A Fully Bayesian Approach to Inference for Coxian Phase-Type Distributions with Covariate Dependent Mean

C.A. McGrory [a,*], A.N. Pettitt [a,b] M.J. Faddy [a]

[a] *Queensland University of Technology*

[b] *Lancaster University*

Abstract

Phase-type distributions represent the time to absorption for a finite state Markov chain in continuous time, generalising the exponential distribution and providing a flexible and useful modelling tool. We present a new reversible jump Markov chain Monte Carlo scheme for performing a fully Bayesian analysis of the popular Coxian subclass of phase-type models; the convenient Coxian representation involves fewer parameters than a more general phase-type model. The key novelty of our approach is that we model covariate dependence in the mean whilst using the Coxian phase-type model as a very general residual distribution. Such incorporation of covariates into the model has not previously been attempted in the Bayesian literature. A further novelty is that we also propose a reversible jump scheme for investigating structural changes to the model brought about by the introduction of Erlang phases. Our approach addresses more questions of inference than previous Bayesian treatments of this model and is automatic in nature. We analyse an example dataset comprising lengths of hospital stays of a sample of patients collected from two Australian hospitals to produce a model for a patient's expected length

of stay which incorporates the effects of several covariates. This leads to interesting conclusions about what contributes to length of hospital stay with implications for hospital planning. We compare our results with an alternative classical analysis of these data.

Key words: Coxian Phase-type model, Phase-type distribution, Reversible jump Markov chain Monte Carlo, Bayesian analysis, Erlang distribution, Covariate Effects

1 Introduction

Phase-type models generalise the exponential distribution and are characterised by an underlying finite Markov chain that has one absorbing state. This underlying Markov process passes through a number of transient states, or phases, until eventually being absorbed. Therefore, the phase-type model is the distribution of the time until absorption for a finite Markov process. This is useful in many application areas: phase-type models have been used to analyse hospital length of stay (LoS) data using maximum likelihood-based approaches ([11], [12], [13], [22] and [32]), they have been successfully used in risk analysis ([1], [2]) and queueing theory ([7]), and they can be fitted using the EM algorithm ([3]). There are several subclasses of phase-type distributions; in this paper we focus on Bayesian inference for the highly versatile and popular Coxian subclass of phase-type models. In the Bayesian litera-

* Corresponding author: School of Mathematical Sciences, Queensland University of Technology, GPO Box 2434, Brisbane, Queensland, 4001, Australia, Tel. (61)7-3138-1287, Fax.: (61)7-3138-1508

Email address: c.mcgrory@qut.edu.au (C.A. McGrory).

ture, phase-type models have been much less well explored. Bayesian Markov chain Monte Carlo techniques for general phase-type models are explored in [8], but this is limited to the fixed dimension case. A reversible jump Markov chain Monte Carlo (RJMCMC) approach which allows the number of transient phases in the model to vary is taken in [4] and [5]. However, in [4] and [5] an alternative mixture representation of the Coxian phase-type model, in terms of a mixed generalised Erlang distribution, is used rather than the matrix exponential formulation used in this paper. Although these two representations are mathematically equivalent, in [4] and [5] a number of latent variables were introduced, which had to be imputed in the RJMCMC scheme. The introduction of latent variables is not necessary here as we use the matrix exponential formulation, which has advantages when seeking reasonable acceptance rates for dimension-changing proposals in RJMCMC algorithms as there are fewer terms involved in the likelihood. It has also been previously noted that there is potential for unreliability when mixture type models are fitted using RJMCMC (see [21], for example).

In this paper we present a novel Bayesian approach in which the Coxian phase-type model is used as a very general residual distribution. The incorporation of a covariate dependent mean into the model has not previously been attempted in the Bayesian literature. In the case of regression models, the use of phase-type distributions allows the error structure of the standard generalised linear model, which is usually a gamma or inverse Gaussian distribution for positive continuous data, to be more flexible to accommodate, for example, long tailedness and a mode near zero simultaneously. This makes these distributions particulary suited to hospital length of stay (LoS) modelling applications, such as the one we consider in this paper, where the data typically

exhibit these features. In this context, phase-type modelling should result in more efficient estimation of the covariate dependence than one would obtain by using a standard exponential family distribution. The phases may or may not have an interpretation in the context of the application, but our focus here is on the estimation of the covariate dependence. The aim is to identify factors leading to increased LoS, which in turn leads to bed occupancy problems, thus having implications for efficient health-care facility and budget planning. This is an active research area and various other techniques have been applied to this problem, examples include the use of classical queuing theory to represent patient flow through various phases of treatment or centers of care (see [14], for example) and the use of a stochastic compartmental modelling approach (see, for instance [30]). Refer to [23] for a useful overview of the directions that research in this area has taken.

In our novel approach we develop an RJMCMC ([17]) analysis of data modelled by a Coxian phase-type distribution. The well-known paper [26] describes how RJMCMC can be used for mixture model analysis and [27] adapts these ideas to the hidden Markov model setting. Our Coxian model differs from the standard Markovian model in that it has additional constraints that must be taken into consideration in the construction of an appropriate RJMCMC algorithm. The difficulties associated with designing an RJMCMC scheme which will adequately explore the posterior are well-known, but we have been able to construct a sampler for this model which traverses the target distribution well. Our modelling of covariate dependency will also be useful in other applications. Another contribution of this paper is to devise an RJMCMC scheme for exploring the inclusion into the phase-type model of an Erlang component, where specific structure leads to a more peaked mode. Using an RJMCMC

scheme we can automatically select the number of transient phases as well as their associated rate parameters, and estimate the covariate dependence (the number of covariates is fixed in our scheme, but this could also be estimated if desired). However, we still have the capability of exploring the important model features mentioned above. We demonstrate our new RJMCMC approach with an application to modelling the effects of several covariates on the length of stay of patients in two Australian hospitals.

In Section 2 we describe the Coxian subclass of phase-type distributions and in Section 3 we describe our Bayesian formulation of the model. In Section 4 we present our RJMCMC methodology. In Section 5 we demonstrate the technique through analysing the hospital LoS data, which leads to conclusions about the effect of several factors on increasing length of stay. Section 6 explores the introduction of Erlang components into the model via RJMCMC and Section 7 concludes the paper.

2 Coxian Phase-Type Distributions

A phase-type distribution describes a Markov process, $\{X(t); t \geq 0\}$, say, where the system moves through some or all of K transient states, or phases, before moving to a single absorbing state $K+1$. See [25] for a full description. The phases are governed by the transition probabilities

$$\begin{aligned} P(X(t+\delta t) = j+1 | X(t) = j) &= \lambda_j \delta t + o(\delta t), \quad j = 1, \ldots, K-1 \\ P(X(t+\delta t) = K+1 | X(t) = j) &= \mu_j \delta t + o(\delta t), \quad j = 1, \ldots, K. \end{aligned}$$

Here δt represents a small time increment. The $\{\lambda_j\}$ are the transition rates between the transient states and the $\{\mu_j\}$ describe the transition from any of

the transient phases to the absorbing state.

In the Coxian phase-type model (see [10]) the system starts in the first phase and then moves through the transient phases sequentially before eventually being absorbed from any one of them. See Figure 1a for an illustration.

The probability density function of the time spent moving through the transient states before absorption is $f(t) = \mathbf{p}\exp\{\mathbf{Q}t\}\mathbf{q}$, where the infinitesimal generator $\mathbf{Q}$ is given by

$$\mathbf{Q} = \begin{pmatrix} -(\lambda_1+\mu_1) & \lambda_1 & 0 & \dots & 0 & 0 \\ 0 & -(\lambda_2+\mu_2) & \lambda_2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -(\lambda_{K-1}+\mu_{K-1}) & \lambda_{K-1} \\ 0 & 0 & 0 & \dots & 0 & -\mu_K \end{pmatrix},$$

and the vectors $\mathbf{p}$ and $\mathbf{q}$ take the forms $\mathbf{p} = (1\ 0\ \dots\ 0)$ and $\mathbf{q} = (\mu_1\ \ \mu_2\ \dots\ \mu_K)^T$. Here $\exp\{\cdot\}$ represents the matrix exponential function and we compute this using Matlab.

The marginal distribution $[\pi_1(t)\dots\pi_K(t)] = \mathbf{p}\exp\{\mathbf{Q}t\}$, describes the probability, $\pi_j(t)$, that the system is in state j, where $j \in [1:K]$, at time point t. The survivor function can be derived from this if it is of interest. The Coxian subclass describes any phase-type distribution with a generator matrix $\mathbf{Q}$ that has real eigenvalues and includes the exponential and Erlang distributions. Reference [19] describes two algorithms for computing a Coxian representa-

tion from a more general phase-type distribution with a generator matrix that has real eigenvalues.

We can introduce covariate dependency into the model so that the mean absorption time is given by the log-linear regression $\exp\{a + \mathbf{b}^T\mathbf{X}\}$, where $\mathbf{X} = (X_1, \cdots, X_c)$ are the covariate values and $\mathbf{b} = (b_1, \cdots, b_c)$ are their coefficients. The expectation of time spent in the system is given by $\mathbf{E}(T) = (-1)\mathbf{p}\mathbf{Q}^{-1}(1\ 1\ \ldots\ 1)^T$ ([25]). Therefore, to incorporate the desired dependency, we scale the transition rate matrix appropriately as $\exp\{-\mathbf{b}^T X\}\mathbf{Q}$, with the intercept term a given by $\exp(a) = (-1)\ \mathbf{p}\ \mathbf{Q}^{-1}(1\ 1\ \ldots\ 1)^T$. In [11] covariates are also incorporated in this way in a classical approach, but this has not been done in previous Bayesian analyses of these distributions.

3 Bayesian Model Formulation for a phase-type Model with an Unknown number of Phases

Given observations comprising absorption times $t_1, ..., t_n$ from a phase-type distribution with K transient states, and putting $\boldsymbol{\theta}_K = (\boldsymbol{\lambda}, \boldsymbol{\mu}, \mathbf{b})$, the likelihood is given by $p(t|\boldsymbol{\theta}_K, K) = \prod_{i=1}^{n} \mathbf{p} \exp\{\mathbf{Q}t_i\}\mathbf{q}$. The transition rates for the transient and the absorbing states are given Gamma prior distributions independent of K: $\lambda_j \sim Ga(\alpha_j, \beta_j)$ and $\mu_j \sim Ga(\gamma_j, \delta_j)$, where $Ga(\cdot, \cdot)$ corresponds to the Gamma probability density function and $\{\alpha_j\}, \{\beta_j\}, \{\gamma_j\}$, and $\{\delta_j\}$ are hyperparameters. The number of phases K and the covariates $\mathbf{b}$ must be assigned prior distributions that are appropriate for the application at hand (we specify these in the context of our application later). Our posterior distribution has the form

$$
\begin{aligned}
p(\boldsymbol{\theta}_K, K|\mathbf{t}) &\propto p(\mathbf{t}|\boldsymbol{\theta}_K, K)p(\boldsymbol{\theta}_K|K)p(K)p(\mathbf{b}) \qquad (1)\\
&= \prod_{i=1}^{n} \mathbf{p} \exp\Big(\exp(-\mathbf{b}^T X)\mathbf{Q}t_i\Big)\mathbf{q} \prod_{j=1}^{K-1} \frac{1}{\Gamma(\alpha_j)} \frac{1}{{\beta_j}^{\alpha_j}} \lambda_j^{\alpha_j - 1} \exp(-\frac{\lambda_j}{\beta_j})\\
&\quad \times \prod_{j=1}^{K} \frac{1}{\Gamma(\gamma_j)} \frac{1}{{\delta_j}^{\gamma_j}} \mu_j^{\gamma_j - 1} \exp(-\frac{\mu_j}{\delta_j}) \times p(K) \times p(\mathbf{b}).
\end{aligned}
$$

4 Reversible Jump Markov Chain Monte Carlo Approach

RJMCMC techniques (see [17]) allow us to fully explore the available parameter space by moving or jumping between models with a varying number of phases. We devised a RJMCMC scheme for this Coxian model, something for which we found no guidance in the literature, which enables transdimensional moves to occur with good acceptance rates. At each iteration of our algorithm we randomly choose, with equal probability, one of the following three move types: perform a fixed dimension parameter update, split a phase in two or combine two existing phases into one, the birth of a new phase or the death of an existing phase.

4.1 Metropolis-Hastings Fixed Dimension Parameter Update Move

We follow standard methods in the literature for updating the rate parameters and the regression parameters via Metropolis-Hastings.

4.2 Dimension Changing Reversible Jump Moves

We denote the current number of phases in the model by K, and the proposed number by K^*, where K^* is restricted to be equal to $K-1$ or $K+1$. We

assume that the maximum potential number of phases is fixed at $K = K_{max}$, say.We propose a change of the model parameters from $\boldsymbol{\theta}_K$ to $\boldsymbol{\theta}_{K^*}$ through a bijective mapping from the parameter space $(\boldsymbol{\theta}_K, u, v)$ to $(\boldsymbol{\theta}_{K^*}, u^*, v^*)$, where u, v, u^* and v^* are auxiliary variables introduced so that dimensionality is the same in the current and proposed parameter spaces. These moves are accepted with probability $min(R, 1)$ where R is given by

$$\frac{p(t|\boldsymbol{\theta}_{K^*}, K^*)p(\boldsymbol{\theta}_{K^*})p(K^*)}{p(t|\boldsymbol{\theta}_K, K)p(\boldsymbol{\theta}_K)p(K)} \times \frac{J_{K^*,K}\ p(u^*, v^*|K^*, K, \boldsymbol{\theta}_{K^*})}{J_{K,K^*}\ p(u, v|K, K^*, \boldsymbol{\theta}_K)} \times \left|\frac{\partial(\boldsymbol{\theta}_{K^*}, u^*, v^*)}{\partial(\boldsymbol{\theta}_K, u, v)}\right|. \quad (2)$$

The first term in (2) is the ratio of the likelihood times prior (see (1)) for the proposed and current parameter values, and the second is the ratio of proposal probabilities. We denote the probability of moving from K to K^* phases by J_{K,K^*}. The third term is the Jacobian for the transformation.

We denote by μ and λ the rate parameters associated with the phase of interest in the model of lower dimension, and by μ_a,μ_b, λ_a and λ_b the rates associated with the two phases of interest in the model of higher dimension. It can be challenging to define a suitable mapping, particularly in the case of practically driven applications, as it is often difficult to obtain good mixing across model dimensions. To obtain reasonable acceptance rates for proposed transdimensional jumps, one requires an appropriately centered proposal distribution with well tuned parameters. However, it is not generally obvious how best to achieve this ([9] makes some suggestions in this regard). Here, we take the approach of constructing our proposal distributions so that the proposed parameters are not too distant from the current parameters, and we take a matching approach to the construction of our mapping between dimension spaces. We ensure that probability of absorption and the mean time

in the phase(s) are matched in the current and proposed dimension spaces corresponding to equations (3) and (4), given below.

$$\frac{\mu}{\mu+\lambda} = \frac{\mu_a}{\mu_a+\lambda_a} + \left(\frac{\lambda_a}{\mu_a+\lambda_a} \times \frac{\mu_b}{\mu_b+\lambda_b}\right) \tag{3}$$

$$\frac{1}{\mu+\lambda} = \frac{1}{\mu_a+\lambda_a} + \frac{1}{\mu_b+\lambda_b}. \tag{4}$$

Split and Combine Moves

The design of our split and combine moves does not allow splits or combines of the final phase. In all other cases we assume equal probabilities of splitting or combining. Figure 1b illustrates these move types graphically. In the combine move we have $(\mu_a, \lambda_a, \mu_b, \lambda_b) \rightarrow (u, v, \mu, \lambda)$. We put $u = \mu_a$ and $v = \lambda_a$, then solving (3) and (4) gives us

$$\mu = \frac{\mu_a\mu_b + \mu_a\lambda_b + \lambda_a\mu_b}{\mu_a + \lambda_a + \mu_b + \lambda_b}$$

$$\lambda = \frac{\lambda_a\lambda_b}{\mu_a + \lambda_a + \mu_b + \lambda_b}.$$

In this case, the Jacobian is given by

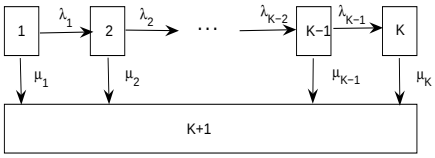
$$\frac{(\mu_a+\lambda_a)^2\lambda_a}{(\mu_a+\lambda_a+\mu_b+\lambda_b)^3}.$$

Our corresponding split move involves the reverse transition (see Figure 1b). In the higher dimension space, we set μ_a and λ_a to be equal to the simulated auxiliary variables u and v, respectively, where $u \sim N_T(2\mu, \sigma^2)$ and $v \sim N_T(2\lambda, \sigma^2)$. Here $N_T(\cdot,\cdot)$ denotes the Normal density function truncated at zero with mean μ and suitable tuned variance σ^2. We simulate from the truncated distribution since we cannot have negative values for the rate parameters. By solving (3) and (4) we obtain

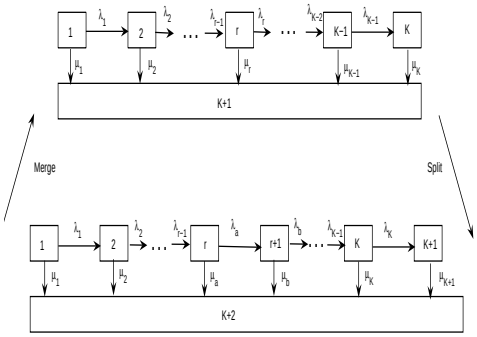
$$\mu_b = \frac{{\mu_a}^2\lambda + \mu_a\lambda_a\lambda - \lambda_a\mu\mu_a - \lambda_a^2\mu}{\lambda_a(-\mu_a - \lambda_a + \mu + \lambda)}$$
$$\lambda_b = -\frac{(\mu_a + \lambda_a)^2\lambda}{\lambda_a(-\mu_a - \lambda_a + \mu + \lambda)}.$$

If either of μ_b or λ_b is negative when calculated the proposal is rejected. The Jacobian for the split move is the reciprocal of the corresponding expression for the combine move. The acceptance ratio for each of the above moves is then given by substituting the appropriate values into (2). These acceptance ratios are reciprocals of one another.

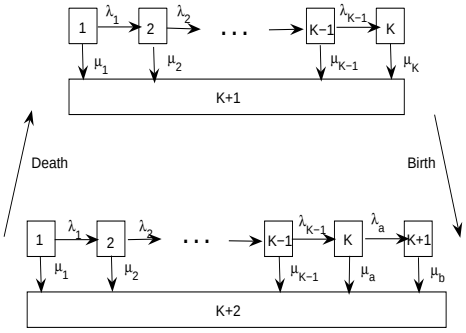
Fig. 1. Diagrammatic representations of (a) the Coxian phase-type model, (b) the effects of our RJMCMC split and merge moves and (c) the effects of our RJMCMC birth and death moves.



(a) Coxian phase-type distribution



(b) Split and Merge Moves



(c) Birth and Death Moves

Birth and Death Moves

These moves are only applied to the final phase in the current model in a given iteration and bring about the birth of a new final phase or the death of the existing final phase with equal probability (provided that $1 < K < K_{max}$). The death move makes the transition $(\mu_a, \lambda_a, \mu_b) \rightarrow (u, v, \mu)$; see Figure 1c. Putting $u = \mu_a$ and $v = \lambda_a$ and solving (3) and (4), we obtain

$$\mu = \frac{(\mu_a + \lambda_a)\mu_b}{(\mu_b + \mu_a + \lambda_a)}.$$

The Jacobian for the death move is given by

$$\frac{(\mu_a + \lambda_a)^2}{(\mu_b + \mu_a + \lambda_a)^2}.$$

In the reverse birth move, we generate u and v from the Normal distribution truncated at 0 with mean μ and variance σ^2 and set $\mu_a = u \sim N_T(\mu, \sigma^2)$ and $\lambda_a = v \sim N_T(\mu, \sigma^2)$. Again, σ^2 is chosen to give reasonable rates of acceptance for the move. To satisfy (3) and (4), we take μ_b to be

$$\mu_b = \frac{(u + v)\mu}{u + v - \mu}.$$

If this results in a negative μ_b, we reject the proposal. The Jacobian for the death move is the reciprocal of that for the corresponding reverse birth move described above. We can obtain the acceptance ratio for the birth and the death moves by substituting the appropriate quantities into (2) and these are of course reciprocals of each other.

5 Application: Modelling Length of Stay in Hospital

The identification of factors that are likely to increase a patient's LoS is a key goal for hospital planners. By addressing issues that lead to a longer LoS, health care costs can be reduced. LoS data are characteristically highly right-skewed making it difficult to fit them with other distributions. See [11] for a discussion of some of the difficulties associated with modelling LoS data. Phase-type distributions provide the flexibility that is required to capture the distributional characteristics of this type of data. The phases may only be artifacts of the modelling, but could have a physical interpretation in relation to the context. However, our focus here is on the estimation and interpretation of the covariate effects.

We applied our method to a dataset previously analysed using classical maximum likelihood techniques ([11]), with our results complementing this analysis. The dataset comprises the lengths of hospital stay of 1901 patients all of whom were at least 18 years of age. These data were collected from two hospitals in S.E. Queensland, Australia, between October 2002 and January 2003. Patients were recruited from a range of specialities, but only those whose admissions were considered uncomplicated contributed to the data.The observed lengths of stay ranged from 0.44 to 170.9 days. The sample mean length of stay was 7.25 days. Information on ten covariates widely believed to be of relevance to length of hospital stay was also available for each patient and was included in our model. Details of the covariate information are given in the Appendix. (Note that this dataset is a part of a larger dataset collected in a prospective study and that [15] and [16] provide further details of the method of collection.)

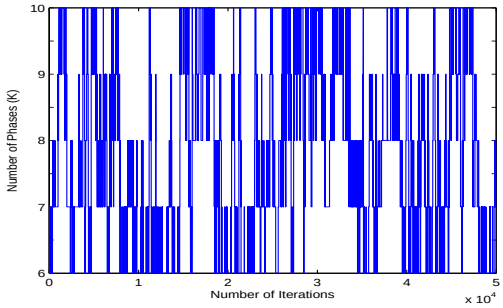
For each patient we also have a predicted length of stay based on the patient's admission category for an uncomplicated admission. This was obtained from the Australian Institute of Health and Welfare. The logarithm of the predicted length of stay, x_0, was incorporated into our model as an offset variable. In this way we are modelling a patient's excess length of stay relative to the prediction and $\exp\{a+\mathbf{b}^T X\}$ as $\mathbf{E}(T/x_0)$ where T is the actual length of stay.

We assigned the covariate coefficients $\mathbf{b}$ uniform priors over the range -10 to 10. We assumed that the maximum potential number of phases in the model was fixed at $K_{max} = 10$ and we chose a uniform prior distribution over 1 to 10 for K. The hyperparameters chosen for the Gamma priors over the rate parameters were also chosen to be uninformative.

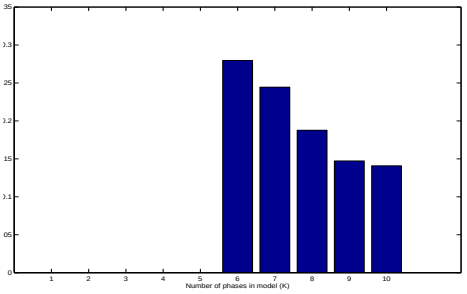
We performed 100 000 iterations of our RJMCMC algorithm and we discarded the first 50 000 of these iterations to allow for a burn-in period. The algorithm was tuned so that acceptance rates for fixed dimension updates of the parameters $\boldsymbol{\mu}$, $\boldsymbol{\lambda}$ and $\mathbf{b}$ were between 30% and 35%. The overall rate of acceptance for the dimension changing moves was around 7% which although low, is reasonably good for RJMCMC. The trace plot for the number of phases, K, over all iterations post burn-in, is given in Figure 2a. The most likely number of phases was six, having posterior probability of 0.27, followed by the seven-phase model which had posterior probability of 0.25 (see Figure 2b for the posterior distribution for K). To further examine convergence, we ran our algorithm from two different starting points and thinned the observations to 1 in 250. We then plotted the posterior probability that the number of phases was six at each iteration point. This plot is shown in Figure 2c. This, together with the trace plot, suggests that the scheme has converged. The posterior estimate of the number of phases as six is in agreement with the classical analysis

[11]. Figure 3 displays the posterior distributions of the parameter estimates from the six-phase fit. Since our main aim is to model the effects of covariates on the mean LoS, the actual parameterisation of the Coxian distribution becomes irrelevant, as long as the residual variation is adequately described. In [5] the authors noted that it was necessary to impose an identifiability constraint in their RJMCMC analysis of the mixture representation of Coxian model in order to obtain identifiability of the rate parameter estimates; such constraints could possibly be considered in our scenario if the rate parameters were of particular interest in the application. However, it is worth noting that in our results the posterior distributions for the rate parameters appear to be unimodal suggesting that identifiability was not a significant problem in our implementation.

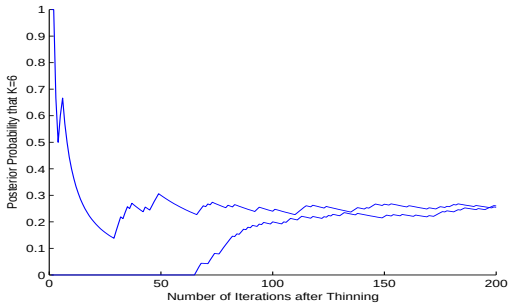
Fig. 2. These plots show the results of 50,000 iterations (after burn-in) of our RJMCMC sampler for the hospital length of stay data. (a) Trace plot of the number of phases (K) in the model at each iteration. (b) The posterior distribution of the number of phases (K). (c) Plot of the estimated posterior probability that the number of phases in the model is six at each iteration of two different runs of the RJMCMC algorithm.



(a)

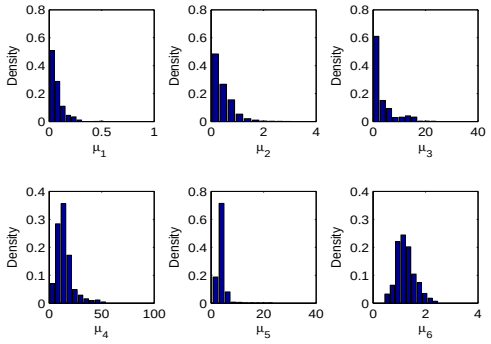


(b)

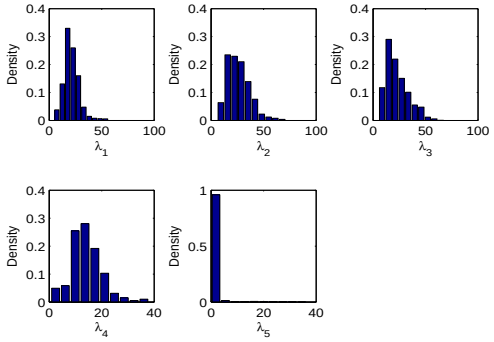


(c)

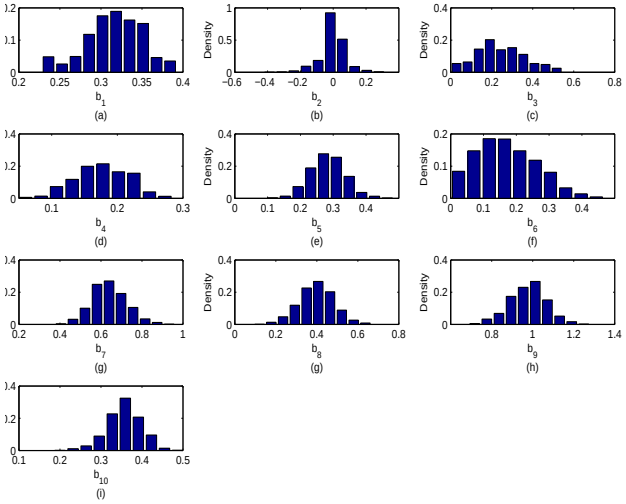
Fig. 3. Posterior distributions of (a) the μ's, (b) the λ's and (c) the b's (after burn-in) from the six phase model fitted in the RJMCMC analysis of the hospital length of stay data.



(a)



(b)



(c)

The estimates of the covariate coefficients $\mathbf{b}$ are of primary interest in the application and the posterior distributions for these are reasonably symmetrical. These can be used to estimate the effect that each of the covariates has on increasing the length of the patient's stay beyond the initial prediction made upon admission. Since the posterior distributions for the parameters exhibit some skewness we used the posterior medians as parameter estimates. The posterior medians (posterior standard deviations in brackets) of the intercept parameter and covariate coefficients are given by

$$\begin{aligned} a &= -1.10(0.04) \\ \mathbf{b} &= [0.32(0.03)\ \ 0.01(0.07)\ \ 0.22(0.12)\ \ 0.18(0.04)\ \ 0.28(0.05) \\ &\qquad\qquad 0.16(0.09)\ \ 0.64(0.08)\ \ 0.39(0.08)\ \ 0.98(0.08)\ \ 0.35(0.04)]. \end{aligned}$$

Our posterior estimates for the covariate coefficients showed some similarity with the maximum likelihood estimates in [11]. Based on our results, we can see that contraction of a health care acquired infection (covariate 9) would be expected to bring about the greatest increase in length of stay, while faecal incontinence (covariate 7) was estimated to be the second most influential factor and sex the least influential.

Inference about the effect of health care acquired infection is useful to hospital planners, as health care acquired infections (HAIs) are widely believed to place a substantial economic burden upon the health system. Moreover, a recent study ([18]) has suggested that HAIs could be prevented in some cases. However, [16] has highlighted that despite this consideration there have been few published studies on the actual impact that the implementation of infection control programs might have in reducing the costs associated with HAIs. To estimate what the economic benefits might be, we must first estimate the effect of HAIs in real terms. Our estimated coefficient for the HAI covariate

was 0.98, with a 95% credible interval of 0.79 to 1.15 for that estimate. Based on our sample of patients, we would estimate that the contraction of an HAI would lead to an increased stay of 13.25 days on average, with 95% credible interval for this estimate of 7.89 days to 15.34 days.

Our estimate of the effect of the pressure ulcer covariate (covariate 6) is also worthy of comment since, as [15] points out, many previous studies have suggested that the development of pressure ulcers in hospital has a fairly significant effect on lengthening stay. This effect has been estimated as ranging from a 7 to a 50 day increase in stay for affected patients (references cited in [15]). However, the authors of [15] suggest that this effect has been overestimated, as they found that the occurrence of pressure ulcers would lead to an estimated median increase in stay of only 4.31 days (with a 95% confidence interval of 1.85 to 6.78 for this estimate.) Our results estimate the coefficient for the pressure ulcer covariate to be 0.16, with a 95% credible interval given by 0.02 to 0.36. This corresponds to an expected increase in LoS of 1.83 days on average, with a 95% credible interval for this estimate of 0.20 to 2.61 days. Therefore, our results also support the view that pressure ulcers may not have as much of a role in increasing LoS as has previously been suggested.

The similarities between our conclusions and those from more classical studies lends support to the ability of our RJMCMC-based sampling scheme to obtain useful model estimates in practical applications. With our method the inference is performed directly, in contrast to the two-tier classical approach ([11]) of model identification and subsequent maximum likelihood parameter estimation. It is also worth noting that we reached this solution from starting values that were easily obtained from a simple generalised linear model fit, rather than multiple iterative searches with different starting values to

determine the maximum likelihood solution.

6 Reversible Jump Scheme for Initial Erlang Phases

In other analyses of data similar to those here, it has been found that an adequate model for the data corresponded to having several of the initial values of μ equal to zero with the associated phases having equal values of λ. This introduces an initial Erlang component leading to a simpler model involving fewer parameters. To explore the effect this might have on our analysis, we conceived a move type which we call the birth of an Erlang phase, the reverse move being the death of an Erlang phase. The essence of this change is to set the current rate parameter μ_1 to be equal to zero. If μ_1 is already zero, then the move is carried out on μ_2 and so on. In this way we have developed an RJMCMC scheme that searches over competing distributional structures. We describe these moves in specific terms in the following sections. As before, the acceptance ratio for these moves is obtained by the substitution of the relevant values into (2).

6.1 Birth and Death of the First Erlang Phase

If μ_1 is currently nonzero, we choose our transformed parameters to satisfy equation (5) corresponding to matching the mean length of time in the first phase before and after it becomes an Erlang phase.

$$\frac{1}{\mu_1 + \lambda_1} = \frac{1}{\lambda_a}. \tag{5}$$

The birth of the Erlang component involves the transition $(\mu_1, \lambda_1) \rightarrow (u, \lambda_a)$.

Figure 4a provides an illustration of this move type. Choosing λ_a to satisfy (5) gives

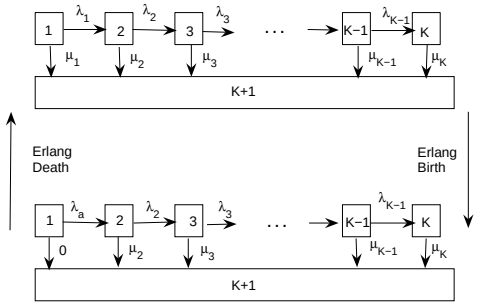
$$\begin{aligned} \lambda_a &= \mu_1 + \lambda_1 \\ u &= \frac{\lambda_1}{\mu_1 + \lambda_1}. \end{aligned}$$

The Jacobian for this move is given by

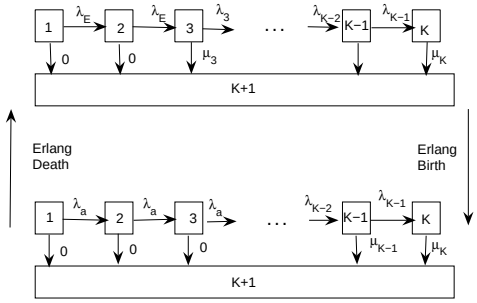
$$\frac{1}{\mu_1 + \lambda_1}.$$

The death of the Erlang phase (see Figure 4a), involves the opposite transition $(u, \lambda_a) \rightarrow (\mu_1, \lambda_1)$. We generate our auxiliary variable u from a uniform proposal distribution $u \sim Un(0, 1)$, where $Un(\cdot, \cdot)$ represents the uniform distribution. Then we put $\mu_1 = u\lambda_a$ and $\lambda_1 = (1 - u)\lambda_a$. This choice for μ_1 and λ_1 satisfies (5). The Jacobian is equal to λ_a (the inverse of the Jacobian for the reverse move).

Fig. 4. Diagrammatic representations of (a) the Erlang birth and death moves when jumping between a general phase model and a one Erlang phase model, and (b) the general Erlang birth and death moves that are performed when there is at least one Erlang phase present in the model.



(a) Birth/death of initial Erlang phase



(b) Example of the birth/death of a general Erlang phase

6.2 Birth and Death of Erlang Phases in General

When one or more of the initial μ's have already been set to zero, birth of another Erlang component must take into account the equal eigenvalue constraint in the Erlang part of the model. This change will involve two components: the rate parameters for the r^{th} phase (the one we are considering for incorporation into the Erlang distributed part of the model) and the rate parameter for the existing Erlang phase or phases. We denote the latter by λ_{E}; refer to Figure 4b for an illustration with $r = 3$. We construct our general Erlang birth/death moves so that the mean time in the phases is matched before and after the transformation corresponding to equation (6) below.

$$\frac{r-1}{\lambda_{\mathrm{E}}} + \frac{1}{\mu_r + \lambda_r} = \frac{r}{\lambda_a} \tag{6}$$

The general birth of an Erlang move increases the number of Erlang phases from $r-1$ to r and involves the transition $(\lambda_{\mathrm{E}}, \mu_r, \lambda_r) \rightarrow (\lambda_a, u, v)$. We put $u = \mu_r$ and $v = \lambda_r$. Then from (6) we obtain

$$\lambda_a = \frac{r\lambda_{\mathrm{E}}(\mu_r + \lambda_r)}{(r-1)(\mu_r + \lambda_r) + \lambda_{\mathrm{E}}}.$$

The Jacobian for this move is given by the following expression

$$\frac{r(r-1)(\mu_r + \lambda_r)^2}{((r-1)(\mu_r + \lambda_r) + \lambda_{\mathrm{E}})^2}.$$

The death move involves the reverse transition $(\lambda_a, u, v) \rightarrow (\lambda_{\mathrm{E}}, \mu_r, \lambda_r)$. Here we put $\lambda_r = v$ and $\mu_r = u$, where $u \sim N_T(0, \sigma^2)$ and $v \sim N_T(\lambda_a, \sigma^2)$. We tune σ^2 to give satisfactory acceptance rates for the move and solve (6) to obtain

$$\lambda_{\mathrm{E}} = \frac{(r-1)\lambda_a(\mu_r + \lambda_r)}{r(\mu_r + \lambda_r) - \lambda_a}.$$

The Jacobian for this move is the inverse of the reverse general birth move.

We continue to use uninformative Gamma priors for the parameters μ and λ in this scheme. However, when some phases in the model currently correspond to an Erlang distribution, the shape and scale parameters of the corresponding Gamma prior are multiplied by the current number of Erlang phases in the model to give the prior distribution for the Erlang rate parameter. This prior was also used in [20].

6.3 Results from Applying Erlang Birth and Death Moves to the Hospital Length of Stay Data

We ran our Erlang birth/death algorithm using the six phase posterior estimates from our initial RJMCMC analysis as a starting point. We performed 10, 000 iterations and discarded the first half of these. We found that the most likely number of Erlang phases was two, having posterior probability of 0.94. The resulting posterior medians (posterior standard deviations given in brackets) of the intercept parameter and covariate coefficients were as follows.

$$\begin{aligned} a &= -1.44(0.04) \\ \mathbf{b} &= [0.37(0.03)\ \ 0.01(0.01)\ \ 0.37(0.12)\ \ 0.17(0.04)\ \ 0.28(0.05) \\ &\qquad\qquad 0.15(0.099)\ \ 0.63(0.08)\ \ 0.40(0.08)\ \ 0.96(0.08)\ \ 0.37(0.04)]. \end{aligned}$$

The regression coefficient posterior medians and standard deviations are very similar for the two models except that the coefficient for x_3 has changed from 0.22 to 0.37 and is more statistically significant. This model is simpler and we

have only used nine parameters to describe the phase-type model and ten to describe the regression part of the model for a dataset of nearly 2000 observations. We have not reported the posterior distributions of the rate parameters as they may be subject to some lack of identifiably, but we note that they have unimodal distributions possibly indicating satisfactory identifiability.

7 Conclusions

Our extension of the reversible jump method to Coxian phase-type modelling with covariate dependent mean provides a fully formal Bayesian method for fitting such distributions to data and extends previous Bayesian analyses of this type of model. Our application to hospital LoS data has demonstrated that our approach can be used to provide valuable statistical inference for real world problems. In particular, posterior distributions for the number of phases and the regression parameters are produced, and we have also indicated that suitable starting values for the RJMCMC algorithm can be easily obtained. These advantages make this Bayesian approach attractive in practice.

We have also devised an RJMCMC method for automatically exploring the structure of the phase-type model to investigate the inclusion of an initial Erlang component which, in our case study, gave an improved and simpler structure for the model. Such modelling can be extended.

The phase-type distributions can be interpreted as providing a flexible and partially parametric extension to standard exponential family models, in particular the gamma density family, while still maintaining a quadratic mean/variance relationship. Hence in the regression context such models should provide for

more robust estimation of regression coefficients. An alternative flexible approach might be provided by fitting a normal mixture to the logarithm of the times, but this needs to be investigated. However, such an approach would not provide the structure of the phase-type model where such a structure may have a useful interpretation (e.g. hospital LoS) and it is doubtful whether it could simultaneously capture the mode near zero and the longtailedness of the data.

In our modelling we have not included the case where the covariates are also selected using the RJMCMC scheme. This would be straightforward to implement, but it would be best to exercise caution in applications as there could be confounding between the selection of the number of phases and the covariates. This requires investigation. Other extensions of this theory include the exploration of cases where we have repeated measures observed on each subject; this could be achieved through the use of a frailty term. If we wished to identify the rate parameters then the approach of [5] could straightforwardly be incorporated into our analysis.

Phase-type models are useful in any application where the data exhibit long tails, and there are many research fields in which this type of data arises in addition to the applications we have already mentioned. For example, phase-type models have been used in web site performance optimisation ([31]), wireless communication system control ([29]), line transect sampling ([28]), gene finding ([24]) and ion channel modelling ([6]).

Acknowledgements

The authors' work was supported by Australian Research Council Discovery and Linkage Grants. Some of the computational resources and services used in this work were provided by the High Performance Computing and Research Support Group, Queensland University of Technology, Australia. We wish to thank the anonymous referees for some helpful comments.

APPENDIX: Covariate Information Used in Modelling the Hospital Length of Stay Data

Covariate	Description	Range
x_0	predicted length of stay in days	1-72
x_1	log of age	2.9-4.61
x_2	sex (male/female)	binary 0/1
x_3	discharge destination (death/survive)	binary 0/1
x_4	admission type (emergency/non-emergency)	binary 0/1
x_5	anti-coagulant therapy during admission	binary 0/1
x_6	pressure ulcer during admission	binary 0/1
x_7	faecal incontinence during admission	binary 0/1
x_8	gastro-intestinal bleeding during admission	binary 0/1
x_9	health care acquired infection	binary 0/1
x_{10}	surgical procedure	binary 0/1

References

[1] Asmussen, S., 2000. Ruin probabilities, Advanced Series on Statistical Science and Applied Probability (Vol. 2), Singapore: World Scientific.

[2] Asmussen, S., and Bladt, M., 1996. Phase-type distributions and risk processes with state-dependent premiums, Scandinavian Actuarial Journal, 96, 19–36.

[3] Asmussen, S., Nerman, O., and Olsson, M., 1996. Fitting phase-type distributions via the EM-algorithm, Scandinavian Journal of Statistics, 23, 419–441.

[4] Ausìn, M.C., Lillo, R.E., Ruggeri, F., and Wiper, M.P., 2003. Bayesian modelling of hospital bed occupancy times using a mixed generalised Erlang distribution, in Bayesian Statistics 7, eds. J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. David, D. Heckerman, A.F.M. Smith and M. West, Oxford: Oxford University Press, pp. 443–451.

[5] Ausìn, M.C., Wiper, M.P., and Lillo, R.E., 2008. Bayesian prediction of the transient behaviour and busy period in short-and long-tailed $GI/G/1$ queueing systems, Computational Statistics and Data Analysis, 52, 1615–1635.

[6] Ball, F.G., Milne, R.K., and Yeo, G.F. 2000. Stochastic models for systems of interacting ion channels, IMA Journal of Medicine and Biology, 17, 263–293.

[7] Bertsimas, D., 1990. An analytic approach to a general class of G/G/c queueing systems, Operations Research, 38, 139–155.

[8] Bladt, M., Gonzalez, A., and Lauritzen, S.L., 2003. The estimation of phase-type related functionals using Markov chain Monte Carlo methods, Scandinavian Actuarial Journal, 4, 280–300.

[9] Brooks, S.P., Guidici, P., and Roberts, G.O., 2003. Efficient construction

of reversible jump Markov chain Monte Carlo proposal distributions (with discussion), Journal of the Royal Statistical Society, Series B, 57, 473–484.

[10] Cox, D.R., and Miller, H.D., 1965. An Introduction to the Theory of Stochastic Processes. London: Methuen & Co..

[11] Faddy, M.J., Graves, N., and Pettitt, A.N., 2009. Modeling length of stay in hospital and other right skewed data: comparison of phase-type, gamma and log-normal distributions, Value in Health, 12, 309–314.

[12] Faddy, M.J., and 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.

[13] Faddy, M.J., and McClean, S.I., 2005. Markov chain modelling for geriatric patient care, Methods of Information in Medicine, 44, 369–373.

[14] Gorunescu, F., McClean, S.I., and Millard, P.H., 2002. A queuing model for bed-occupancy management and planning of hospitals, Journal of the Operational Research Society, 53, 19–24.

[15] Graves, N., Birrell, F. and Whitby, M., 2005. The effect of pressure ulcers on length of hospital stay. Infection Control and Hospital Epidemiology, 26, 293–297

[16] Graves, N., Weinhold, D., Tong, E., Birrell, F., Doidge, S., Ramritu, P, Halton, K. Lairson, D. and Whitby, M., 2007. Effect of healthcare-acquired infection on length of hospital stay and cost, Infection Control and Hospital Epidemiology, 28, 280–292.

[17] Green, P.J., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, Biometrika, 82, 711–732.

[18] Harbarth, S., Sax, H., and Gastmeier,P., 2003. The preventable proportion of

nosocomial infections: an overview of published reports, Journal of Hospital Infection, 54, 258–266.

[19] He, Q., and Zhang, H., 2006. Spectral polynomial algorithms for computing bi-diagonal representation for phase type distributions and matrix-exponential distributions, Stochastic Models, 22, no. 2, 289–317.

[20] Insua, D.R., Wiper, M. and Ruggeri, F., 1998. Bayesian analysis of $M/Er/1$ and $M/H_K/1$ queues, Queueing Systems, 30, 289–308.

[21] Jasra, A., Stephens, D.A. and Holmes, C.C., 2007. Population-based reversible jump Markov chain Monte Carlo, *Biometrika*, 94, 787–807.

[22] Marshall, A.H., and McClean, S.I., 2003. Conditional phase-type distributions for modelling patient length of stay in hospital, International Transactions in Operational Research, 10, 565–576.

[23] Marshall, A., Vasilakis, C., and El-Darzi, E., 2005. Length of stay-based patient flow models: recent developments and future directions, Health Care Management Science, 8, 213–220.

[24] Munch, K. and Krogh, A., 2006. Automatic generation of gene finders for eukaryotic species, BMC Bioinformatics, 7, 263–275.

[25] Neuts, M.F. (1981. Matrix Geometric Solutions in Stochastic Models, Baltimore: Johns Hopkins University Press.

[26] Richardson, S., and Green, P.J., 1997. On Bayesian analysis of mixtures with an unknown number of components (with discussion) Journal of the Royal Statistical Society Series B, 59, 731–792.

[27] Robert, C.P., Rydén, T., and 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, 62, 57–75.

[28] Skaug, H.J., 2006. Markov Modulated Poisson Processes for Clustered Line Transect Data, Environmental and Ecological Statistics, 13, 199–211.

[29] Tan, H., Nunez-Queija, R., Gabor, A.F., Boxma, O.J., 2009. Admission control for differentiated services in future generation CDMA networks, Performance Evaluation, In Press, available online.

[30] Taylor, G.J., McClean, S.I., and Millard, P.H., 2000. Stochastic models of geriatric patient bed-occupancy behaviour, Journal of the Royal Statistical Society, Series A, 163, 39–48.

[31] Van der Weij, W. Bhulai, S., Van der Mei, R. 2009. Dynamic thread assignment in web server performance optimization, Performance Evaluation, 66, 301–310.

[32] Xie, H., Chaussalet, T.J., and Millard, P.H. , 2005. A continuous time Markov model for the length of stay of elderly people in institutional long-term care, Journal of the Royal Statistical Society, Series A, 168, 51–61.