Inferencia bayesiana para un modelo de colas general

Conchi Ausín Rosa Lillo Michael Wiper

Departamento de Estadística
Universidad Carlos III de Madrid

Las Palmas de Gran Canaria - 31 de enero de 2005

Transacciones comerciales en un banco Israeli

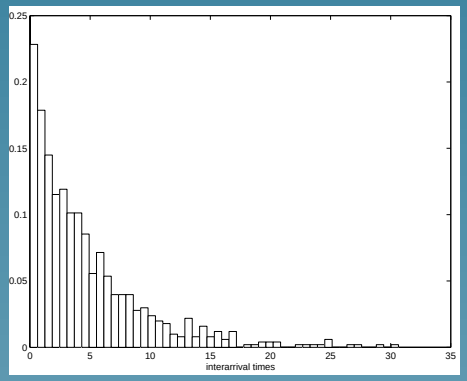
El banco tiene un único servidor de transacciones comerciales y luego tenemos un sistema (FIFO) de colas G/G/1.

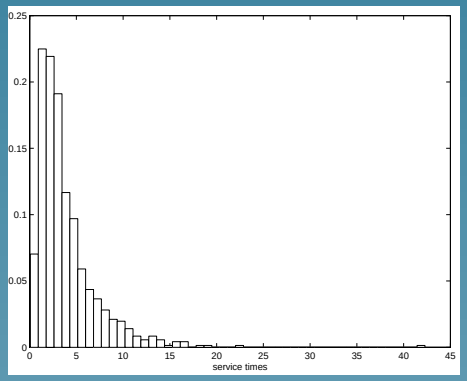
Temas de interés:

- Modelizar tiempos entre llegadas y tiempos de servicio.
- Estabilidad del sistema.
- Distribuciones transitorias (y estacionarias).

Los datos

Tiempos entre llegadas y tiempos de servicio durante una semana.





No tenemos un sistema M/M/1.

Modelización de tiempos entre llegadas y tiempos de servicio

Modelos simples (exponencial, híper-exponencial o Erlang) no se ajustan bien a estos datos.

¿Cuál es la alternativa?

- Modelización no-paramétrica: no sabemos nada sobre las propiedades del sistema de colas.
- Modelización semi-paramétrica:

La distribución coxiana o MGE

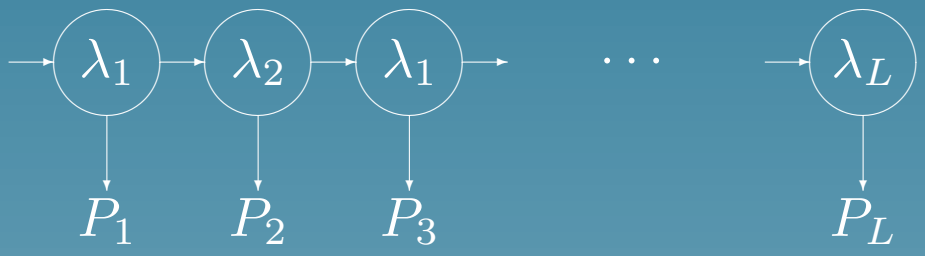
Si X es un tiempo entre llegadas:

$$X = \begin{cases} Y_1 & \text{con probabilidad } P_1 \\ Y_1 + Y_2 & \text{con probabilidad } P_2 \\ Y_1 + Y_2 + Y_3 & \text{con probabilidad } P_3 \\ \vdots & \vdots \\ \sum_{r=1}^{L} Y_r & \text{con probabilidad } P_L \end{cases}$$

donde $Y_r \sim \text{Exp}(\lambda_r)$ y $\sum_{r=1}^{L} P_r = 1$.

Representación tipo fase

La distribución está compuesta de varias fases exponenciales, donde es posible salir en cualquiera fase.



La distribución coxiana como mixtura

Se tiene

$$f\left(x \mid L, \mathbf{P}, \lambda\right) = \sum_{r=1}^{L} P_r f_r\left(x \mid \lambda_1, ..., \lambda_r\right), \tag{1}$$

donde $f_r\left(x \mid \lambda_1, ..., \lambda_r\right)$ es la densidad Erlang generalizada,

$$f_r\left(x \mid \lambda_1, ..., \lambda_r\right) = \sum_{j=1}^{r} \left(\prod_{\substack{i=1 \\ i \neq j}}^{r} \frac{\lambda_i}{\lambda_i - \lambda_j} \right) \lambda_j \exp\left(-\lambda_j x\right), \tag{2}$$

Propiedades atractivas

- Es una distribución tipo fase (PH).
- Incluye las distribuciones exponencial ($L = 1$) y Erlang ($\lambda_r = \lambda$, $P_r = 0$ para $r < L$) como casos particulares.
- Incluye el modelo híper-exponencial.
- Aumentando el valor de L, se aproxima cualquiera distribución con soporte $\Re^+$.

Inferencia bayesiana para la distribución coxiana

Observamos una muestra de tiempos entre llegadas, $\{x_1, \ldots, x_n\}$ (y otra muestra de tiempos entre servicios).

Para cada X_i, sea Z_i una indicadora de la fase en que sale. Luego

$$\begin{aligned} P(Z_i = r|L, \mathbf{P}) &= P_r \\ f(x_i|z_i = r, \boldsymbol{\lambda}) &= f_r\left(x_i \mid \lambda_1, ..., \lambda_r\right) \end{aligned}$$

es la densidad Erlang generalizada (2).

Ausín et al (2003) usan distribuciones a priori propias sobre los parámetros $L, \boldsymbol{\lambda}, \mathbf{P}$ y condicionan sobre las otras variables latentes Y_{ir} = tiempo pasado en fase r. Aquí introducimos una esquema más flexible.

Reparameterización del modelo

Supongamos que L es conocido y impongamos una restricción de orden $\lambda_1 > \ldots > \lambda_L$. Entonces, podemos reparameterizar el modelo:

$$\lambda_i = \tau_i \lambda_{i-1} \text{ para } i \geq 1$$

donde $\tau_1 = 1$ y $0 < \tau_i < 1$ para $i = 2, \ldots, L$. Luego, la distribución Erlang generalizada es

$$f_r(x \mid \mathbf{P}, \lambda_1, \boldsymbol{\tau}) = \sum_{j=1}^{r} \left(\prod_{\substack{i=1 \\ i \neq j}}^{r} \frac{\prod_{k=1}^{i} \tau_k}{\prod_{k=1}^{i} \tau_k - \prod_{k=1}^{j} \tau_k} \right) \lambda_1 \prod_{k=1}^{j} \tau_k \exp\left(-\lambda_1 \prod_{k=1}^{j} \tau_k x \right)$$

Distribuciones a priori

Distribuciones a priori poca informativas:

$$
\begin{aligned}
f(\lambda_1) &\propto \frac{1}{\lambda_i} \\
\tau_i &\sim \text{Beta}(1{,}1, 1{,}1) \quad \text{para } i = 1, \ldots, L \\
\mathbf{P} &\sim \text{Dirichlet}\,(1, \ldots, 1)
\end{aligned}
$$

Distribuciones a posteriori

$$
\begin{aligned}
P(Z_i = r \mid x_i, \lambda_1, \boldsymbol{\tau}, \mathbf{P}) &\propto P_r f_r\left(x_i \mid \lambda_1, \tau_2, \ldots, \tau_r\right) \\
\mathbf{P} \mid \mathbf{z} &\sim \text{Dirichlet}(1 + n_1, ..., 1 + n_L) \quad \text{donde } n_r = \#\{z = r\} \\
f\left(\lambda_1 \mid \mathbf{x}, \mathbf{z}, \boldsymbol{\tau}\right) &\propto \frac{1}{\lambda_1} \prod_{i=1}^{n} f_{z_i}\left(x_i \mid \lambda_1, \tau_2, \ldots, \tau_{z_i}\right) \\
f\left(\tau_r \mid \mathbf{x}, \mathbf{z}, \lambda_1, \boldsymbol{\tau}_{-r}\right) &\propto \prod_{\substack{i=1 \\ z_i \geq r}}^{n} f_{z_i}\left(x_i \mid \lambda_1, \ldots, \tau_2, \ldots, \tau_{z_i}\right), \quad \text{para } r = 2, ..., L.
\end{aligned}
$$

Muestreo Gibbs

1. Valores iniciales $\mathbf{P}^{(0)}$, $\lambda_1^{(0)}$, $\boldsymbol{\tau}^{(0)}$. $t = 1$.
2. Muestrear $Z_i^{(t)} \sim Z_i \mid x_i, \lambda_1^{(t-1)}, \boldsymbol{\tau}^{(t-1)}, \mathbf{P}^{(t-1)}$
3. Muestrear $\mathbf{P}^{(t)} \sim \mathbf{P} \mid \mathbf{z}^{(t)}$.
4. Muestrear $\lambda_1^{(t)} \sim \lambda_1 \mid \mathbf{x}, \mathbf{z}^{(t)}, \boldsymbol{\tau}^{(t-1)}$.
5. Muestrear $\tau_r^{(t)} \sim \tau_r \mid \mathbf{x}, \mathbf{z}^{(t)}, \lambda_1^{(t)}, \boldsymbol{\tau}_{-r}^{(t-1)}$ para $r = 2, \ldots, L$.
6. $t = t + 1$. Ir a 2.

Pasos Metropolis

λ_1 y $\boldsymbol{\tau}$ tienen distribuciones no estandares pero es fácil usar el método Metropolis:

- Para λ_1, se genera un candidato $\tilde{\lambda}$ de una distribución gamma con media $\lambda_1^{(t-1)}$.
- Para τ_r, se genera un candidato de una mixtura de dos distribuciones beta con media $\tau_r^{(t-1)}$.
- Se acepta el candidato $\tilde{\lambda}$ con probabilidad

$$\text{mín}\left\{1, \frac{f(\tilde{\lambda} \mid \mathbf{x}, \mathbf{z}^{(t)}, \boldsymbol{\tau}^{(t-1)})}{f(\lambda_1^{(t-1)} \mid \mathbf{x}, \mathbf{z}^{(t)}, \boldsymbol{\tau}^{(t-1)})} \frac{g(\lambda_1^{(t-1)} \mid \tilde{\lambda})}{g(\tilde{\lambda} \mid \lambda_1^{(t-1)})}\right\}$$

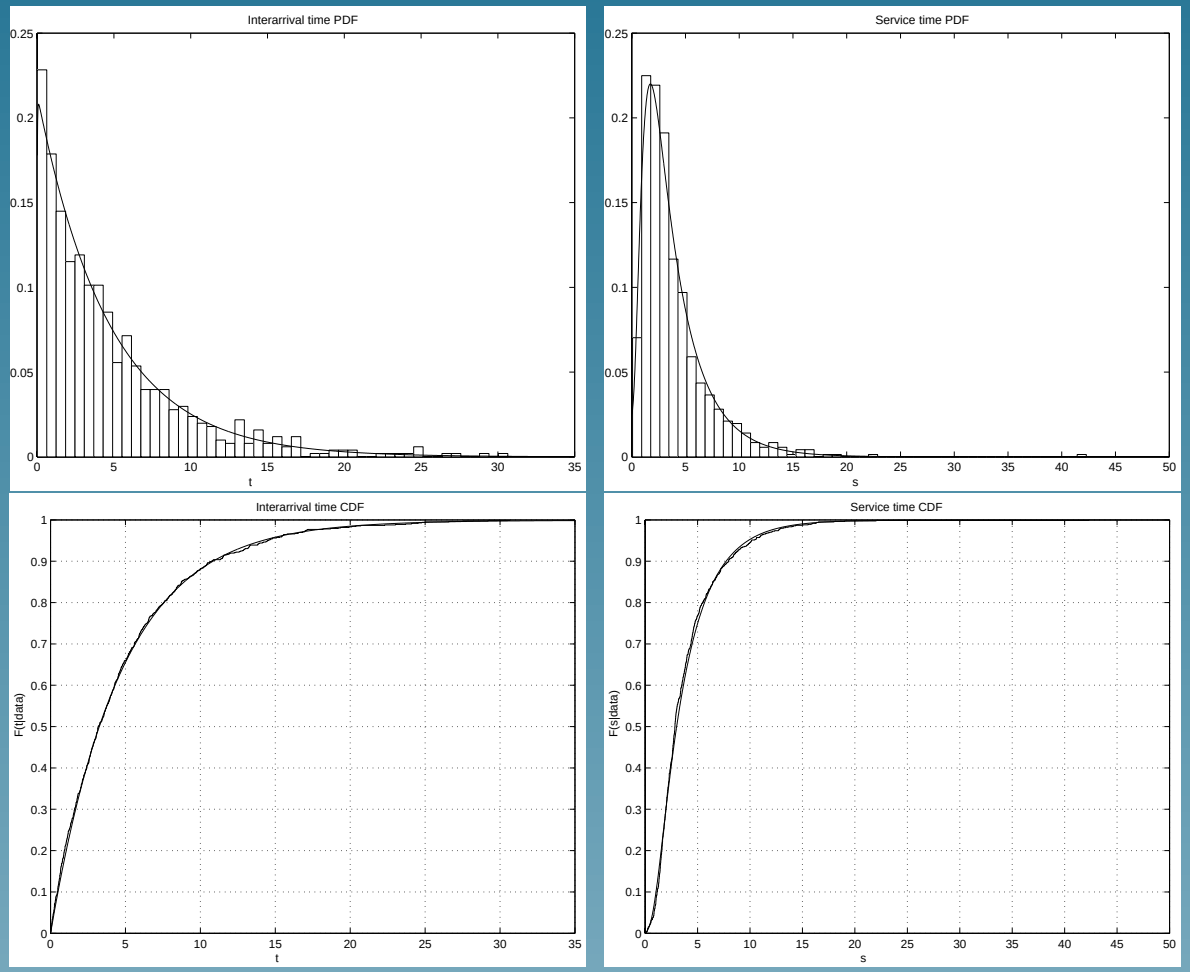
Extension a L desconocido

Definimos una distribución a priori $P(L)$ y usamos un algoritmo salto reversible (Green 1995, Richardson y Green 1997) para muestrear la distribución a posteriori.

Idea:

- En cada paso del algoritmo Gibbs, proponemos un cambio (crecimiento o decrecimiento) del parámetro L.
- Si proponemos un decrecimiento, se combinan dos componentes adyacentes de la mixtura.
- Para un crecimiento, se divide un componente en 2.

Densidades y funciones de distribución ajustadas



El sistema de colas $MGE/MGE/1$

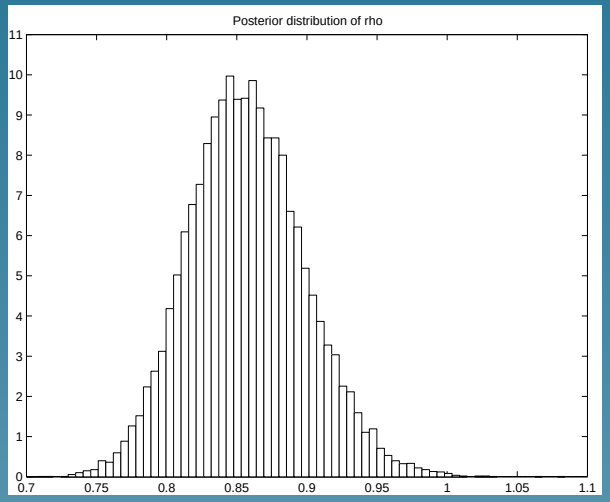
Es el sistema estable?

Dados los parámetros, la intensidad de tráfico del sistema es el tiempo de servicio esperado partido por el tiempo esperado entre llegadas

$$\begin{aligned}\rho &= \frac{E[S|L_S, \mathbf{P}_S, \boldsymbol{\lambda}_S]}{E[A|L_A, \mathbf{P}_A, \boldsymbol{\lambda}_A]} \\ &= \frac{\sum_{r=1}^{L_S}(1-\sum_{i=1}^{r-1} P_{Si})\frac{1}{\lambda_{Sr}}}{\sum_{r=1}^{L_A}(1-\sum_{i=1}^{r-1} P_{Ai})\frac{1}{\lambda_{Ar}}}\end{aligned}$$

Luego se estima la intensidad de tráfico mediante la muestra MCMC.

Resultados



El gráfico muestra un histograma de las valores de ρ generados por el algoritmo MCMC.

Probabilidad de estabilidad y esperanza de ρ

- La probabilidad de que el sistema sea estable es muy alta; $P(\rho < 1|\text{data}) = 0{,}999$.
- La intensidad de tráfico esperada es bastante alta; $E[\rho|\text{data}] = 0{,}86$.

Estimación de las características del sistema

Adaptamos resultados debidos a Bertsimas y Nakazato (1992) para estimar las distribuciones transitorias de tamaño de la cola, y tiempo de espera de clientes y la duración de un periodo de ocupación. Todas las distribuciones dependen de las transformadas de Laplace Stieltjes (LT) de las distribuciones de tiempos entre llegadas y servicios.

$$f_A^\star(s) = \sum_{r=1}^{L_A} P_{Ar} \prod_{i=1}^{r} \left(\frac{\lambda_{Ai}}{\mu_{Ai} + s} \right)$$

$$f_S^\star(s) = \sum_{r=1}^{L_S} P_{Sr} \prod_{i=1}^{r} \left(\frac{\lambda_{Si}}{\lambda_{Si} + s} \right)$$

Duración de un periodo de ocupación

Dados $\{(L_A, \mathbf{P}_A, \boldsymbol{\lambda}_A), (L_S, \mathbf{P}_S, \boldsymbol{\lambda}_S)\}$, la LT de la función de distribución de la duración de un periodo de ocupación, B, es:

$$\begin{aligned} F_B^\star(s) &= \int_0^\infty e^{-st} F_B(t) dt \\ &= 1 - (1 - f_B^\star(s)) \frac{\prod_{k=1}^{M} (s + \mu_k)}{\prod_{r=1}^{M} (s - x_r(s))} \end{aligned}$$

donde $x_r(s)$ $(r = 1, \ldots, L_S)$ son las raices de la ecuación

$$\left\{ \begin{array}{c} f_A^\star(s - x) f_B^\star(x) = 1 \\ \mathrm{Re}(x) > 0 \end{array} \right\} \tag{3}$$

donde $f_A^\star(s)$ y $f_B^\star(s)$ son las LT's de las distribuciones de tiempos entre llegadas y tiempos de servicio.

Distribuciones transitorias del número de clientes en el sistema y el tiempo de espera

- Supongamos que el sistema sea estable.
- Supongamos que el sistema está vacio cuando se abre el banco.
- Fórmulas complicadas que dependen de productos tensoriales de transformadas de Laplace.
- Además dependen de las raices $x_r(s)$.

Problemas

Para calcular las distribuciones predictivas del número de clientes etc. en cada iteración del algoritmo MCMC, necesitamos

1. Invertir transformadas de Laplace

2. Encontrar las raices de la ecuación (3).

Puede ser muy costoso en tiempo.

Inversión de transformadas

Podemos considerar dos posibilidades:

- En cada iteración MCMC, calculamos la LT y hallamos la LT predictiva como promedio de LT's, invertiendola al final. Es un método más rápido pero bastante impreciso.

- Calculamos y invertimos la LT en cada iteración. Hallar la distribución predictiva como promedio de las distribuciones halladas en cada iteración.

Utilizamos el eficiente algoritmo de Hosono (1981) para invertir las transformadas.

Búsqueda de raices

Encontrar las raices es fácil dados los parámetros mediante cálculo simbólico en Mathematica, pero complicado dentro de un algoritmo MCMC.

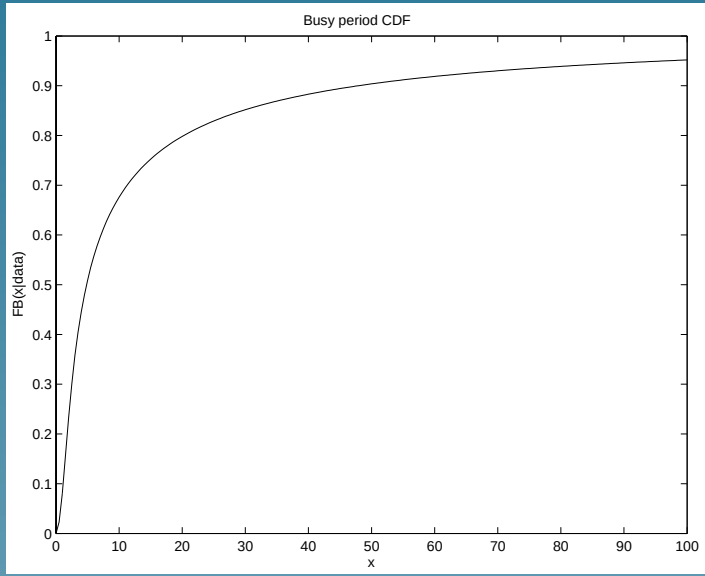
Teorema

El problema de calcular los raices de (3) equivale al problema de encontrar los L_S raices con parte real positiva del polinomio de orden $L_A + L_S$

$$\begin{aligned}\mathcal{P}(x) \quad = \quad & \left[\sum_{r=1}^{L_A}\sum_{t=1}^{L_S} P_{Ar}P_{Ss}\left(\prod_{i=1}^{r}\lambda_{Ai}\right)\left(\prod_{i=1}^{L_A}(\lambda_{Ai}+s-x)\right)\right. \\ & \left.\left(\prod_{j=1}^{t}\lambda_{Sj}\right)\left(\prod_{j=t+1}^{L_S}(\lambda_j+x)\right)\right] - \\ & \left(\prod_{r=1}^{L_A}(\lambda_{Ar}+s-x)\right)\left(\prod_{t=1}^{L_S}(x+\lambda_{St})\right)\end{aligned}$$

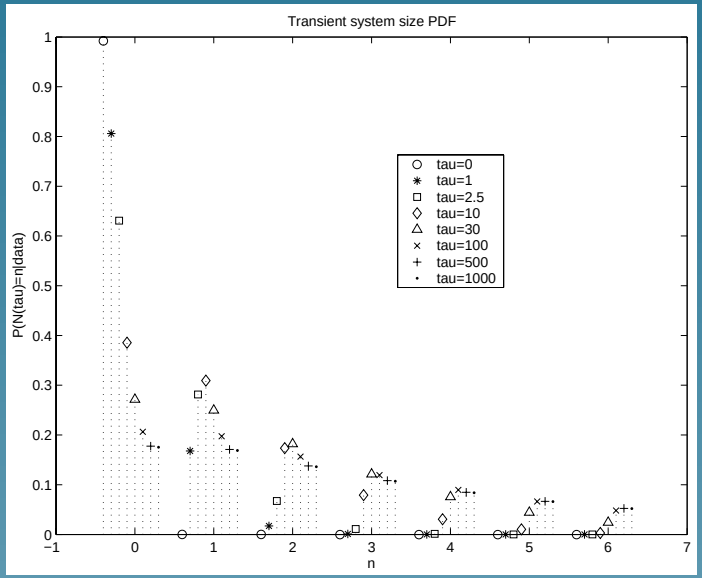
donde es fácil calcular los coeficientes del polinomio. Sacar raices de un polinomio no es tan difícil usando el algoritmo de Laguerre.

La distribución de un periodo de ocupación



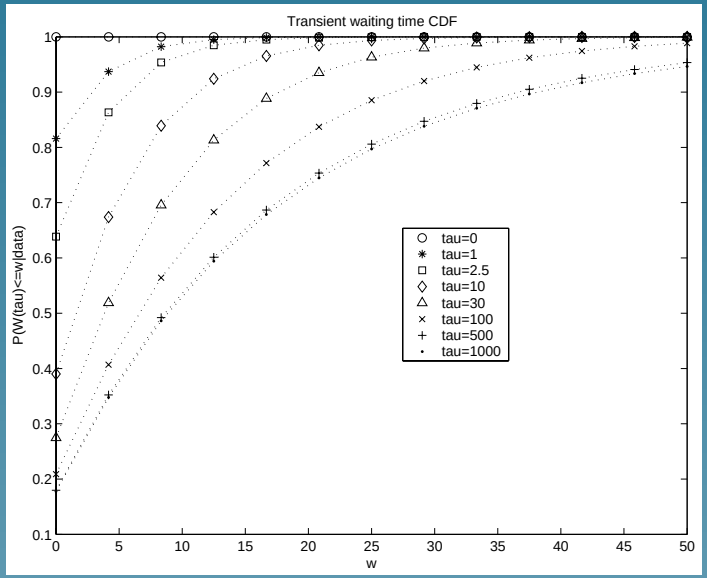
El gráfico Existe una 70 % probabilidad de que un periodo de ocupación dure menos de 10 minutos y una probabilidad de 10 % de que dure más de una hora.

Números de clientes



La distribución transitoria converge lentamente a la distribución estacionaria.

Tiempo de espera



Otra vez la convergencia es lenta. Análisis estacionaria del sistema parece insuficiente.

Extensiones

- Demostrar que la distribución a posteriori es propia y que la esperanza $E[X \mid \mathbf{x}]$ existe.
- No se han considerado efectos temporales. Existe una ligera dependencia entre llegadas.
- Combinamos con métodos de análisis de decisiones para decidir si se necesitan más servidores.

Bibliografía

1. Ausín, C., Lillo, R., Ruggeri, F. y Wiper, M. (2003). Bayesian modeling of hospital bed occupancy times using a mixed generalized Erlang distribution. En *Bayesian Statistics 7*, 443-452.

2. 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.

3. Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711-732.

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

5. Richardson, S. y Green, P. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J.R.S.S., B*, **61**, 57-75.

BAYESIAN STATISTICS 7, pp. 000–000
J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid,
D. Heckerman, A. F. M. Smith and M. West (Eds.)
© Oxford University Press, 2003

Bayesian Modeling of Hospital Bed Occupancy Times using a Mixed Generalized Erlang Distribution

M. CONCEPCIÓN AUSÍN
Universidad Carlos III de Madrid, Spain
causin@est-econ.uc3m.es

ROSA E. LILLO
Universidad Carlos III de Madrid, Spain
lillo@est-econ.uc3m.es

FABRIZIO RUGGERI
CNR-IMATI, Italy
fabrizio@iami.mi.cnr.it

MICHAEL P. WIPER
Universidad Carlos III de Madrid, Spain
mwiper@est-econ.uc3m.es

SUMMARY

In this paper we model the distribution of length of stay in hospital of geriatric patients. We assume that length of stay has a mixed generalized Erlang distribution and, given data from patients in a geriatric ward of a London hospital, reversible jump methods are used to estimate the predictive distribution of length of stay in hospital. We also address the problem of optimizing the number of beds in hospital with respect to average cost per unit of time where costs are based on lost demand, the number of unoccupied beds and the different types of patients in the hospital.

Keywords: MIXED GENERALIZED ERLANG DISTRIBUTION; $M/G/C$ LOSS SYSTEM; REVERSIBLE JUMP MCMC.

1. INTRODUCTION

The distribution of the time that geriatric patients spend in hospital seems to have a complex behaviour. Patients are admitted and subject to an initial acute care which may be followed by a number of different stages of treatment. Most patients are discharged (or die) after a relatively short period of time. However, some patients may remain in hospital over a long time period receiving continuous attention. This suggests that the distribution of lengths of stay in hospital is likely to be very heterogeneous.

Figure 1 shows boxplots of the distribution of the lengths of stay of 1092 geriatric patients at St. George's Hospital in London over the period 1965-1984. The data are a subset of the sample analyzed by e.g. Taylor *et al.* (2000) and may be downloaded

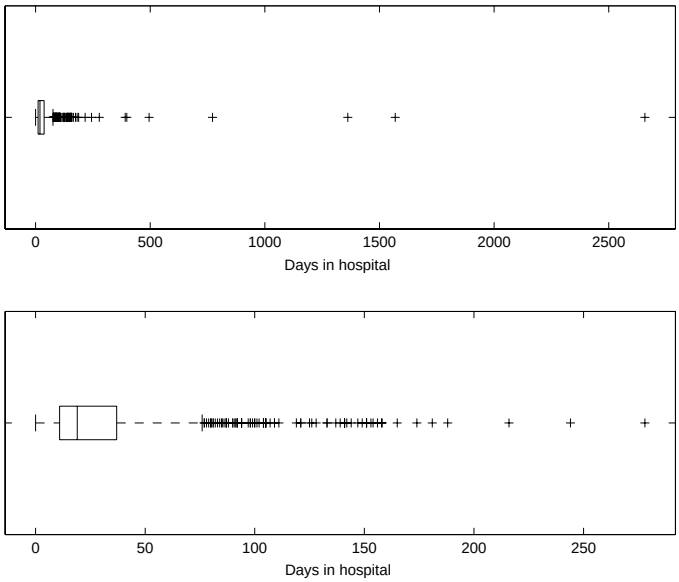


Figure 1. *Boxplots of days spent in hospital for all patients (top) and for patients staying under 300 days (bottom).*

from the Royal Statistical Society, `http://www.blackwellpublishers.co.uk/rss/`. We can observe from the boxplots that the data are very highly skewed to the left with a number of large outliers.

The diverse patterns of duration of stay in hospital require distinct resources and suitable organization and consequently, for hospital planning, it is important to capture the shape of the distribution of the length of stay. As patients undergo various stages of treatment in hospital, this suggests the use of phase–type distributions, see Neuts (1981), for modeling the distribution of stay. Such models essentially assume that the total time spent by a patient in hospital can be decomposed into a number of different, exponential, phases. See also Faddy (1990,1994).

A number of authors have used classical techniques to fit a variety of phase–type models to (versions of) the St. George's Hospital data. For example, Harrison and Millard (1991) and Gorunescu *et al.* (1999) considered exponential mixture models and Faddy and McClean (1999) used the more general class of mixed, generalized Erlang (MGE) distributions although they assumed a fixed number of phases. One of the main difficulties with such classical models is that they are not designed to deal with situations where the model dimension, e.g. the number of phases in the MGE model, is unknown. However, in such problems, Bayesian inference can often be carried out using variable dimension MCMC methods. Here, in Section 2 we also assume an MGE model and we show how Bayesian inference can be undertaken using reversible jump techniques under the assumption that the number of phases are unknown.

One of the motivations for studying the length of stay in hospital is to attempt to optimize the number of beds, c. The decision of how many beds to put in a ward will typically be based on the both the costs of maintaining beds and the (opportunity) costs of losing prospective patients because the ward is full, assuming that in this case, prospective patients are lost rather than put on a waiting list. Gorunescu *et al.* (1999) analyze this problem using classical techniques. Here, in Section 3, we assess the optimization of the number of beds assuming a slightly more complicated cost function that also takes into account the different costs for short and long stay patients.

In Section 4, we finish with a brief discussion and some possible extensions to other fields.

2. MODELING AND INFERENCE FOR TIME SPENT IN HOSPITAL

In this section, we consider modeling and inference for the time spent in hospital by a patient, T. We assume that T follows a MGE distribution, that is

$$T = \begin{cases} X_1 & \text{with probability } P_1 \\ X_1 + X_2 & \text{with probability } P_2 \\ \vdots & \vdots \\ X_1 + \ldots + X_L & \text{with probability } P_L \end{cases} \tag{1}$$

where $X_r \sim \text{Ex}(\mu_r)$ for $r = 1, \ldots, L$ and $\sum_{r=1}^{L} P_r = 1$.

Thus, from (1), the density of T has a mixture form;

$$f(t \,|\, L, \boldsymbol{P}, \boldsymbol{\mu}) = \sum_{r=1}^{L} P_r f_r(t \,|\, \boldsymbol{\mu})$$

where $f_r(t \,|\, \boldsymbol{\mu})$ is the density function of a sum of r exponentials, or a generalized Erlang distribution

$$f_r\left(t \,|\, L, \boldsymbol{\mu}\right) = \sum_{j=1}^{r} \left(\prod_{s \neq j} \left(\frac{\mu_s - \mu_j}{\mu_s \mu_j} \right)^{-1} \right) \mu_j^{2-r} e^{-\mu_j t}. \tag{2}$$

see e.g. Johnson and Kotz (1970). Note that in (2) it is assumed that all μ_r, $r = 1, \ldots, L$ are distinct. Alternative formulae are available in the case where there is some repetition; see Johnson and Kotz (1970).

This distribution is a phase–type distribution of order L and contains the exponential ($L = 1$) and Erlang ($P_L = 1$, $\mu_1 = \ldots = \mu_L$) distributions as special cases. By increasing the number of phases, it is possible to approximate any (strictly continuous) density function over the positive real line using an MGE distribution. In our case, we will assume that all parameters, including L, are unknown.

Given that we observe the times spent in hospital of n patients, $\boldsymbol{t} = \{t_1, ..., t_n\}$, independently, we now wish to carry out Bayesian inference for this model. The likelihood function takes a very complicated form but can be simplified by introducing the latent variables; Z_i is the phase in which the i'th patient leaves the hospital and X_{ir} is the time spent by the i'th patient in phase r, for $r = 1, \ldots, Z_i$ where $\sum_{r=1}^{Z_i} X_{ir} = T_i$. Then, we have

$$f(t_i, z_i, x_{i1}, \ldots, x_{iz_i} \,|\, \boldsymbol{\theta}) = P_{z_i} \prod_{r=1}^{z_i} \mu_r \exp(-\mu_r x_{ir})$$

where $\boldsymbol{\theta} = (L, \boldsymbol{P}, \boldsymbol{\mu})$.

In order to carry out Bayesian inference, we also need to define prior distributions for the model parameters $\boldsymbol{\theta}$. We assume the following prior dependence structure,

$$f(\boldsymbol{\theta}) = f(L) f(\boldsymbol{P} \,|\, L) f(\boldsymbol{\mu} \,|\, L).$$

For the number of phases, L, many prior distributions can be considered. Here, we assume a translated Poisson distribution; $L - 1 \sim \text{Po}(1)$. This choice has the advantage

of penalizing overparameterisation by giving low probability a priori to large numbers of phases, L.

Conditional on L, we use semi–conjugate prior distributions for $(\boldsymbol{P}, \boldsymbol{\mu})$; a Dirichlet distribution for the weights, $\boldsymbol{P} \,|\, L \sim \text{Di}(1, \ldots, 1)$ and independent exponential distributions, $\mu_r \sim \text{Ex}(0.01)$, for $r = 1, ..., L$. These prior distributions are used mainly for convenience. A possible extension would be to use a hierarchical structure for the prior on $\boldsymbol{\mu}$.

2.1 *Sampling from the Posterior*

Given the data and prior distributions defined earlier, the posterior distribution can be sampled via typical MCMC methods for mixture distributions, using a Gibbs sampler to sample the parameter distributions conditional on L and a reversible jump method to let the chain move through the posterior distribution of L.

Given the latent variables, it is straightforward to show that the conditional posterior of $\boldsymbol{P}$ is still a Dirichlet distribution and the conditional distributions of the μ_r are gammas.

The latent variables can be sampled by sampling the two components of

$$f(z_i, x_{i1}, \ldots, x_{iz_i} \,|\, t_i, \boldsymbol{\theta}) = f(x_{i1}, \ldots, x_{iz_i} \,|\, t_i, z_i, \boldsymbol{\theta}) f(z_i \,|\, t_i, \boldsymbol{\theta})$$

for $i = 1, \ldots, n$.

The conditional posterior $f(z_i \,|\, t_i, \boldsymbol{\theta})$ is easily derived from (2). It is slightly more complicated to sample from $f(x_{i1}, \ldots, x_{iz_i} \,|\, t_i, z_i, \boldsymbol{\theta})$. This distribution is a product of z_i exponentials restricted to the subspace $\sum_{r=1}^{z_i} x_{ir} = t_i$. Assuming, without loss of generality, that $\min\{\mu_r\} = \mu_j$ and assuming that $\mu_r \neq \mu_j$ for all $r \neq j$, we have

$$f\left(x_{i1}, \ldots, x_{ij-1}, x_{ij+1}, \ldots, x_{i,z_i} \,|\, t_i, z_i, \boldsymbol{\theta}\right) \propto \prod_{\substack{r=1 \\ r \neq j}}^{z_i} (\mu_r - \mu_j) \exp\left\{-(\mu_r - \mu_j)\, x_{ir}\right\}$$

defined over the region $\boldsymbol{R} = \left\{x_{i1} + \ldots + x_{ij-1} + x_{ij+1} + \ldots + x_{i,z_i-1} \leq t_i\right\}$. (A similar formula is available in the case where $\boldsymbol{\mu}$ has multiple minima.) This density can be sampled straightforwardly by, e.g. rejection sampling, see e.g. Ripley (1987).

In order to sample over different numbers of phases L, we can follow the split–combine procedure for mixtures of unknown number of components introduced by Richardson and Green (1997). The proposed moves change the order, L, by one unit in such a manner that in the MGE model, any phase can be split into two or any two consecutive phases can be combined into one. The corresponding changes to the parameters are done such that the marginal distribution of T is preserved. If a combine move is proposed, the latent variables $\left(z_i, x_{i1}, \ldots, x_{iz_i}\right)$ (for $i = 1, \ldots, n$) are changed in the natural way and if a split move is proposed, they are modified analogously to the Gibbs allocation technique described earlier.

2.2 *Results for the St. George's Hospital Data*

Figure 2 shows a histogram of the observed data truncated onto times less than 300 days. The predictive density (solid line) calculated from an MCMC run of 100000 iterations in equilibrium is overlaid. This is compared with a classical density estimate (dashed line) based on the use of an EM-algorithm by Asmussen *et al.* (1996) to estimate the

MGE model with $L = 5$ phases. This method was used in an earlier study of these data by Faddy and McClean (1999). Both density estimates appear very similar. This result is to be expected here as fairly uninformative priors have been used within the Bayesian model.

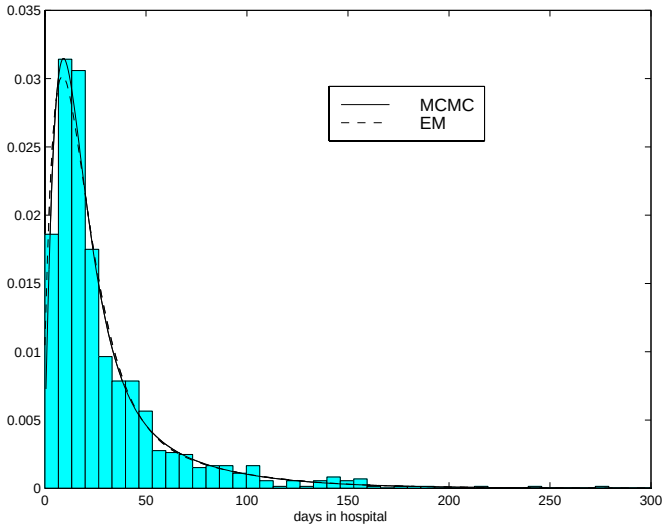


Figure 2. *Densities estimated using the MCMC algorithm and the EM algorithm for a fixed number of phases equal to 5.*

The total computation time for the MCMC algorithm was approximately one and a half hours on a Unix workstation. For the EM algorithm, the computation times are sensitive to the value of L chosen. For $L = 1$, computation is almost immediate whereas for $L = 10$, this can take 20 minutes or more. Thus, the overall computation times for the two methods are comparable.

Notice that one advantage of the use of the variable dimension MCMC approach as opposed to the use of the EM method is that the number of phases L does not have to be fixed a priori. Also the uncertainty in the model is reflected directly in the Bayesian approach via the posterior distribution for L, see Figure 3, which means that the predictive density is an average over the various different models. A final advantage of the Bayesian approach is that it gives a natural way of electing an optimum number of beds in the hospital, and of assessing the uncertainty in this selection, via the use of decision theory methods. See Section 3.

Figure 3 shows the posterior probabilities for the total number of phases L. We can observe that the highest probabilities are assigned to $L = 5$ and $L = 6$. This corresponds well with the results of Faddy and McClean (1999). It should be noted that there is little sensitivity to the choice of prior for L in the value of the posterior mode of L. Experiments using different translated Poisson priors have produced very similar results. For example, varying the prior mean of L between 2 and 4, we still find that most of the posterior probability mass is concentrated on the values of $L = 4$ to 8, although the posterior mode varies between 5 and 7. Small changes in the prior distributions of the remaining parameters appear to have little effect. Furthermore, as we would expect, the predictive density estimation of the time spent in hospital by a patient is unaffected by these changes to the priors.

Finally, note that the mean number of days spent by a patient in hospital is estimated from the MCMC output to be around 37.3 (s.d. $= 4$).

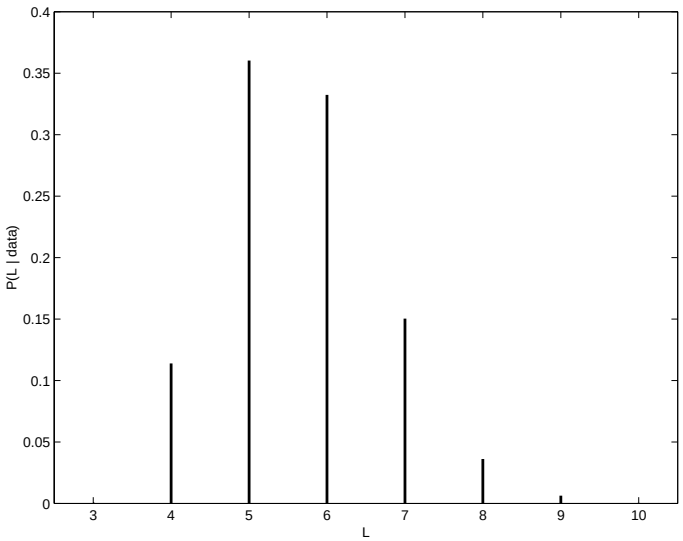


Figure 3. *Probabilities for different numbers of phases L.*

3. OPTIMIZING THE NUMBER OF BEDS IN HOSPITAL

In this section, our aim is to formulate a cost function which can be used to optimize the number of beds in the hospital, c. We assume here that patients arrive at the hospital according to a simple Poisson process with rate λ and that when no beds are available, these patients do not join a waiting list but instead are lost to the system, being admitted to other hospitals for example. The most important costs to be considered will be due to having insufficient beds or having too many empty beds and we therefore need to estimate the probabilities that various numbers of beds are occupied.

Under these conditions, the number of patients in the hospital can be modeled as a $M/G/c$ loss system, that is, a queueing system with Poisson arrivals, general service distribution, c servers and fixed capacity c, i.e. no queueing; see e.g. Tijms (1990). Here the general service time is identified with the length of the stay and its distribution is approximated by the MGE model, introduced in Section 2. The capacity of the system is equal to the number of beds, c and subsequently there is no queueing.

The average number of patients arriving per unit of time is λ, and the mean time that patients stay in hospital (service time) is,

$$\mathrm{E}[T \mid \boldsymbol{\theta}] = \sum_{r=1}^{L} \left(1 - \sum_{j=1}^{r-1} P_j \right) \frac{1}{\mu_r}.$$

The *offered load*, a of the system is defined to be the mean number of arrivals in a service (patient stay) time, that is $a = \lambda \mathrm{E}[T \mid \boldsymbol{\theta}]$.

The stationary distribution of the number of busy servers (beds), N, in a $M/G/c$ loss system is given by, see e.g., Tijms (1990),

$$P(N = j \mid a) = \frac{a^j/j!}{\sum_{k=0}^{c} a^k/k!}, \qquad k = 0, ..., c.$$

This is a truncated Poisson distribution defined on $[0, c]$. The probability that an arriving patient finds all c beds occupied is given by $B(c, a)$ where

$$B(c, a) = \frac{a^c/c!}{\sum_{k=0}^{c} a^k/k!} \tag{3}$$

see e.g. Cohen (1976). This formula, known as *Erlang's loss formula*, corresponds to the fraction of prospective patients that is lost due to having insufficient beds.

The mean number of busy beds, can also be obtained, see for example, Tijms (1990),

$$\mathrm{E}[N \,|\, a] = a\,[1 - B(c, a)]\,. \tag{4}$$

We can now deal with the problem of deciding the optimal number of beds in hospital. We will consider three possible costs. Firstly we assume an opportunity cost, $\pi > 0$, for each patient that is lost because no beds are available. The proportion of patients that can not be admitted is given by the Erlang loss formula and hence, the average cost per day caused by lost patients is

$$\pi\lambda B\,(c, a)\,.$$

Secondly we assume a holding cost, $h > 0$, for each empty bed per day. The average number of empty beds is equal to the total number of beds, c minus the average number of busy beds, given by (4). Thus, the average cost per unit of time due to unoccupied beds is given by,

$$h\,\{c - a[1 - B(c, a)]\}$$

Finally we consider a patient cost, r, for each patient in hospital (busy bed) per day. We consider different costs associated with the pattern of length of stay. Assume, for example, different costs, r_S, r_M, r_L, for 'short–stay' (S), 'medium–stay' (M) and 'long–stay' (L) patients, respectively. This will typically be the case in practice when some patients will enter the hospital for urgent and expensive treatment such as operations etc. and may then leave relatively quickly when cured, whereas others may enter because of general poor health, when they may spend a long period in hospital but without needing intensive treatment. By using the service time distribution, T, we can easily estimate the proportion of patients of every class, p_S, p_M and p_L. From (4), the average number of beds occupied by each type of patient at any given time will by given by $p_k a[1 - B(c, a)]$; for $k = S, M$ or L. Then, the average cost per day due to occupied beds will be,

$$ra[1 - B(c, a)]$$

where $r = r_S p_S + r_M p_M + r_L p_L$.

Combining these cost functions we have that the average cost function per unit of time given that there are c beds is

$$g(c) = \pi\lambda B(c, a) + h\{c - a[1 - B(c, a)]\} + ra[1 - B(c, a)]$$

Note that as c grows, the Erlang loss formula $B(c, a)$ approaches zero and thus, the cost function will approach be approximately linearly increasing for large c. It can also be shown that a sufficient, but not necessary condition for the cost function to have a unique minimum (or two equal, successive minima) is that $\pi\lambda + ha - ra > 0$.

Finally, in the case where we have a sample of patient stay times and the model parameters are unknown, as described in Section 2, it is straightforward to calculate the expected cost function given the MCMC output in the usual way. Results for the St. George's Hospital data are presented in the following subsection.

3.1 *Results for the St. George's Hospital Data*

There is no direct information on the arrival process of patients in the St. George's Hospital data. Instead, we assume here that the arrival parameter λ is known. Supposing that $\lambda = 1.5$ which approximately corresponds to the values used by Gorunescu *et al.* (1999), then recalling from Section 2.2 that the predictive mean estimate of the length stay in hospital was around 37.3, then the offered load can be estimated to be $\mathrm{E}[a \mid \boldsymbol{t}] = 37.3 \times 1.5 = 55.95$ which corresponds to the average number of patients who arrive during a mean length of stay in hospital.

Figure 4 illustrates the estimated mean cost function (solid line), the median (thick solid line) and an 80% predictive interval (dashed lines), where we assume the following costs: $h = 1, \pi = 200, r_S = 4, r_M = 2$ and $r_L = 5$ units. Note that the values for h and π are based on those of Gorunescu *et al.* (1999).

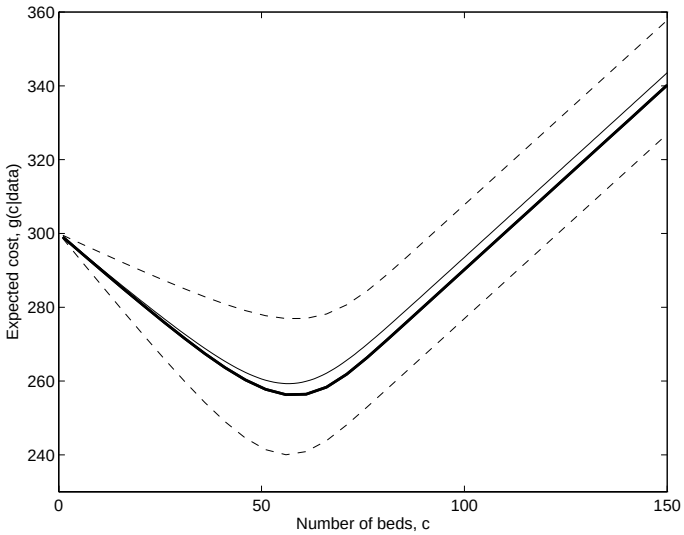


Figure 4. *Mean and median cost functions and 80% predictive interval.*

As can be observed, the optimal number of beds is estimated to be 57 when the minimum cost is 259.31 units. However, the maximum variance is reached near the minimum number of beds; specifically when $c = 58$ and there is some uncertainty in the range 55-60.

It is also important to explore the sensitivity of these results to different interarrival rates of patients. As can be shown, the optimal value of c increases with λ as does the average cost and its variance. Table 1 illustrates the optima for λ between 1 and 2. There is a reasonably large degree of sensitivity to λ.

Table 1. *Optimal numbers of beds for different interarrival rates.*

λ	Optimal c	$\mathrm{E}[g(c)]$	$s.d.[g(c)]$
1.0	38	174.26	15.22
1.5	57	259.31	23.28
2.0	75	344.08	31.37

As we noted earlier in Section 2.2, the predictive distribution of the stay in hospital is insensitive to the prior distribution for the number of phases L. The same result applies to the predicted cost function and so the election of the optimal number of beds is not strongly influenced by this prior. Note however that we could also check the sensitivity to the chosen values of the cost parameters h, π, etc. could also be explored, although this is not done here.

4. DISCUSSION

In this paper, we have developed a Bayesian methodology to make inference about the distribution of the length of stay of patients in hospital and to enable us to optimize the number of beds in the hospital. A number of modifications and extensions are possible.

Firstly, although the MGE distribution seems to fit the data reasonably well, it would also be possible to consider alternatives which might be more flexible, e.g. mixtures of gamma or Erlang distributions, see e.g. Wiper *et al.* (2000), or a general phase–type model, see e.g. Bladt *et al.* (2001). Another possibility is to consider the effects of covariates such as the age of patients and year of admission on the number of days spent in hospital. See also Faddy and McClean (1999).

Secondly, the assumption that all patients who cannot find a bed are lost may be overly restrictive. It is possible to consider alternative systems to model the throughput of patients such as a simple $M/G/c$ queueing model. In this case, we could assume a waiting list of patients and compute the stationary distribution of the system. For queueing systems with general service distributions, this is complex but in the case where the service time distribution is phase–type, this can be carried out using matrix–geometric methods; see e.g. Ausín *et al.* (2002). In this case, the cost function should be modified to take account of waiting times.

Finally, note that this procedure could be applied to other types of data, e.g. stays in hotels and teletraffic data problems.

REFERENCES

Ausín, M.C., Wiper, M.P. and Lillo, R.E. (2002). Bayesian estimation for the $M/G/1$ queue using a phase type approximation. To be published in *J. Statist. Planning and Inference* **.**

Asmussen, S., Nerman, O. and Olsson, M. (1996). Fitting phase type distributions via EM algorithm. *Scandinavian J. Statist.* **23**, 419–441.

Bladt, M., Gonzalez, A. and Lauritzen, S.L. (2001). The estimation of phase–type related functionals through Markov chain Monte Carlo methods. To be published in *Scandinavian Actuarial J.*

Cohen, J.W. (1976). *On Regenerative Processes in Queueing Theory.* Berlin: Springer.

Faddy, M.J. (1990). Compartmental models with phase–type residence time distributions. *Appl. Stochastic Models and Data Anal.* **6**, 121–127.

Faddy, M.J. (1994). Examples of fitting structured phase–type distributions. *Appl. Stochastic Models and Data Anal.* **10**, 247–255.

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

Gorenescu F., McClean S.I. and Millard, P.H. (1999). Using a M/PH/C Queue to Optimize Hospital Bed Occupancy. *Proceedings of the Applied Stochastic Models and Data Analysis Conference, Lisbon*, 106–111.

Harrison, G.W., 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.

Johnson, N.L., Kotz, S. (1970). *Distributions in statistics: continuous univariate distributions.* New York: Wiley.

Neuts, M.F. (1981). *Matrix–geometric solutions in stochastic models.* Baltimore: John Hopkins University Press.

Richardson, S. and Green, P. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Statist. Soc. B* **59**, 731–792.

Ripley, B.D. (1987). *Stochastic Simulation.* New York: Wiley.

Taylor, G.J., McClean, S.I. and Millard, P.H. (2000). Stochastic Models of Geriatric Patient Bed Occupancy Behaviour. *J. Roy. Statist. Soc. A* **163**, 39–48.

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

Wiper, M.P., Rios Insua, D. and Ruggeri, F. (2001). Mixtures of gamma distributions with applications. *J. Comp. Graph. Statist.* **10**, 440–454.

ACKNOWLEDGEMENTS

Some of the work for this paper was carried out while Conchi Ausín and Mike Wiper were visiting CNR–IMATI in November to December 2001. Conchi Ausín, Rosa Lillo and Mike Wiper also acknowledge support from the Spanish Ministry of Science and Technology through grant BEC2000-0167.