pares de parámetros de la distribución del tiempo entre llegadas y de servicio. En la Subsección 5.2.2, se proponen dos procedimientos diferentes de estimación para las distribuciones predictivas mencionadas, $\Pi_n(\tau \mid \mathbf{t}, \mathbf{s}, \rho < 1)$, $\Gamma_w(\tau \mid \mathbf{t}, \mathbf{s}, \rho < 1)$ y $F_B(x \mid \mathbf{t}, \mathbf{s}, \rho < 1)$. En las ilustraciones que se muestran a continuación, se hace uso del primero de los procedimientos propuestos puesto que, tal como se detalla en el apartado 5.2.2.2, de este modo, se obtiene, generalmente, una estimación más precisa aunque el coste computacional sea un poco mayor.

Como se indicó en la Sección 5.1, para calcular las distribuciones transitorias de $N(\tau)$ y de $W(\tau)$, se supone que el sistema está vacío en el instante inicial, $\tau = 0$, y que la únicas probabilidades iniciales distintas de cero son, $\pi_{0,j}(0)$, para $i = 1, ..., L$, véase (5.8). En el caso de la distribución transitoria de $W(\tau)$, se requiere además que estas probabilidades iniciales, $\boldsymbol{\pi}_0(0)$, verifiquen la condición dada en (5.16). Aunque esta condición no es necesaria para calcular la distribución transitoria de $N(\tau)$, en esta Sección, se asume también para este caso con objeto de fijar unos valores para las probabilidades iniciales, $\boldsymbol{\pi}_0(0)$, dados en (5.19). Con este supuesto, las expresiones para los valores de D_r, dados en (5.10), se simplifican del modo siguiente, véase Bertsimas y Nakazato (1992),

$$D_r = \frac{1}{sE[A_1]} \frac{(-1)^M S^*_{1,M}(0)}{S^*_{1,M}(x_r(s))} x_r(s) \prod_{\substack{k=1 \\ k \neq r}}^{M} \frac{x_r(s)}{x_r(s) - x_k(s)}, \qquad \text{para } r = 1, ..., M,$$

donde A_1 es la variable que representa el tiempo entre las llegadas al sistema.

Con respecto a los parámetros utilizados para el algoritmo de Hosono (1981), en esta Sección, se ha fijado el valor de $a = 6$ para la aproximación dada en (5.39) en el apéndice. Puesto que las transformadas, $\Pi^*_n(s \mid \boldsymbol{\theta}^{(j)})$, $\Gamma^*_w(s \mid \boldsymbol{\theta}^{(j)})$ y $F^*_B(s \mid \boldsymbol{\theta}^{(j)})$, invertidas en cada iteración MCMC, verifican que sus inversas están acotadas por $M = 1$, ya que corresponden a las probabilidades de que haya n clientes en el sistema, de que el tiempo de espera sea menor que w y de que el periodo de ocupación sea menor que x, respectivamente, se obtiene un error de aproximación igual a 6.144×10^{-6}, véase (5.43) en el apéndice. Además, para fijar los parámetros de truncamiento, se ha utilizado el procedimiento propuesto por Bertsimas y Nakazato (1992), expuesto esquemáticamente al final del Apéndice 5.5, lo que da lugar a un error total de inversión menor que 10^{-5} veces el valor de la probabilidad de interés. En particular, en el paso 1, se ha fijado un valor inicial $k = 50$, que ha resultado ser suficiente en la mayoría de los casos.

En la Tabla 5.1, se muestra la probabilidad a posteriori de que se verifique la condición de equilibrio, $\rho < 1$, calculada a partir de (5.22) y la estimación de la intensidad de tráfico, ρ, obtenida mediante (5.24), en los dos sistemas de colas considerados. Nótese que el verdadero valor de la intensidad de tráfico conocidos los parámetros de los sistemas simulados M/M/1 y Weib/MGE/1 es 0.6 y 0.3338, respectivamente. En ambos casos, se observa que la probabilidad, $P(\rho < 1 \mid, \mathbf{t}, \mathbf{s})$, es lo suficientemente elevada como para asumir que existen las distribuciones estacionarias. Sin embargo, por diversas razones, que se han comentado a lo largo del Capítulo, puede resultar interesante el análisis del comportamiento transitorio del sistema evaluado desde el instante inicial, que se muestra a continuación.

	M/M/1	Weib/MGE/1
$P(\rho < 1 \mid, \mathbf{t}, \mathbf{s})$	0.9987	0.9999
$E[\rho \mid \mathbf{t}, \mathbf{s}]$	0.5837	0.3454

Tabla 5.1: Estimaciones de la probabilidad a posteriori de que exista equilibrio en el sistema y valores esperados de la intensidad de tráfico en cada uno de los dos sistemas de colas simulados.

En las Figuras 5.2 y 5.3, se muestra la evolución de las estimaciones de la distribución transitoria,

$$\Pi_n(\tau \mid \mathbf{t}, \mathbf{s}, \rho < 1) = \Pr(N(\tau) = n \mid \mathbf{t}, \mathbf{s}, \rho < 1),$$

del número de clientes presentes en los dos sistemas de colas, obtenidas mediante (5.37), evaluadas en distintos instantes de tiempo, τ. En el caso del sistema M/M/1, se muestra también, en línea continua, el valor de la distribución estacionaria del tamaño del sistema cuando se conocen los parámetros, que se distribuye geométricamente según (1.10). Obsérvese que, a medida que aumenta el valor de τ, la distribución converge a su valor estacionario. En el caso del sistema Weib/MGE/1, no se muestra la distribución estacionaria en línea continua puesto que no se conoce, en este caso, su valor explícito dados los parámetros, aunque se aprecia la convergencia de las distribuciones transitorias estimadas a un valor estacionario. En particular, obsérvese que, para valores muy grandes de τ, la estimación de la probabilidad de que el sistema esté vacío, $\Pi_0(\tau \mid \mathbf{t}, \mathbf{s}, \rho < 1)$, es igual a uno menos el valor medio a posteriori de ρ, que se indica en la Tabla 5.1, lo cual es congruente con el resultado conocido que afirma que, en cualquier sistema GI/G/1, la probabilidad estacionaria de que el sistema se encuentre vacío en un instante aleatorio es $(1-\rho)$, véase (1.7). Por otro lado, nótese también que la convergencia a la distribución estacionaria en el sistema M/M/1 es más lenta que en el sistema Weib/MGE/1, lo que era de esperar puesto que el valor de la intensidad de tráfico es mayor en el sistema M/M/1.

En las Figuras 5.4 y 5.5, se muestran las estimaciones de la función de distribución transitoria,

$$\Gamma_w(\tau \mid \mathbf{t}, \mathbf{s}, \rho < 1) = \Pr(W(\tau) \leq w \mid \mathbf{t}, \mathbf{s}, \rho < 1),$$

del tiempo de espera en cola en el sistema M/M/1 y el sistema Weib/MGE/1, respectivamente, calculadas a partir de (5.34). De nuevo, se consideran valores de τ variando desde el instante inicial, $\tau = 0$, hasta $\tau = 100$, instante en el que, en ambos sistemas, la distribución ha convergido a su valor estacionario. Como antes, en el sistema M/M/1, se ilustra, en línea continua, el valor de la función de distribución estacionaria cuando los parámetros son conocidos, que es proporcional a una distribución exponencial, véase (1.11). Y, por los motivos mencionados anteriormente, esta distribución no se muestra en el caso del sistema Weib/MGE/1. Obsérvese que, para valores grandes de τ, las estimaciones de la probabilidad estacionaria de que el tiempo de espera en cola sea nulo, $\Gamma_0(\tau \mid \mathbf{t}, \mathbf{s}, \rho < 1)$, son congruentes con los resultados teóricos que afirman que, si el proceso de llegadas es Markoviano, como es el caso del sistema M/M/1, esta probabilidad coincide con la probabilidad de que el sistema esté vacío en un instante aleatorio y es igual a $(1-\rho)$. Sin embargo, si el proceso de llegadas no es de Poisson, como es el caso del sistema Weib/MGE/1, estas probabilidades no coinciden. Nuevamente, se aprecia que la convergencia a la distribución estacionaria en el sistema M/M/1 es más lenta que en el sistema Weib/MGE/1.

Por último, en las Figuras 5.6 y 5.7, se muestra la estimación de la función de distribución predictiva de la longitud del periodo de ocupación, $F_B(x \mid \mathbf{t}, \mathbf{s}, \rho < 1)$, para cada sistema de colas, obtenidas mediante (5.38). En el caso del sistema M/M/1, se ilustra también, con línea continua, la función de distribución cuando los parámetros son conocidos cuyo valor se obtiene integrando la función de densidad dada en (1.13).

El tiempo que se requiere para obtener las estimaciones que se han mostrado en esta Sección varía según el caso, dependiendo, fundamentalmente, de la cantidad de valores que se deseen estimar y de la simplicidad del modelo de colas que se esté considerando. El primer motivo es evidente ya que, por ejemplo, el tiempo que se necesita para obtener las estimaciones de $\Pi_n(\tau \mid \mathbf{t}, \mathbf{s}, \rho < 1)$, presentadas en las Figuras 5.2 y 5.3, es menor que el que se necesita para obtener las de $\Gamma_w(\tau \mid \mathbf{t}, \mathbf{s}, \rho < 1)$, presentadas en las Figuras 5.4 y 5.5, ya que para en las primeras se estiman $6 \times 4 = 24$ valores, en cada caso, mientras que en las segundas $6 \times 13 = 78$. Sin embargo, la estimación individual de cada uno de estos valores es mayor en el primer caso que en el segundo ya que evaluar la transformada $\Pi_n^*(s \mid \boldsymbol{\theta}^{(j)})$ es más costoso que evaluar la transformada $\Gamma_w^*(s \mid \boldsymbol{\theta}^{(j)})$. El segundo factor del que depende el coste computacional es, como se ha mencionado, la simplicidad del modelo de colas. Teniendo en cuenta que el tiempo que se necesita para invertir la transformada de una determinada

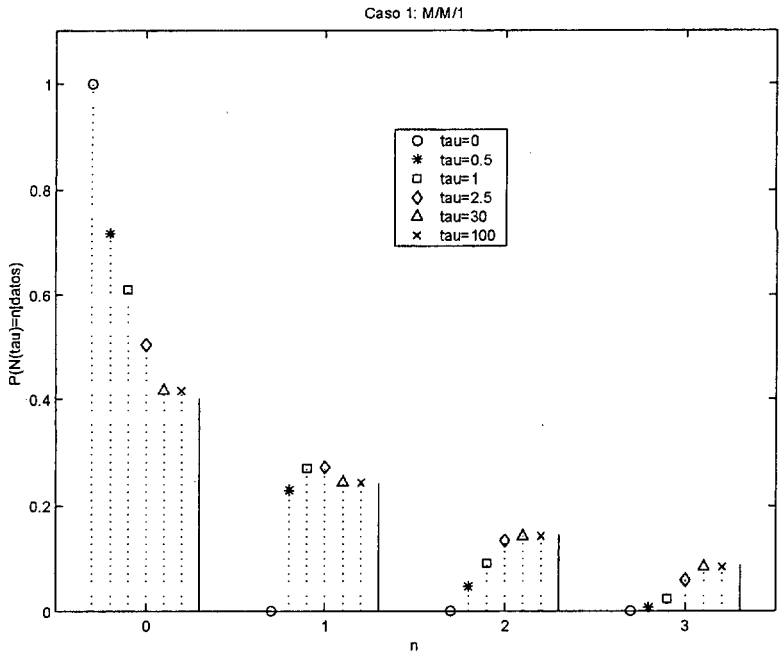

Figura 5.2: Estimación de la distribución transitoria del número de clientes en el sistema M/M/1 desde el instante inicial, $\tau = 0$, hasta la convergencia a la distribución estacionaria. Se muestra, en línea continua, el valor verdadero de la distribución estacionaria, conocidos los parámetros.

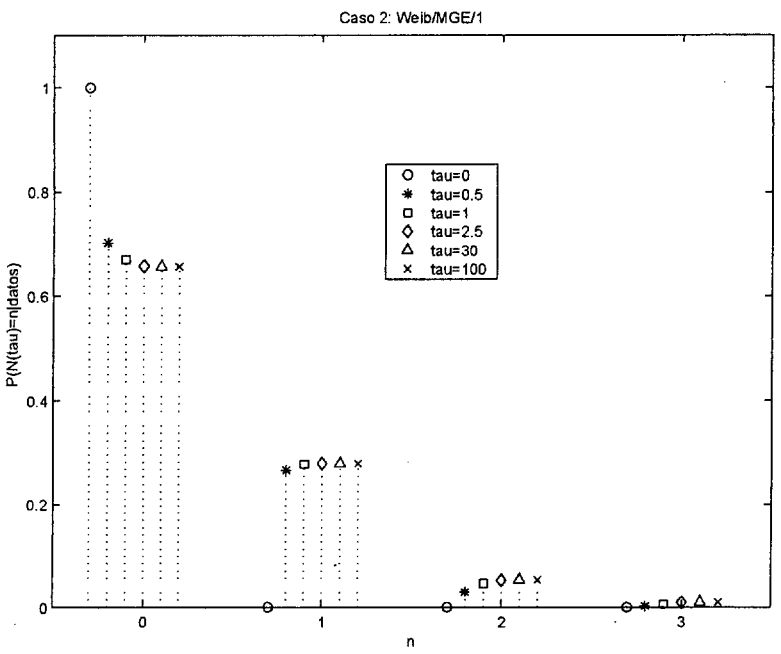

Figura 5.3: Estimación de la distribución transitoria del número de clientes en el sistema Weib/MGE/1 desde el instante inicial, $\tau = 0$, hasta la convergencia a la distribución estacionaria. No se muestra el valor verdadero de la distribución estacionaria conocidos los parámetros, puesto que no se conoce su valor explícito.

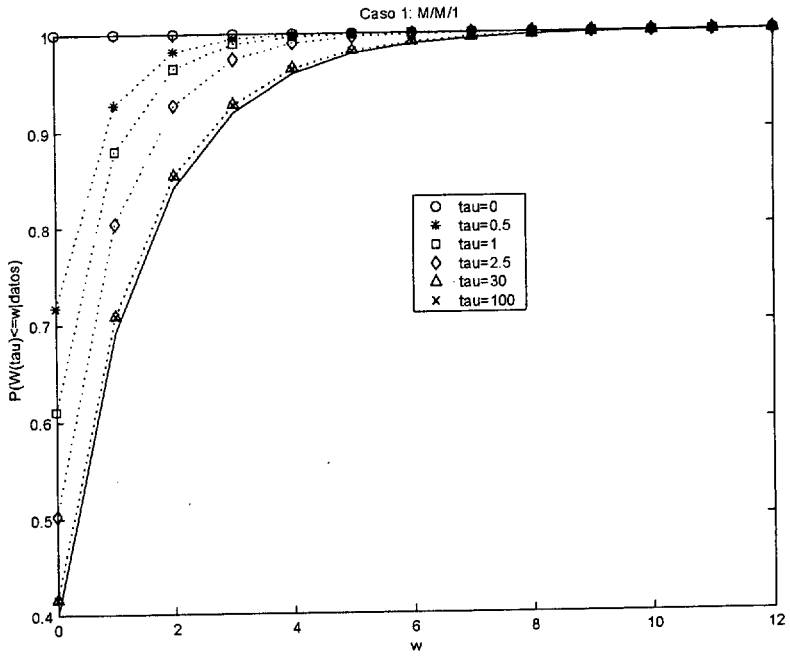

Figura 5.4: Estimación de la distribución transitoria del tiempo de espera en colas en el sistema M/M/1 desde el instante inicial, $\tau = 0$, hasta la convergencia a la distribución estacionaria. Se muestra, en línea continua, el valor verdadero de la distribución estacionaria, conocidos los parámetros.

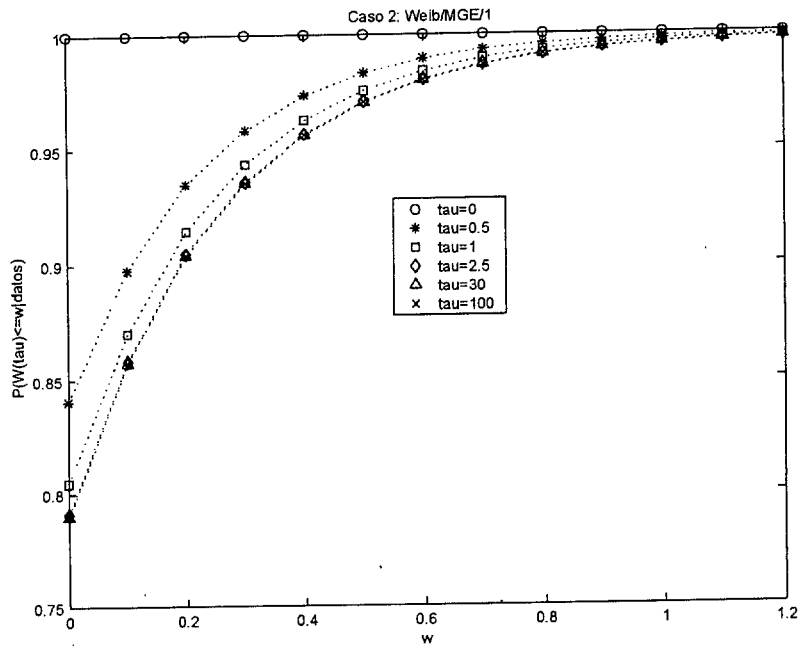

Figura 5.5: Estimación de la distribución transitoria del tiempo de espera en cola en el sistema Weib/MGE/1 desde el instante inicial, $\tau = 0$, hasta la convergencia a la distribución estacionaria. No se muestra el valor verdadero de la distribución estacionaria conocidos los parámetros, puesto que no se conoce su valor explícito.

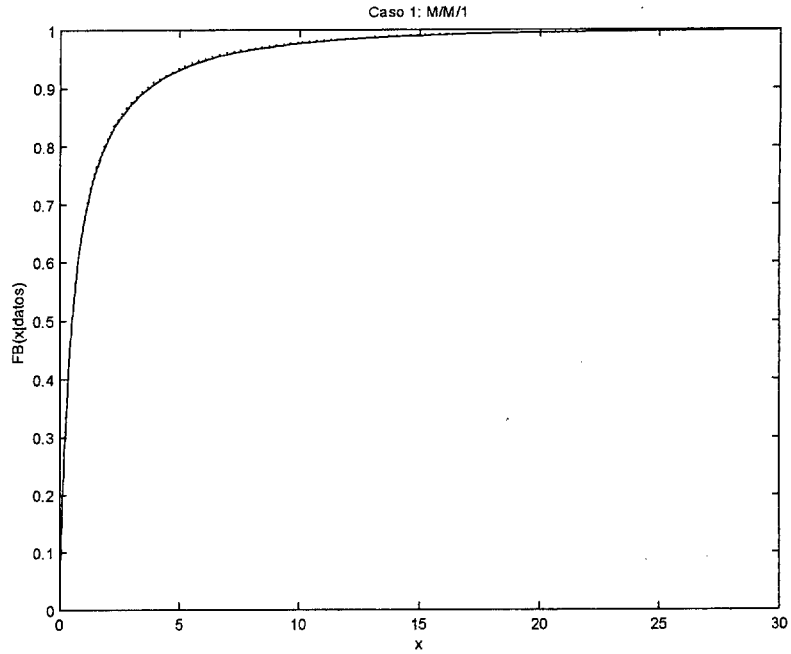

Figura 5.6: Estimación de la distribución estacionaria de la longitud del periodo de ocupación en el sistema M/M/1. Se muestra, en línea continua, el valor verdadero de esta distribución, conocidos los parámetros.

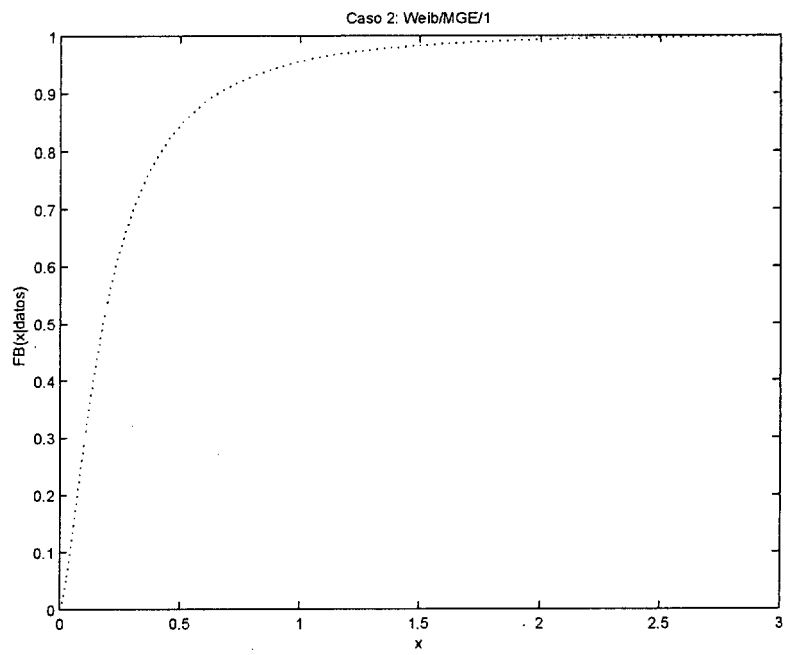

Figura 5.7: Estimación de la distribución estacionaria de la longitud del periodo de ocupación en el sistema Weib/MGE/1. No se muestra el valor verdadero de la distribución estacionaria conocidos los parámetros, puesto que no se conoce su valor explícito.

probabilidad aumenta con el número de parámetros del modelo de colas, la estimación de las distribuciones asociadas al sistema M/M/1 serán menos costosas ya que el número de iteraciones MCMC con dimensiones reducidas de los parámetros, $L^{(j)}$ y $M^{(j)}$, véase (5.21), es mucho mayor que en el sistema Weib/MGE/1. Concretamente, ninguna de las estimaciones mostradas en los gráficos de esta Sección requiere más de 20 minutos para su realización en un Pentium IV.

5.4. Comentarios y extensiones

En este Capítulo, se han propuesto diferentes métodos Bayesianos para la estimación de la distribución transitoria del número de clientes presentes en un sistema GI/G/1, de la distribución transitoria del tiempo de espera en cola y de la distribución de la longitud del periodo de ocupación. Para ello, se ha aproximado este sistema de colas general con la familia de modelos MGE/MGE/1 y se han utilizado los resultados de Bertsimas y Nakazato (1992) que permiten obtener expresiones para las transformadas de Laplace de las cantidades mencionadas cuando se conocen los parámetros del sistema MGE/MGE/1. Se ha ilustrado la metodología propuesta con dos sistemas de colas simulados.

Como se comentó en la Sección 5.3, el coste computacional asociado a la estimación de las distribuciones transitorias del tamaño del sistema, $N(\tau)$, y del tiempo de espera, $W(\tau)$, se incrementa considerablemente a medida que aumenta el número de instantes, τ, y el rango de valores de la variable para los que se realiza la estimación. Una alternativa menos costosa puede consistir en evaluar en cada instante, τ, la esperanza de $N(\tau)$ y de $W(\tau)$, o un número determinado de momentos. Para ello, se puede hacer uso de varios resultados que se muestran en Bertsimas (1990) relacionados con la función generadora de momentos de $N(\tau)$ y la transformada de Laplace-Stieljes de $W(\tau)$. Sin embargo, cabe esperar que para valores elevados de τ los momentos de las distribuciones de $N(\tau)$ y de $W(\tau)$ tiendan a infinito puesto que, en el estado estacionario, no existen los momentos de las distribuciones predictivas de estas cantidades ya que los sistemas de colas MGE/MGE/1 contienen a los modelos MGE/M/1 y M/MGE/1 en los que, como ya se ha comentado en esta tesis, se sabe que aparece el problema de la no existencia de momentos. Otra alternativa puede basarse en la estimación de la mediana, o de un número determinado de cuantiles, de las distribuciones predictivas en cada instante, τ. Sin embargo, conocidos los parámetros de un sistema MGE/MGE/1, la obtención de la mediana de $N(\tau)$ o de $W(\tau)$ no es trivial y, aunque se puede aproximar numéricamente, este procedimiento no es el más recomendable en un método de estimación MCMC puesto que no se espera un descenso tan favorable en el coste computacional y como es natural, las estimaciones serán peores. Actualmente, se está trabajando en todas estas cuestiones que abordan problemas teóricos y numéricos.

Por último, los procedimientos expuestos en este Capítulo podrían extenderse a sistemas de colas más generales GI/G/c mediante aproximaciones basadas en modelos de colas MGE/MGE/c y haciendo uso de los resultados obtenidos en Bertsimas (1990) para calcular algunas medidas de este sistema con más servidores cuando se conocen sus parámetros. Aunque las técnicas utilizadas en Bertsimas (1990) son similares a las empleadas en Bertsimas y Nakazato (1992), no se consideran las mismas cantidades de interés en el modelo de colas. En particular, en Bertsimas (1990), no se analiza el comportamiento transitorio del sistema y los procedimientos propuestos requieren un coste mucho mayor de tiempo y memoria, que se incrementa con el número de servidores.

5.5. Apéndice: Inversión numérica de transformadas de Laplace mediante el algoritmo de Hosono.

En este Apéndice, se introduce brevemente el algoritmo propuesto por Hosono (1981) para la inversión numérica de transformadas de Laplace. Este algoritmo se ha utilizado en este Capítulo para invertir numéri-

camente las transformadas de la distribución del número de clientes en el sistema, del tiempo de espera en cola y del periodo de ocupación.

Existen muchos otros procedimientos para la inversión numérica de transformadas de Laplace, como los basados en el método de las series de Fourier y otras técnicas que se recopilan en Abate y Whitt (1992) y Conesa (2000). Sin embargo, se ha optado por este algoritmo, fundamentalmente, porque Bertsimas y Nakazato (1992) lo han aplicado con éxito para el caso en el que los parámetros del sistema sean conocidos y además, es muy rápido y fácil de implementar. Al igual que se hace en Bertsimas y Nakazato (1992), se han obtenido aproximaciones muy precisas al comparar, para el sistema M/M/1, las distribuciones del tamaño del sistema y del periodo de ocupación conocidas con las aproximaciones obtenidas invirtiendo numéricamente sus transformadas con el algoritmo de Hosono (1981). No ha sido posible realizar estas comparaciones con otros sistemas que no sean el modelo M/M/1 ya que no se han obtenido hasta ahora expresiones explícitas para sus distribuciones dados los parámetros del sistema. Nótese que, como se comentó anteriormente, las soluciones conocidas para el modelo de colas M/M/1 están expresadas en términos de funciones de Bessel modificadas, véase Gross y Harris (1985).

El algoritmo de Hosono (1981) es esencialmente equivalente al algoritmo de EULER, que es una variante del método de las series de Fourier, véase Abate y Whitt (1992), y que fue desarrollado por Dubner y Abate (1968). La estrategia fundamental del algoritmo de Hosono (1981) se basa en la siguiente aproximación,

$$\exp(s) \approx \frac{\exp(a)}{2\cosh(a-s)} = \frac{e^a}{2}\sum_{n=-\infty}^{\infty}\frac{(-1)^n i}{s-a-i(n-0.5)\pi}. \tag{5.39}$$

Teniendo en cuenta que una función, $f(\tau)$, se puede expresar en términos de su transformada de Laplace, $f^*(s)$, mediante,

$$f(\tau) = \frac{1}{2\pi i}\int_{\nu-i\infty}^{\nu+i\infty} f^*(s)\exp(s\tau)\,d\tau, \tag{5.40}$$

Hosono (1981) muestra que, bajo unas condiciones muy suaves para $f^*(s)$, se puede obtener una buena aproximación de $f(\tau)$ si se reemplaza el valor de $\exp(s\tau)$ en (5.40) por su aproximación obtenida a partir de (5.39) para valores elevados de a. Concretamente, el resultado de esta aproximación viene dado por,

$$f(\tau) \approx f_{ec}(\tau, a) = \frac{e^a}{\tau}\sum_{n=1}^{\infty} F_n, \tag{5.41}$$

donde,

$$F_n = (-1)^n \operatorname{Im}\left[f^*\left(\frac{a+i\pi(n-0.5)}{\tau}\right)\right]. \tag{5.42}$$

El valor de a se utiliza en Hosono (1981) como una medida de precisión en la aproximación obtenida. En particular, si $f(\tau)$ está acotado, es decir, si $|f(\tau)| \leq M$, para todo τ, el error de aproximación verifica que,

$$|f_{ec}(\tau, a) - f(\tau)| \leq M\frac{e^{-2a}}{|1-e^{-2a}|}. \tag{5.43}$$

Nótese que, como se comentó en la Sección 5.3, todas las transformadas que se han invertido en este Capítulo verifican que sus inversas están acotadas por $M = 1$ puesto que se trata de las probabilidades, $\Pr(N(\tau) = n)$, $\Pr(W(\tau) \leq w)$ y $\Pr(B \leq \tau)$, y por tanto se puede obtener fácilmente una medida del error de aproximación.

Obsérvese que, en la práctica, no es posible evaluar la serie infinita $f_{ec}(\tau, a)$, dada en (5.41), y es necesario truncarla de modo que se tenga un número finito de sumandos. Como es natural, Hosono (1981) recomienda no truncarla arbitrariamente y propone la siguiente aproximación, que denomina de orden (l, m),

$$f_{ec}^{(l,m)}(\tau, a) = \frac{e^a}{\tau}\left(\sum_{n=1}^{l-1} F_n + 2^{-m-1}\sum_{n=0}^{m-1} A_{mn}F_{l+n}\right),$$

donde A_{mn} se obtiene de forma recursiva según,

$$A_{mm} = 1, \qquad A_{mn-1} = A_{mn} + \binom{m+1}{n},$$

y donde l y m son parámetros de truncamiento, los cuales pueden determinarse según unas cotas establecidas en Hosono (1981) para controlar el error originado al truncar $f_{ec}(\tau, a)$.

Hosono (1981) afirma que, en términos generales, las condiciones para que el algoritmo funcione adecuadamente se verifican si la función $f(\tau)$ es lo suficientemente suave. De hecho, los puntos de discontinuidad de $f(\tau)$ originan errores en la inversión.

Bertsimas y Nakazato (1992) proponen el siguiente procedimiento para implementar el algoritmo de Hosono (1981) que permite obtener un error total en la inversión numérica menor que $10^{-a+1}|f(\tau)|$.

ALGORITMO HOSONO.

Para cada valor de τ,

1. Fijar $s_n = \tau^{-1}(a + i\pi(n - 0.5))$, definir $F_n = (-1)^n \operatorname{Im}[f^*(s_n)]$ y encontrar un k tal que,

$$\left| \sum_{r=0}^{a} \binom{a}{r} \frac{e^a}{\tau} F_{k+r} \right| < \left(\frac{2}{e^2} \right)^a.$$

2. Calcular los valores de,

$$C_n = 0.5^a \sum_{r=0}^{a-n-1} \binom{a}{r}, \quad \text{para } n = 0, ..., a-1.$$

3. Aproximar,

$$f(\tau) \approx \frac{e^a}{\tau} \left(\sum_{n=1}^{k-1} F_n + \sum_{r=0}^{a-1} C_r F_{k+r} \right).$$

Capítulo 6

Conclusiones y extensiones.

En esta tesis, se han propuesto distintos procedimientos para el análisis Bayesiano de modelos de colas con diferentes características y con distribuciones generales para el tiempo entre las llegadas y/o el tiempo de servicio en el sistema. Con el fin de aproximar estas distribuciones desconocidas, se han desarrollado varios métodos de estimación Bayesiana de densidades basados en modelos de mixturas de tipo PH con un número desconocido de componentes, que se ha abordado mediante métodos MCMC de dimensión paramétrica variable. Haciendo uso de las propiedades de los sistemas en los que intervienen distribuciones PH y de algunos resultados clásicos de la Teoría de Colas, se han obtenido estimaciones de las distribuciones estacionarias y transitorias de varias medidas de interés en los modelos de colas, como el número de clientes en el sistema o el tiempo de espera en cola. Además, se han propuesto métodos Bayesianos para el diseño de sistemas de colas con varios servidores, que se ha ilustrado con conjuntos de datos reales.

En cada Capítulo de esta tesis se ha incluido una Sección con los comentarios y extensiones relativos al tema específico considerado en cada uno de ellos. En este Capítulo, se pretende exponer, brevemente, algunas de las extensiones generales en la misma línea de trabajo de esta memoria, pero que requieren procedimientos Bayesianos y modelos de colas más sofisticados que los abordados en esta tesis. Estas líneas de investigación se han concretado en dos proyectos específicos que son el análisis tráfico de datos en internet y el estudio de los centros de llamadas (*call centers*).

Como es sabido, recientemente, el número de usuarios y el movimiento de información en internet ha aumentado drásticamente. Consecuentemente, la predicción del tráfico en internet ha adquirido una enorme importancia ya que su conocimiento constituye una herramienta útil para que los proveedores puedan evitar problemas como los originados por el bloqueo de los servidores, tomar decisiones como algunas cuestiones publicitarias, etc. Sin embargo, como sucede con otras redes de telecomunicaciones, como los centros de llamadas mencionados, se han desarrollado pocos estudios estadísticos relacionados con el tráfico en internet.

Las características principales de las redes de tráfico locales (LAN) y de grande escala (WAN) son la auto similitud y la llegada de mucha información simultánea (*burstiness*), véase, por ejemplo, Leland et al. (1994) y Willinger et al. (1995). Además, se ha mostrado que las llegadas de paquetes de internet (IP) se producen, en general, con características similares que no se ajustan a un proceso de Poisson, véase Paxson y Floyd (1995) y Crovella y Bestavros (1996). Con el fin de describir procesos de llegadas con este tipo de características, se han introducido diferentes modelos tales como el movimiento browniano fraccional, véase Vehel y Riedi (1997) y el proceso BMAP, que es un proceso MAP con llegadas en grupos, considerado en Klemm et al. (2003). Se puede encontrar una revisión de la literatura relativa a esta materia en Jagerman et al. (1996).

Sin embargo, una limitación importante del uso de estos procesos es que la inferencia y la estimación

de los parámetros es, generalmente, muy difícil debida a la complejidad de la función de verosimilitud. En particular, no existe apenas literatura destinada a la inferencia Bayesiana en el proceso BMAP o en el movimiento browniano fraccional.

El interés en el estudio del tráfico de internet no se limita al análisis de los procesos de llegadas de los paquetes IP, sino que se centra fundamentalmente en los sistemas de colas originados en su transmisión. Se han hecho varios estudios probabilísticos de los modelos de colas BMAP/G/1, véase, por ejemplo, Lucantoni (1993). Sin embargo, lamentablemente se han obtenido todavía pocos resultados sobre las características de este sistema cuando sus parámetros son conocidos. Por otro lado, tampoco se han desarrollado muchos estudios para modelos de colas cuyo proceso de llegadas es un movimiento browniano fraccional, un ejemplo se encuentra en Narayan (1998).

Por consiguiente, una línea de investigación futura por donde proseguir el trabajo de la tesis consiste en desarrollar técnicas de inferencia Bayesiana para los procesos de llegada de tráfico de IP y aplicar estos procedimientos a varias muestras reales de datos de internet. Para llevar a cabo la inferencia, se requiere el desarrollo de algoritmos numéricos aproximados que funcionen rápidamente en la práctica, ya que el cálculo de distribuciones a posteriori mediante los métodos MCMC puede ser muy costoso debido al elevado tamaño muestral que se encuentra generalmente en los datos de internet. Además, el análisis del proceso de formación de colas cuando los paquetes se encuentran en estado de espera requiere la combinación de resultados propios de la Teoría de Colas con métodos de simulación y con los métodos Bayesianos que se hayan desarrollado previamente.

La otra línea de investigación que se ha destacado en estas extensiones generales es el análisis del funcionamiento de los centros de llamadas. Un centro de llamadas es un conjunto de recursos, constituido, generalmente, por equipamiento informático, de personal y sistemas de telecomunicación, que posibilita la oferta de servicios mediante la comunicación telefónica, integrada con la informática, entre agentes (también llamados teleoperadores) y clientes. El mercado de las empresas de marketing telefónico es un sector en pleno desarrollo. En los últimos años, se ha registrado una tasa media anual de crecimiento aproximada del 20% en Estados Unidos, véase Koole y Mandelbaum (2002), y del 15% aproximadamente, en España, donde a finales de 2002, el sector estaba integrado, por alrededor de 60 operadores, el empleo generado se situó en unos 44.000 trabajadores, con una facturación media por empleado cercana a los 20.000 euros, según un estudio realizado por la consultora DBK.

Aunque el estudio de los centros de llamadas comprende muchas áreas de investigación, tales como sectores tecnológicos de la ICT (*Information and Communication Technology*) o diferentes ciencias sociales, la Teoría de Colas y la Investigación Operativa han constituido los pilares fundamentales para su desarrollo. Nótese que la interpretación de un centro de llamadas como un sistema de colas resulta muy natural. En esta dirección, existe una literatura muy amplia enfocada en la gestión y la optimización en los centros de llamadas, véase, por ejemplo, Harris et al. (1987) y Sze (1984) y recientemente, se han producido varias contribuciones considerando redes de colas para describir centros de llamadas más sofisticados que surgen en la actualidad, véase, por ejemplo, en Bhulai y Koole (2000) y en Kogan et al. (1997).

Sin embargo, como sucede para modelos de colas en general, se han realizado muy pocos estudios estadísticos del funcionamiento de los centros de llamadas. En Gans et al. (2003), se recopilan aproximadamente 200 publicaciones de carácter científico relacionadas con estas entidades, de las cuales únicamente 16 se destinan a su estudio estadístico. Además, en la mayoría de los casos, el modelo utilizado es el sistema M/M/c, que, en algunas ocasiones, recibe el nombre de Erlang C. Una excepción es Brown et al. (2002), donde se considera un conjunto de datos reales procedentes de un centro de llamadas en una entidad bancaria que se describe mediante un modelo de colas en el que las llegadas se producen según un proceso de Poisson no homogéneo y el tiempo de servicio sigue una distribución lognormal. El problema fundamental en este caso es que no existen muchos resultados clásicos sobre este modelo de colas que permitan estimar las distribuciones de las características de interés.

El análisis estadístico de los centros de llamadas es complicado ya que para obtener una descripción adecuada de su funcionamiento es necesario, como en el tráfico de IP, considerar modelos de colas en los que los tiempos entre llegadas (y a menudo los tiempos de servicio) no sean independientes entre sí. Además, es necesario incorporar en el modelo la posibilidad de abandono de clientes impacientes, así como el reintento de algunos de estos clientes. Los sistemas deben de tener capacidad finita puesto que, en la mayoría de los casos, el número de teleoperadores, aunque sean automáticos, es finito. Por último, es muy frecuente en los centros de llamadas encontrar diferentes prioridades en la atención de los clientes, por ejemplo, en centros de llamadas que se han implantado, recientemente, para gestionar la cita previa de pacientes en atención sanitaria primaria.

Consecuentemente, otra de las líneas de investigación futura o extensiones generales de la tesis es el desarrollo de métodos Bayesianos para modelos de colas con las características mencionadas propias de los centros de llamadas. Como es natural, algunos de los procedimientos que se desarrollen para la inferencia en procesos de llegadas de paquetes en internet se pueden aplicar también a los centros de llegadas ya que comparten muchas características como la dependencia del tiempo y las llegadas en grupos.

Bibliografía

Abate, J. y Whitt, W. (1987). Transient behavior of the M/M/1 queue: Starting at the origin. *Queueing Systems: Theory and Applications*, 2:41–65.

Abate, J. y Whitt, W. (1988). Transient behavior of the M/M/1 queue via laplace transforms. *Advances in Applied Probability*, 20:145–178.

Abate, J. y Whitt, W. (1992). The Fourier-series method for inverting transforms of probability distributions. *Queueing Systems*, 10:5–88.

Abramowitz, M. y Stegun, I. A. (1964). *Handbook of Mathematical Functions.* Dover, Nueva York.

Allen, A. O. (1990). *Probability, Statistics and Queueing Theory with Computer Science Applications.* Academic Press, Boston.

Armero, C. (1985). Bayesian analysis of M/M/1/FIFO/∞ queues. En Bernardo, J. M., DeGroot, M. H., Lindley, D. V., y Smith, A. F. M., eds., *Bayesian Statistics 2*, págs. 3–23. North-Holland, Amsterdam.

Armero, C. (1994). Bayesian inference in Markovian queues. *Queueing Systems*, 15:419–426.

Armero, C. y Bayarri, M. J. (1994a). Bayesian prediction in M/M/1 queues. *Queueing Systems*, 15:401–417.

Armero, C. y Bayarri, M. J. (1994b). Prior assesments for prediction in queues. *The Statistician*, 43:139–153.

Armero, C. y Bayarri, M. J. (1996). Bayesian questions and answers in queues. En Bernardo, J. M., Berger, J. O., Dawid, A. P., y Smith, A. F. M., eds., *Bayesian Statistics 5*, págs. 613–618. Oxford University Press, Oxford.

Armero, C. y Bayarri, M. J. (1997). A Bayesian analysis of a queueing system with unlimited service. *Journal of Statistical Planning and Inference*, 58:241–261.

Armero, C. y Bayarri, M. J. (1999). Dealing with uncertainties in queues and network of queues: a Bayesian approach. En Ghosh, S., eds., *Multivariate Analysis, Design of Experiments, and Survey Sampling*, págs. 579–608. Marcel Dekker, Nueva York.

Armero, C. y Conesa, D. (1998). Inference and prediction in bulk arrival queues and queues with service in stages. *Applied Stochastic Models and Data Analysis*, 14:35–46.

Armero, C. y Conesa, D. (2000). Prediction in Markovian bulk arrival queues. *Queueing Systems*, 34:327–350.

Armero, C. y Conesa, D. (2003). Statistical performance of a multiclass bulk production queueing system. *European Journal of Operational Research.* Aceptado.

Asmussen, S. (1987). *Applied Probability and Queues.* John Wiley and Sons, Nueva York.

Asmussen, S. (1992). Phase-type representations in random walk and queueing problems. *The Annals of Probability*, 20(2):772–789.

Asmussen, S. (1997). Phase-type distributions and related point processes: Fitting and recent advances. En Alfa, A. S. y Chakravarthy, S., eds., *Matrix-Analytic Methods in Stochastic Models*, págs. 137–149. Marcel Dekker.

Asmussen, S. (2000). *Ruin Probabilities*. World Scientific Publishing, Singapur.

Asmussen, S. y Moller, J. R. (2001). Calculation of the steady-state waiting time distribution in GI/PH/c and MAP/PH/c queues. *Queueing Systems*, 37:9–29.

Asmussen, S., Nerman, O., y Olsson, M. (1996). Fitting phase-type distributions via the EM algorithm. *Scandinavian Journal of Statistics*, 23:419–441.

Ausín, M. C., Lillo, R. E., Wiper, M. P., y Ruggeri, F. (2003). Bayesian modeling of hospital bed occupancy times using a mixed generalized erlang distribution. En Bernardo, J. M., Berger, J. O., Dawid, A. P., y West, M., eds., *Bayesian Statistics 7*, págs. 443–452. Oxford University Press, Oxford.

Ausín, M. C., Wiper, M. P., y Lillo, R. E. (2004). Bayesian estimation for the M/G/1 queue using a phase type approximation. *Journal of Statistical Planning and Inference*, 118:83–101.

Bagchi, T. P. y Cunningham, A. A. (1972). Bayesian approach to the design of queueing systems. *Infor: Information Systems and Operational Research*, 10:36–46.

Banks, J., Carson, J. S., y Nelson, B. L. (1996). *Discrete-Event System Simulation*. Prentice-Hall, Englewood Cliffs.

Bell, C. (1971). Characterization and computation of optimal operating policies for operating an M/G/1 queueing system with removable server. *Operations Research*, 19:208–218.

Bertsekas, D. y Gallager, R. (1992). *Data networks*. Prentice-Hall International, Londres.

Bertsimas, D. (1990). An analytic approach to a general class of G/G/c queueing systems. *Operations Research*, 38:139–155.

Bertsimas, D. y Nakazato, D. (1992). Transient and busy period analysis of the G/G/1 queue: The method of stages. *Queueing Systems*, 10:153–184.

Bhat, U., Miller, G. K., y Rao, S. S. (1997). Statistical analysis of queueing systems. En Dshalalow, J. H., eds., *Frontiers in Queuing*, págs. 351–394. CRC Press, Boca Raton.

Bhattacharya, S. K. y Singh, N. (1994). Bayesian estimation of the traffic intensity in $M/E_k/1$ queue. *Far-East Journal of Mathematical Sciences*, 2:57–62.

Bhulai, S. y Koole, G. M. (2000). A queueing model for call blending in call centers. En *Proceedings of the 39th IEEE CDC, IEEE Control Society*, págs. 1421–1426.

Bladt, M., Gonzalez, A., y Lauritzen, S. L. (2003). The estimation of phase-type related functionals using Markov Chain Monte Carlo methods. *Scandinavian Actuarial Journal. To appear.*

Blomqvist, N. (1967). The covariance function of the M/G/1 queueing system. *Skandinavisk Aktuarietidskrift*, 50:157–174.

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., y Zhao, L. (2002). Statistical analysis of a telephone call center: A queueing-science perspective. Technical Report 03-12, Wharton School Center for Financial Institutions, University of Pennsylvania.

Butler, R. y Huzurbazar, A. (2000). Bayesian prediction of waiting times in stochastic models. *Canadian Journal of Statistics*, 28:311–325.

Cappé, O., Robert, C. P., y Rydén, T. (2003). Reversible jump MCMC converging to birth-and-death MCMC and more general continuous time samplers. *Journal of the Royal Statistical Society, Series B. To appear.*

Casella, G. y Robert, C. P. (1996). Rao-blackwellization of sampling schemes. *Biometrika*, 83:81–94.

Clarke, A. B. (1957). Maximum likelihood estimates in a simple queue. *Annals of Mathematical Statistics*, 28:1036–1040.

Conesa, D. V. (2000). *Inferencia y predicción en colas con ingresos o servicios en grupos.* Tesis Doctoral, Universitat de València.

Coolen, F. y Coolen-Schrijner, P. (2003). A nonparametric predictive method for queues. *European Journal of Operational Research*, 145:405–422.

Cox, D. R. (1955). A use of complex probabilities in the theory of stochastic processes. *Proceedings of the Cambridge Philosophical Society*, 51:313–319.

Cox, D. R. (1966). Some problems of statistical analysis connected with congestion. En Smith, W. L. y Wilkinson, W. E., eds., *Proceedings of the Symposium on Congestion Theory*, págs. 289–316. University of North Carolina Press, Chapel Hill.

Crovella, M. y Bestavros, A. (1996). Self-similarity in World Wide Web Traffic: Evidence and possible causes. En *Proceedings of SIGMETRICS'96: The ACM Internacional Conference on Measurement and Modeling of Computer Systems*, págs. 160–169.

Cumani, A. (1982). On the canonical representation of homogenous Markov processes modelling failure-time distributions. *Microelectronics and reliability*, 22(3):583–602.

Dempster, A. P., Laird, N. M., y Rubin, D. B. (1997). Maximum likelihood from incomplete data via the em algorith. *Journal of the Royal Statistical Society, Series B*, 99:1–38.

Diebolt, J. y Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, Series B*, 56:363–375.

Dubner, H. y Abate, J. (1968). Numerical inversion of Laplace transforms by relating them to the finite Fourier cosine transform. *Journal of the Association for Computing Machinery*, 15:115–123.

Erlang, A. K. (1909). The theory of probabilities and telephone conversations. *Nyt Tidsskrift for Matematik B*, 20.

Faddy, M. J. y McClean, S. I. (1999). Analysing data on lengths of stay of hospital patients using phase-type distributions. *Applied Stochastic Models in Business and Industry*, 15:311–317.

Feldmann, A. y Whitt, W. (1998). Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. *Performance Evaluation*, 31(8):963–976.

Fishman, G. S. (2001). *Discrete-Event Simulation: Modeling, Programming, and Analysis.* Springer-Verlag, Berlin.

Gander, W. y Gautschi, W. (2000). Adaptive quadrature - revisited. *BIT*, 40(1):84–101.

Ganesh, A., Green, P., O'Connell, N., y Pitts, S. (1998). Bayesian network management. *Queueing Systems*, 28:267–282.

Gans, N., Koole, G., y Mandelbaum, A. (2003). Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5:79–141.

Gelfand, A. E. y Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409.

Gilks, W., Richardson, S., y Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in practice*. Chapman and Hall, Londres.

Gorenescu, F., McClean, S. I., y Millard, P. H. (1999). Using a M/PH/C queue to optimise hospital bed occupancy. *Proceedings of the Applied Stochastic Models and Data Analysis Conference, Lisbon*, págs. 106–111.

Grassman, W. K. (1977). Transient solutions in markovian queueing systems. *Computers & Operations Research*, 4:47–53.

Green, P. J. (1995). Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732.

Gross, D. y Harris, C. M. (1985). *Fundamentals of queueing theory*. John Wiley and Sons, Nueva York.

Gross, D., Masi, D., Shortle, J., y Fischer, M. (2002). Difficulties in simulating queues with pareto service. En Yücesan, E., Chen, C. H., Snowdon, J. L., y Charnes, J. M., eds., *Proceedings of the 2002 Winter Simulation Conference*, págs. 407–415.

Gruet, M.-A., Philippe, A., y Robert, C. P. (1999). MCMC control spreadsheets for exponential mixture estimation. *Journal of Computational and Graphical Statistics*, 8:298–317.

Hajek, B. (1983). The proof of a folk theorem on queuing delay with applications to routing in networks. *Journal of the Association for Computing Machinery*, 30:834–851.

Halfin, S. (1982). Linear estimators for a class of stationary queueing processes. *Operations Research*, 30:515–529.

Harris, C. M. (1968). The pareto distribution as a queue service discipline. *Operations Research*, 16:307–313.

Harris, C. M., Hoffman, K. L., y Saunders, P. (1987). Modeling the irs telephone taxpayer information system. *Operations Research*, 35:504–523.

Harrison, G. W. y Millard, P. H. (1991). Balancing acute and long term care: the mathematics of throughput in departments of geriatric medicine. *Methods of information in medicine*, 30:221–228.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109.

Hosono, T. (1981). Numerical inversion of Laplace transform and some applications to wave optics. *Radio Science*, 16:1015–1019.

Huzurbazar, A. V. (1999). Flowgraph methods for generalized phase type distributions with non-exponential waiting times. *Scandinavian Journal of Statistics*, 26:145–157.

Jagerman, D., Melamed, B., y Willinger, W. (1996). Stochastic modelling of traffic processes. En Dshalalow, J. H., eds., *Frontiers in Queuing: Models, Methods and Problems*. CRC Press.

Jagers, A. A. y Van Doorn, E. A. (1986). On the continued erlang loss function. *Operations Research Letters*, 5:43–46.

Johnson, M. A. y Taaffe, M. R. (1991). An investigation of phase-distribution moment-matching algorithms for use in queueing models. *Queueing Systems*, 8:129–148.

Johnson, N. L. y Kotz, S. (1970). *Distributions in statistics. Continuous univariate distributions.* John Wiley and Sons, Nueva York.

Kendall, D. G. (1953). Stochastic processes occurring in the theory of queues and their analysis by the method of imbedded markov chains. *Annals of Mathematical Statistics*, 24:338–354.

Kitaev, M. Y. y Rykov, V. V. (1995). *Controlled queueing systems.* CRC Press, Boca Raton, FL.

Kleinrock, L. (1975). *Queueing systems.* John Wiley and Sons, Nueva York.

Klemm, A., Lindemann, C., y Lohmann, M. (2003). Modeling IP traffic using the Batch Markovian Arrival Process. *Performance Evaluation*, 54:149–173.

Kogan, Y., Levy, Y., y Milito, R. A. (1997). Call routing to distributed queues: Is FIFO better than MED? *Telecommunications Systems*, 7:299–312.

Koole, G. y Mandelbaum, A. (2002). Queueing models of call centers: An introduction. *Annals of Operations Research*, 113:41–59.

Lang, A. y Arthur, J. L. (1997). Parameter approximation for phase-type distributions. En Alfa, A. S. y Chakravarthy, S., eds., *Matrix-Analytic Methods in Stochastic Models*, págs. 151–206. Marcel Dekker.

Law, A. y Kelton, W. D. (1991). *Simulation Modeling and Analysis.* McGraw-Hill, Nueva York.

Lehoczky, J. (1990). Statistical methods. En Heyman, D. P. y Sobel, M. J., eds., *Stochastic Models*, págs. 255–293. Elsevier/North-Holland, New York y Amsterdam.

Leland, W., Taqqu, M., Willinger, W., y Wilson, D. (1994). On the self-similar nature of Ethernet traffic. *IEEE/ACM Transactions in Networking*, 2:1–15.

Lillo, R. E. (2000a). Characterization and computation of optimal policies for an M/G/1 priority queue. *The Belgian Journal of Operations Research, Statistics and Computer Science*, 38:45–57.

Lillo, R. E. (2000b). On optimal exhaustive policies for the M/G/1 queue. *Operations Research Letters*, 27:39–46.

Lillo, R. E. (2000c). On the optimal control of M/G/1 systems under the cycle criterion. *Systems & Control Letters*, 41:29–39.

Lillo, R. E. (2000d). Optimal operating policies for an M/G/1 exhaustive server-vacation model. *Methodology and Computing in Applied Probability*, 2:153–167.

Lillo, R. E. (2000e). Stability and irreducibility of queueing systems with finite capacity. *Queueing Systems*, 35:129–139.

Lillo, R. E. (2001). Optimal control of an M/G/1 queue with impatient priority customers. *Naval Research Logistics*, 48:200–209.

Lillo, R. E. y Neuts, M. F. (1999). Two service units with interference in the access to servers. *Journal of Applied Mathematics and Stochastic Analysis*, 12(4):357– 370.

Lucantoni, D. (1993). The BMAP/G/1 queue: A tutorial. En Donatiello, L. y Nelson, R., eds., *Models and Techniques for Performance Evaluation of Computer and Communication Systems*, págs. 330–358. Springer, Nueva York.

McGrath, M. F., Gross, D., y Singpurwalla, N. D. (1987). A subjective Bayesian approach to the theory of queues I - modeling. *Queueing Systems*, 1:317–333.

McGrath, M. F. y Singpurwalla, N. D. (1987). A subjective Bayesian approach to the theory of queues II - inference and information in m/m/1 queues. *Queueing Systems*, 1:335–353.

Mengersen, K. L. y Robert, C. P. (1996). Testing for mixtures: A Bayesian entropic approach. En Bernardo, J. M., Berger, J. O., Dawid, A. P., y Smith, A. F. M., eds., *Bayesian Statistics 5*, págs. 255–276. Oxford University Press.

Muddapur, M. V. (1972). Bayesian estimates of parameters in some queueing models. *Annals of the Institute of Statistical Mathematics*, 24:327–331.

Narayan, O. (1998). Exact asymptotic queue length distribution for fraccional Brownian traffic. *Advances in Performance Analysis*, 1:39–63.

Nelson, R. (1995). *Probability, Stochastic Processes and Queueing Theory.* Springer-Verlag, Nueva York.

Neuts, M. F. (1973). The single server queue in discrete time-numerical analysis i. *Naval Research Logistics Quarterly*, 20:297–304.

Neuts, M. F. (1977). Algorithms for the waiting time distributions under various queue disciplines in the M/G/1 queue with service time distributions of phase type. *TIMS Studies in the Management Sciences*, 7:177–197.

Neuts, M. F. (1981). *Matrix Geometric Solutions in Stochastic Models.* Johns Hopkins University Press, Baltimore.

Neuts, M. F. (1989). *Structured stochastic matrices of M/G/1 type and their applications.* Marcel Dekker, Nueva York.

Paxson, V. y Floyd, S. (1995). Wide-area traffic: The failure of Poisson modeling. *IEEE Transactions in Networking*, 3:226–244.

Peskun, P. H. (1973). Optimum Monte-Carlo sampling using Markov chains. *Biometrika*, 60:607–612.

Press, W. H., Teukolsky, S. A., Vetterling, W., y Flannery, B. P. (1989). *Numerical recipes : the art of scientific computing : (FORTRAN version).* Cambridge University Press, Cambridge.

Ralston, A. y Rabinowitz, P. (1978). *A first course in numerical analysis.* McGraw-Hill, Nueva York.

Ramaswami, V. (1982). The busy period of queues which have a matrix geometric steady state probability vector. *Opsearch*, 19:238–361.

Ramaswami, V. y Lucantoni, D. M. (1985). Stationary waiting time distribution in queues with phase type service and in quasi-birth-and-death processes. *Communications in Statistics. Stochastic Models*, 1:125–136.

Reynolds, J. F. (1973). On estimating the parameters of a birth-death process. *The Australian Journal of Statistics*, 15:35–43.

Richardson, S. y Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B*, 59:731–792.

Ripley, B. D. (1987). *Stochastic simulation.* John Wiley and Sons, Nueva York.

Robert, C. P., Rydén, T., y Titterington, D. M. (2000). Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. *Journal of the Royal Statistical Society, Series B*, 61(1):57–75.

Rodrigues, J. y Galvão, J. (1998). A note on Bayesian analysis in M/M/1 queues derived from confidence intervals. *Statistics*, 31:35–42.

Ríos, D., Wiper, M. P., y Ruggeri, F. (1998). Bayesian analysis of M/Er/1 and M/H_k/1 queues. *Queueing Systems*, 30:289–308.

Ríos, S., Ríos-Insua, S., y Ríos-Insua, M. J. (1989). *Procesos de decisión multicriterio*. Eudema, Madrid.

Ruggeri, F., Wiper, M. P., y Ríos, D. (1996). Bayesian models for correlations in M/M/1 queues. Technical Report 97.3, CNR-IAMI, Milán.

Rydén, T. (1996). An EM algorithm for estimation in Markov-Modulated Poisson Processes. *Computational Statistics and Data Analysis*, 21:431–447.

Rydén, T. (2000). Statistical estimation for Markov-modulated Poisson processes and Markovial arrival processes. En Latouche, G. y Taylor, P., eds., *Advances in Algorithmic Methods for Stochastic ModelsMatrix Analytic Methods for Stochastic Models*, págs. 329–350. Notable Publications.

Schruben, L. y Kulkarni, R. (1982). Some consequences of estimating parameters for the M/M/1 queue. *Operations Research Letters*, 1:75–78.

Scott, S. L. y Smyth, P. (2003). The Markov Modulated Poisson Process and Markov Poisson Cascade with applications to web traffic data. En Bayarri, M. J., Berger, J. O., Bernardo, J. M., Dawid, A. P., Heckerman, D., Smith, A. F. M., y West, M., eds., *Bayesian Statistics 7*, págs. 671–680. Oxford University Press.

Sengupta, B. (1989). Markov processes whose steady state distribution is matrix-exponential with an application to the GI/PH/1 queue. *Advances in Applied Probability*, 21:159–180.

Smith, A. F. M. y Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, 55:3–23.

Squillante, M. S. (1998). Matrix-analytic methods in stochastic parallel-server schedulling models. En Alfa, A. S. y Chakravarthy, eds., *Advances in Matrix Analytic Methods for Stochastic Models*, págs. 311–340. Notable Publications.

Stephens, M. (2000a). Bayesian analysis of mixture models with an unknown number of components — an alternative to reversible jump methods. *The Annals of Statistics*, 28(1):40–74.

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

Sze, D. Y. (1984). A queueing model for telephone operator staffing. *Operations Research*, 32:229–249.

Taylor, G. J., McClean, S. I., y Millard, P. H. (2000). Stochastic models of geriatric patient bed occupancy behaviour. *Journal of the Royal Statistical Society, Series B*, 163(1):39–48.

Tebaldi, C. y West, M. (1997). Bayesian inference on network traffic using link count data. *Journal of the American Statistical Association*, 93:557–576.

Thiruvaiyaru, D. y Basawa, I. V. (1992). Empirical Bayes estimation for queueing systems and networks,. *Queueing Systems*, 11:179–202.

Tijms, H. C. (1990). *Stochastic modelling and analysis : a computational approach*. John Wiley and Sons, Chichester.

Vehel, J. y Riedi, R. (1997). Fractional Brownian motion and data traffic modeling: The other end of the spectrum. En Vehel, L. y Lutton, E. y Tricot, C., eds., *Fractals in Engineering*. Springer.

Willinger, W., Taqqu, M., Leland, W., y Wilson (1995). Self-similarity in high-speed packet traffic: analysis and modelling of Ethernet traffic measurements. *Statistical Science*, 10:67–85.

Wiper, M. P. (1998). Bayesian analysis of Er/M/1 and Er/M/c queues. *Journal of Statistical Planning and Inference*, 69:65–79.

Wiper, M. P., Ríos, D., y Ruggeri, F. (2001). Mixtures of gamma distributions with applications. *Journal of Computational and Graphical Statistics*, 10:440–454.

Wolff, R. W. (1965). Problems of statistical inference for birth and death queueing models. *Operations Research*, 13:343–357.