# On The Analysis of Hospital Infection Data Using Markov Models

[1]**Ross, J.V.** **and** [2]**Taimre, T.**

[1]King's College, Cambridge CB2 1ST, United Kingdom. E-Mail: j.v.ross@warwick.ac.uk
[2]Department of Mathematics, The University of Queensland, Queensland 4072, Australia. E-Mail: ttaimre@maths.uq.edu.au

***Keywords:*** *Count data, Markov processes, parameter estimation, hospital epidemiology, nosocomial infections.*

**EXTENDED ABSTRACT**

We present a general approach to estimating parameters of continuous-time Markov chains from discretely sampled data. This methodology is combined with a new stochastic model for transmission of hospital-acquired infections — one which accounts for dynamic bed occupancy — providing a method for estimating the parameters of such systems. We pay particular attention to the conditions under which modelling dynamic bed occupancy is necessary. We additionally provide a new method for incorporating partial observability, and compare the results from this method to the commonly used existing approach. These results are anticipated to have wide application in studying nosocomial infections, and for assessing the efficacy of possible management strategies designed to decrease the prevalence of such infections.

Hospital-acquired infections caused by transmissible nosocomial pathogens have been widely studied. This is due to the detrimental effects of such pathogens on patient health resulting in high costs in terms of loss of life and demands on health-care resources. Reports in the United Kingdom state that $1$ in $10$ patients admitted to hospital will acquire a nosocomial infection, resulting in approximately $5000$ deaths and costing the National Health Service one billion pounds per annum (Inweregbu *et al.*, 2005).

Reports of nosocomial infections are continuing to rise, placing increased importance upon their study (see Nimmo *et al.* (2003) for evidence in Australia). This has focussed attention on developing strategies for limiting the prevalence of such infections, for example possibly through improved hygiene practices amongst health-care workers (e.g. handwashing), selective antibiotic use, and isolation (human-human distancing) strategies.

To understand the dynamics of nosocomial infection transmission, and to evaluate the efficacy of these possible control strategies, we must estimate parameters from data consisting of counts of symptomatic individuals over time. Cooper *et al.* (2003) undertook a systematic review of studies using interrupted time series (ITS) data for evaluating intervention strategies for controlling hospital infections. Out of $24$ ITS studies presenting any form of statistical analysis, only one incorporated dependencies between observations.

This led Cooper and Lipsitch (2004) to present a method for parameter estimation using structured hidden Markov models, representing a significant advance on the methods commonly used at that time. The underlying model of this method is a continuous-time Markov chain, a type of model used extensively in theoretical studies but not appearing nearly as widely in applied studies, perhaps due in part to a lack of easily-understood estimation procedures for calibrating these models to real-world systems. Cooper and Lipsitch (2004) provide a method of estimation for such models, which additionally goes some way to incorporating partial observability, an aspect inherent in modelling nosocomial pathogens.

Our purpose here is two-fold. First we provide a new approach to incorporating partial observability. This is to overcome a limitation in the existing hidden Markov model approaches, where symptomatic individuals do not remain identified between observation points. This arises from the major drawback of using hidden Markov models, in that the current output (observation) is assumed to be statistically independent of the previous output. A comparison of our approach to the existing hidden Markov model approach is undertaken demonstrating the effectiveness of our approach.

Secondly, we provide a model and accompanying methodology for addressing an important problem often encountered when analysing hospital infection data, namely dynamic bed occupancy. We investigate when it is necessary to incorporate dynamic bed occupancy in estimation procedures, and provide clear methodology for how this may be effected in practice.

This paper provides results and establishes improved methodology useful for studying hospital-acquired infections and for assessing the efficacy of possible management strategies; we hope that these results are used to determine optimal management strategies for mitigating the prevalence of nosocomial infections.

## 1 INTRODUCTION

Hospital-acquired infections caused by transmissible nosocomial pathogens have been widely studied. This is due to the detrimental effects of such pathogens on patient health resulting in high costs in terms of loss of life and demands on health-care resources. Reports of such infections continue to rise, placing increased importance upon their study (see e.g. Nimmo *et al.*, 2003).

To understand the dynamics of nosocomial infection transmission, and to evaluate possible control strategies, we must estimate parameters from data consisting of counts of symptomatic individuals over time. Cooper *et al.* (2003) undertook a systematic review of studies using interrupted time series (ITS) data for evaluating intervention strategies for controlling hospital infections. Out of 24 ITS studies presenting any form of statistical analysis only one incorporated dependencies between observations.

This led Cooper and Lipsitch (2004) to present a method for parameter estimation using structured hidden Markov models. The underlying model of this method is a continuous-time Markov chain. Continuous-time Markov chains have been proposed as theoretical models for an array of biological systems. However, their usage in applied modelling is not as extensive, most likely due to a lack of clear statistical procedures for fitting the models to data.

While the method presented by Cooper and Lipsitch (2004) incorporates partial observability, an aspect inherent in modelling nosocomial pathogens, they also cite a number of limitations in their approach to parameter estimation. One of these is: "Difficulties may also occur if there are large fluctuations in the total population size; while the observation model could readily cope with such fluctuations by varying denominators, changes in the dimension of the underlying Markov chain would be harder to accommodate." (page 234).

We overcome this limitation by developing a new stochastic model for the underlying Markov chain. The new model specifically incorporates dynamic bed occupancy. We compare estimates derived from incorporating dynamic bed occupancy, to those derived from assuming full ward occupancy, to deduce conditions under which the modelling of dynamic bed occupancy is necessary.

Finally, we consider an alternative approach to incorporating partial observability. This is to overcome another limitation in the hidden Markov model approaches which have appeared recently, where symptomatic individuals do not remain identified between observation points (e.g. Cooper and Lipsitch, 2004 & McBryde *et al.*, 2007). This arises from the major drawback of using hidden Markov models, in that the current output (observation) is assumed to be statistically independent of the previous output. Our approach is to explicitly model the number of observed colonised (symptomatic) indidviduals, in addition to the underlying actual number of colonised individuals, and then use marginal distributions for estimation purposes, consequently incorporating dependencies between observations. A comparison of the accuracy of these two approaches is undertaken.

## 2 MARKOV MODELS

Most models used for infectious disease modelling belong to a class of processes known as Markov population processes (Kingman, 1969), a type of continuous-time Markov chain. The evolution of such a Markov chain is governed by transition rates which we place in a matrix $Q = (q(m,n), m,n \in S)$, with $q(m,n)$ being the rate of transition from state $m$ to state $n$, for $n \neq m$, and $q(m,m) = -q(m)$, where $q(m) := \sum_{n \neq m} q(m,n)$ $(< \infty)$, is the total rate at which we leave state $m$. $S$ is the state space (set of all possible values the process may take on) and the model is specified by writing down $Q$.

As an example, consider the model for hospital infections presented by Cooper and Lipsitch (2004), being similar to the model of Pelupessy *et al.* (2002). Specifically, the model is a continuous-time Markov chain $(m(t), t \geq 0)$ ($m(t)$ is the number of colonised patients at time $t$) taking values in $S = \{0, 1, \ldots, N\}$ with transition rates

$$q(m, m+1) = \beta \frac{m}{N}(N-m) + \nu\mu(N-m)$$

$$q(m, m-1) = (1-\nu)\mu m,$$

where $\beta$ is the colonisation rate (per contact rate of infection transmission), $\mu$ is the per individual rate of discharge from the ward, $\nu$ is the probability of a new patient having already been colonised on admission and $N$ is the number of beds in the ward. We note that this model assumes that a discharged patient is immediately replaced by a new patient and thus assumes that all $N$ beds are occupied at any one time. The model we present in the next section removes this assumption by allowing for dynamic bed occupancy.

We will assume throughout the paper that $Q$ is *regular*, so that there is a unique transition function $P(t)$ with entries $p_{ij}(t)$ corresponding to the probability that the process moves from state $i$ to state $j$ in time $t$. Regularity is guaranteed if $Q$ is *bounded* in the sense that $q(m) \leq \alpha$ (for all $m$), for some constant $\alpha$, a condition that is trivially satisfied when there are finitely many states. In this case we may write $P(t) = \exp(Qt)$, where exp is the matrix exponential. In most cases the transition function cannot be evaluated explicitly. However there exists packages for efficient evaluation of the

required matrix exponential, provided the state space is not too large.

We will suppose that there is a parameter (or vector of parameters) $\theta$, contained in some parameter space $\Theta$, that must be estimated. We will allow the dependence on the parameters $\theta$ to be made explicit in our notation by writing $Q(\theta)$ for the transition rates and $P(\theta;t) = \exp(Q(\theta)t) = (p_{ij}(\theta;t),\, i,j \in S)$ for the transition function. We will also write $p_i(\theta;t)$ for the probability that the process is in state $i$ at time $t$; $p(\theta;t) = (p_i(\theta;t),\, i \in S)$ is taken here to be the stationary distribution as we assume the process is in stationarity. Given a set of $s$ observations $i_k = m(t_k)$ $(k = 1,\ldots,s)$ of the state of the process at times $(0 \leq\ )\, t_1 < \cdots < t_s$, the likelihood of observing them is

$$L(\theta) = p_{i_1}(\theta;t_1) \prod_{k=2}^{s} p_{i_{k-1},i_k}(\theta;t_k - t_{k-1}). \quad (1)$$

We may then calculate the maximum likelihood estimator (MLE) $\widehat{\theta}$ of the parameters $\theta$ by maximising the likelihood (1) over the parameter space $\Theta$ for the given observations. As noted previously, the transition function cannot usually be evaluated explicitly, but progress can be made by computing the transition probabilities numerically. This, combined with a numerical search algorithm over the parameter space $\Theta$, allows us to compute the desired MLE, provided it exists. For results concerning identifiability, and of existence and uniqueness, of the maximum likelihood estimator, we refer readers to Bladt and Sorensen (2005); essentially non-existence occurs when the sampling interval is too large in comparison to the rate of process dynamics.

To compute the required matrix exponentials, we use the `mexpv` and `padm` functions from the EXPOKIT package (Sidje, 1998) for MATLAB. Any one of a range of numerical optimisation techniques can be used to maximise (1). We use the *Cross-Entropy Method* (Rubinstein and Kroese, 2004), which has proved to be particularly effective for maximizing the likelihood functions that we consider (Ross *et al.*, 2006). This combination provides a useful tool for fitting continuous-time Markov chains to real systems, provided that the parameter space and the maximum population size is not too large. Thus, the method is typically ideal for calibrating models to hospital wards when modelling nosocomial infection dynamics and when assessing the efficacy of management strategies.

## 3 DYNAMIC BED OCCUPANCY

In this section we present a new stochastic model which explicitly models dynamic bed occupancy, and thus overcomes a limitation of methods currently in use (e.g. Pelupessy *et al.*, 2002 & Cooper and Lipsitch, 2004).

Our model is a two-dimensional continuous-time Markov chain defined as follows. Denoting by $n(t)$ and $m(t)$, respectively, the number of occupied beds and the number of colonised patients at time $t$, $\{p(t) = (n(t),m(t)),\, t \geq 0\}$ is assumed to be a Markov chain taking values in $S = \{(n,m) : 0 \leq m \leq n \leq N\}$ with non-zero transition rates

$$q((n,m),(n,m+1)) = \frac{\beta}{N} m(n-m),$$

corresponding to colonisation of a patient within the ward,

$$q((n,m),(n+1,m+1)) = \gamma 1_{\{n<N\}} \nu,$$

corresponding to a colonised individual being admitted to the ward,

$$q((n,m),(n-1,m-1)) = \mu_1 m,$$

corresponding to a colonised patient being discharged from the ward,

$$q((n,m),(n+1,m)) = \gamma 1_{\{n<N\}} (1-\nu),$$

corresponding to a non-colonised individual being admitted to the ward and

$$q((n,m),(n-1,m)) = \mu_2 (n-m),$$

corresponding to a non-colonised patient being discharged from the ward; the total number of beds in the ward is denoted by $N$ with the parameters $\beta$ and $\nu$ as before being the colonisation rate and probability of a new patient having already been colonised on admission, respectively, and $\mu_1$, $\mu_2$, $\gamma$ and $1_{\{\cdot\}}$ are, respectively, the rate at which colonised patients leave the ward, the rate at which non-colonised patients leave the ward, the rate of admission of new patients and the indicator function (which takes the value one when $\{\cdot\}$ is satisfied and zero otherwise).

We note that explicit modelling of bed occupancy allows for a more realistic model of infection transmission. Firstly it incorporates dynamic bed occupancy, and thus removes the need for the assumption of full ward occupancy, and its consequential effect on the rate of disease transmission. Additionally it allows for different rates of discharge of colonised and non-colonised patients, thus allowing a wider range of nosocomial infections to be modelled.

We also note that the rates of discharge $\mu_1$ and $\mu_2$, and the rate of admission $\gamma$, may be estimated from hospital administration data, and thus the only parameters requiring estimation from counts of symptomatic individuals are the rate of transmission $\beta$ and the probability of being already colonised on admission $\nu$. Estimation of these parameters is undertaken as outlined in the previous section — the specific parameters used are presented in the Results section to follow.

In the Results section, we also compare the results of estimation using the model presented above (with equal rates of discharge, $\mu_1 = \mu_2$, for fair comparison) to the model used by Cooper and Lipsitch (2004) which assumes full ward occupancy. This comparison is used to deduce conditions under which it is necessary to incorporate dynamic bed occupancy, and thus when it is necessary to use the methodology we have presented here.

## 4 PARTIAL OBSERVABILITY

In this section we present a new stochastic model which explicitly models the partial observability process, and discuss how this may be used for parameter estimation from counts of symptomatic individuals.

For clarity of exposition we present our method with respect to the one-dimensional model used by Cooper and Lipsitch (2004). However, the method may be equally applied to any model, including the dynamic bed occupancy model presented in the previous section — all we are doing is dividing colonised patients into two classes: those that display symptoms, and those that do not.

Our modified model is a two-dimensional continuous-time Markov chain defined as follows. Denoting by $m(t)$ and $n(t)$, respectively, the number of colonised patients and the number of symptomatic patients at time $t$, $\{p(t) = (m(t), n(t)),\, t \geq 0\}$ is assumed to be a Markov chain taking values in $S = \{(m, n) : 0 \leq n \leq m \leq N\}$ with non-zero transition rates

$$q((m,n),(m+1,n)) = \left[\frac{\beta}{N}m(N-m) + \nu\mu(N-m)\right](1-\delta),$$

corresponding to colonisation of a patient who does not display symptoms, or discharge of a non-colonised patient and admission of a colonised, asymptomatic patient,

$$q((m,n),(m+1,n+1)) = \left[\frac{\beta}{N}m(N-m) + \nu\mu(N-m)\right]\delta,$$

corresponding to colonisation of a patient who displays symptoms, or discharge of a non-colonised patient and admission of a patient who displays symptoms,

$$q((m,n),(m-1,n)) = (1-\nu)\mu(m-n),$$

corresponding to discharge of a colonised, asymptomatic patient and admission of a non-colonised patient,

$$q((m,n),(m-1,n-1)) = (1-\nu)\mu n,$$

corresponding to discharge of a symptomatic patient and admission of a non-colonised patient,

$$q((m,n),(m,n+1)) = \nu\mu(m-n)\delta,$$

corresponding to discharge of an asymptomatic patient and admission of a symptomatic patient and

$$q((m,n),(m,n-1)) = \nu\mu n(1-\delta),$$

corresponding to discharge of a symptomatic patient and admission of a colonised, asymptomatic patient; $\delta$ is the probability of a colonised patient being symptomatic, and thus identified as colonised, and all other parameters are as for the original model — $\beta$ is the colonisation rate, $\mu$ is the per individual rate of discharge from the ward, and $\nu$ is the probability of a new patient being already colonised on admission.

We use the marginal distribution of this process for parameter estimation since we only have data on the number of symptomatic patients $n(t)$ (the second dimension of the process). More specifically, we find $\theta = (\beta, \nu, \delta)$ which maximises the likelihood

$$L(\theta) = \left(\sum_{j=i_1}^{N} p_{(j,i_1)}(\theta; t_1)\right) \prod_{k=2}^{s}\left[\sum_{j=i_{k-1}}^{N}\sum_{l=i_k}^{N} p_{(j,i_{k-1}),(l,i_k)}(\theta; t_k - t_{k-1})\right].$$

We note that our approach to incorporating partial observability imposes a specific mechanism upon the observability process — namely a fixed probability of being symptomatic for every colonised patient. However, this mechanism should be ideal for nosocomial infection modelling, and additionally, as mentioned, it has the advantage of accounting for dependencies between symptomatic individuals.

A comparison of this approach to the hidden Markov approach of Cooper and Lipsitch (2004) will be undertaken in the next section.

## 5 RESULTS AND DISCUSSION

### 5.1 One-dimensional Model

Before progressing to our investigation of dynamic bed occupancy and partial observability, we first perform an investigation of the one-dimensional model used by Cooper and Lipsitch (2004) to assess if there is any bias in the procedure.

The data used consists of 20 independent simulations of the model, using parameters: colonisation rate $\beta = 0.255$, per patient rate of leaving $\mu = 0.125$, probability of a new patient being already colonised on admission $\nu = 0.028$ and $N = 16$ beds; the units for rates are days$^{-1}$. These parameters are in the range of typical values for nosocomial infections (Cooper and Lipsitch, 2004).

Each data set was collected by starting a simulation in state $m(0) = 10$ corresponding to 10 colonised patients, and the simulation was run for approximately $17,200$ transitions (corresponding to approximately $10,000$ days of data). The measurements were then taken daily, and the final $s = 50$ days worth of data taken as the data set. This method will produce data which is distributed approximately stationary; one could calculate the stationary distribution of the process and take the initial state of the simulation from this distribution and simulate for the desired data set length, but this makes little difference here and simulations are relatively cheap.

The unknown parameters requiring estimation are colonisation rate $\beta$ and probability of a new patient being already colonised on admission $\nu$, with $\mu$ assumed to be known (estimated) precisely. The minimum, maximum, median and mean of the parameters estimated from these 20 data sets are presented in Table 1.

| | $\hat{\beta}$ | $\hat{\nu}$ |
|---|---|---|
| Minimum | 0.0000 | 0.0088 |
| Maximum | 0.3399 | 0.3298 |
| Median | 0.1646 | 0.0779 |
| Mean | 0.1739 | 0.1159 |

**Table 1.** Parameter estimates using the model used by Cooper and Lipstich (2004); true parameters: colonisation rate $\beta = 0.255$ and probability of a patient being already colonised on admission $\nu = 0.028$.

In Figure 1 we plot all 20 estimates (crosses), the mean (circle), the median (square), the true value (dot), along with $50\%$ (dotted) and $95\%$ (solid) confidence ellipses; the ellipses are drawn using the covariance matrix of the 20 joint estimates. We note that maximum likelihood estimators are asymptotically normally distributed, so confidence ellipses may also be derived using this fact (Ross *et al.*, 2006), and additionally via other approaches (Cooper and Lipsitch, 2004).

These results suggest that the estimation procedure is biased, even using a data set of 50 daily observations. In particular estimates appear to underestimate the contribution of within-ward transmission, and overestimate the contribution from imported infection. However, we note that 6 of the 20 estimates for $\beta$ were within $\pm 10\%$ of the true value; only 4 of the 20 were within $\pm 30\%$ of the true parameter value for $\nu$.

### 5.2 Dynamic Bed Occupancy

We now investigate the accuracy of estimates using our dynamic bed occupancy model. The data used consists of 20 independent simulations for the dynamic bed occupancy model, using parameters: colonisation rate $\beta = 0.255$, per colonised patient rate of leaving $\mu_1 = 0.125$ and per non-colonised patient rate of leaving $\mu_2 = 0.125$, probability of a new patient being already colonised upon admission $\nu = 0.028$, rate of admission of new patients $\gamma = 2$ and $N = 16$ beds; once again the units for rates are days$^{-1}$.

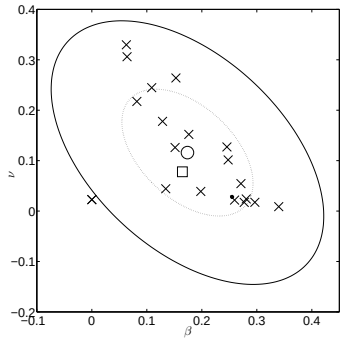


**Figure 1.** Maximum likelihood estimates of $(\beta, \nu)$ (crosses) along with the mean (circle), median (square) and true (dot) parameter values, and confidence ellipses ($50\%$ (dotted), $95\%$ (solid)) using the model used by Cooper and Lipsitch (2004).

Each data set was collected using the same approach, simulating for $10,000$ transitions from the initial state $(n(0), m(0)) = (10, 10)$ and taking the last $s = 50$ days worth of data. Due to space limitation we cannot present these data sets here, however we note that there exists a large amount of variation within, and between, them.

The unknown parameters requiring estimation are colonisation rate $\beta$ and probability of a new patient being already colonised on admission $\nu$, with all other parameters assumed to be known (estimated) precisely. The minimum, maximum, median and mean of the parameters estimated from these 20 data sets are presented in Table 2.

| | $\hat{\beta}$ | $\hat{\nu}$ |
|---|---|---|
| Minimum | 0.0000 | 0.0105 |
| Maximum | 0.2956 | 0.2864 |
| Median | 0.1602 | 0.0446 |
| Mean | 0.1702 | 0.0899 |

**Table 2.** Parameters estimates using the dynamic bed occupancy model; true parameters: colonisation rate $\beta = 0.255$ and probability of a patient being already colonised on admission $\nu = 0.028$.

These results demonstrate that there is a large degree of variation in the estimates and it appears that the estimates are again biased. Thus, when this method is used in practice, a simulation study such as this should be performed with parameters in the region of interest; biased corrected estimates may then be provided. We note that increasing the length of the observations will reduce bias. We also note, for future reference, that it appears that the colonisation rate $\beta$ is again underestimated and the probability of a

new patient being already colonised on admission $\nu$ is again overestimated.

In Figure 2 we plot all 20 estimates (crosses), the mean (circle), the median (square), the true value (dot), along with $50\%$ (dotted) and $95\%$ (solid) confidence ellipses; the ellipses are drawn using the covariance matrix of the 20 joint estimates.

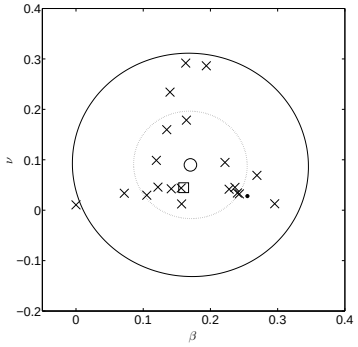


**Figure 2.** Maximum likelihood estimates of $(\beta, \nu)$ (crosses) along with the mean (circle), median (square) and true (dot) parameter values, and confidence ellipses ($50\%$ (dotted), $95\%$ (solid)) using the dynamic bed occupancy model.

We now use only data on the number of colonised patients $m(t)$ from these 20 data sets, and estimate the colonisation rate $\beta$ and probability of a patient being already colonised on admission $\nu$ using the one-dimensional model used by Cooper and Lipsitch (2004). This is to assess the impact of ignoring dynamic bed occupancy for parameter values used in our simulations. The minimum, maximum, median and mean of the parameters estimated from these 20 data sets are presented in Table 3.

| | $\hat{\beta}$ | $\hat{\nu}$ |
|---|---|---|
| Minimum | 0.0000 | 0.0092 |
| Maximum | 0.1966 | 0.2884 |
| Median | 0.1298 | 0.0681 |
| Mean | 0.1218 | 0.0964 |

**Table 3.** Parameters estimates using the one-dimensional model used by Cooper and Lipsitch (2004), from 20 simulated data sets from the dynamic bed occupancy model; true parameters: colonisation rate $\beta = 0.255$ and probability of a patient being already colonised on admission $\nu = 0.028$.

In Figure 3 we plot all 20 estimates (crosses), the mean (circle), the median (square), the true value (dot), along with $50\%$ (dotted) and $95\%$ (solid) confidence ellipses; the ellipses are drawn using the covariance matrix of the 20 joint estimates.

We note that ignoring dynamic bed occupancy has resulted in estimates which are further biased; the colonisation rate $\beta$ is further underestimated by, on average, an additional $0.0484$, and the probability of a patient being already colonised on admission $\nu$

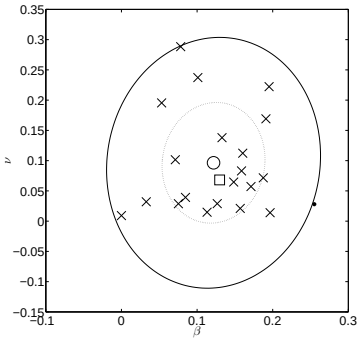


**Figure 3.** Maximum likelihood estimates of $(\beta, \nu)$ (crosses) along with the mean (circle), median (square) and true (dot) parameter values, and confidence ellipses ($50\%$ (dotted), $95\%$ (solid)) using the one-dimensional model used by Cooper and Lipsitch (2004).

is slightly further overestimated by, on average, an additional $0.0066$. This shows that ignoring dynamic bed occupancy can result in incorrect estimates. The underestimation of colonisation rate $\beta$ is expected whenever full ward occupancy is assumed and not actually realised in the data; this can be seen from inspection of the form of the models infection rate. Here we have assumed a rate of admission $\gamma = 2$, which corresponds to a lapse, on average, of $1/2$ a day before an empty bed is re-occupied. This is not an unrealistic assumption for some wards and demonstrates that care must be taken when estimating parameters for systems in which dynamic bed occupancy is a true feature of the system.

### 5.3 Partial Observability

We now investigate the accuracy of estimates using our partial observability process, and compare these to estimates using the hidden Markov model approach of Cooper and Lipsitch (2004). The data used consists of 20 independent simulations of the partial observability model, using parameters: colonisation rate $\beta = 0.255$, per patient rate of leaving $\mu = 0.125$, probability of a colonised patient presenting symptoms $\delta = 0.75$, probability of a new patient being already colonised on admission $\nu = 0.028$ and $N = 16$ beds; the units for rates are once again days$^{-1}$.

Each data set was collected using the same approach, simulating for $10,000$ transitions from the initial state $(n(0), m(0)) = (10, 10)$ and taking the last $s = 50$ days worth of data.

The unknown parameters requiring estimation are colonisation rate $\beta$, probability of a new patient being already colonised on admission $\nu$ and the probability of a colonised patient being symptomatic $\delta$, with all other parameters assumed to be known (estimated) precisely. The minimum, maximum, median and mean of the parameters estimated using our partial

observability method applied to these 20 data sets are presented in Table 4.

| | $\hat{\beta}$ | $\hat{\nu}$ | $\hat{\delta}$ |
|---|---|---|---|
| Minimum | 0.2115 | 0.0000 | 0.0062 |
| Maximum | 1.201 | 1.0000 | 0.8065 |
| Median | 0.7790 | 0.2836 | 0.3653 |
| Mean | 0.6885 | 0.3405 | 0.3699 |

**Table 4.** Parameter estimates using the partial observability process; true parameters: colonisation rate $\beta = 0.255$, probability of a patient being already colonised on admission $\nu = 0.028$ and probability of a colonised patient displaying symptoms $\delta = 0.75$.

We note that there exists a large amount of variation in these estimates; the colonisation rate $\beta$ is consistently overestimated, the probability of a new patient being already colonised on admission $\nu$ is consistently, and typically substantially, overestimated and the probability of colonised patient displaying symptoms $\delta$ is consistently underestimated. Since these are the best estimates that can be achieved given the data (and without correcting for bias), it appears that a larger data set is required to produce reliably accurate estimates. However, using the mean estimates – $\beta = 0.6885, \nu = 0.3405$ and $\delta = 0.3699$ – (or median estimates) produces an accurate estimate of the mean number of symptomatic patients, and provides a better estimate of the variability in the infection dynamics process than the mean (and median) estimates produced using the hidden Markov approach, presented below (Table 5).

The minimum, maximum, median and mean of the parameters estimated using Cooper and Lipstich's hidden Markov method applied to these 20 data sets are presented in Table 5; we note that $\lambda$ is parameter which is similar to $\delta$ in our method – see Section 3 of Cooper and Lipsitch (2004) for details.

| | $\hat{\beta}$ | $\hat{\nu}$ | $\hat{\lambda}$ |
|---|---|---|---|
| Minimum | 0.7836 | 0.8035 | 0.0625 |
| Maximum | 2.087 | 1.000 | 0.0666 |
| Median | 1.447 | 0.9770 | 0.0638 |
| Mean | 1.412 | 0.9510 | 0.0640 |

**Table 5.** Parameter estimates using the hidden Markov model approach applied to the partial observability data; true parameters: colonisation rate $\beta = 0.255$ and probability of a patient being already colonised on admission $\nu = 0.028$.

Once again a large amount of variation exists in these estimates. The colonisation rate $\beta$ is overestimated, and typically substantially, in all data sets. Similarly the probability of a new patient being already colonised on admission $\nu$ is overestimated, and substantially, in all data sets, with almost all estimates close to 1. These results suggest that the estimates provided in papers using the above hidden Markov model approach need to be viewed with care.

Future work will investigate the number of observations required such that reliably consistent and accurate estimates may be obtained using our partial observability process. We are also investigating methods to overcome another limitation encountered whenever using Markov chains for parameter estimation purposes: "when the state space becomes large (corresponding to a large number of beds) the algorithm becomes slow and numerical problems may occur." (page 234, Cooper and Lipsitch, 2004).

**Acknowledgements**. Thanks to Matt Keeling for valuable discussions. The support of the Leverhulme Trust and the Australian Research Council Centre of Excellence for Mathematics and Statistics of Complex Systems is gratefully acknowledged.

## 6 REFERENCES

M. Bladt and M. Sorensen. Statistical inference for discretely observed Markov jump processes. *J. Roy. Statist. Soc., Ser B*, 67:395–410, 2005.

B. Cooper and M. Lipsitch. The analysis of hospital infection data using hidden Markov models. *Biostatistics*, 5:223–237, 2004.

B.S. Cooper, S.P. Stone, C.C. Kibbler, B.D. Cookson, J.A. Roberts, G.F. Medley, G.J. Duckworth, R. Lai, and S. Ebrahim. Systematic review of isolation policies in the hospital management of methicillin-resistant *Staphylococcus aureus*. *Health Technology Assessment*, 7:1–194, 2003.

K. Inweregbu, J. Dave, and A. Pittard. Nosocomial infections. *Continuing Education in Anaesthesia, Critical Care & Pain*, 5:14–17, 2005.

J. F. C. Kingman. Markov population processes. *J. Appl. Probab.*, 6:1–18, 1969.

E. S. McBryde, A. N. Pettit, B. S. Cooper, D. L. S. McElwain. Characterising an outbreak of *Vancomycin-resistant enterococci* using hidden Markov models. *Journal of the Royal Society Interface* 4, 745-754, 2007.

G. Nimmo, J. Bel and P. Collignon. Fifteen years of surveillance by the Australian Group for Antimicrobial Resistance (AGAR). *Commun. Dis. Intel.* 27 (Suppl.), S47-S54, 2003.

I. Pelupessy, M.J.M. Bonten, and O. Diekmann. How to assess the relative importance of different colonization routes of pathogens within hospital settings. *Proc. Natl. Acad. Sci. USA*, 99:5601–5605, 2002.

J.V. Ross, T. Taimre, and P.K. Pollett. On parameter estimation in population models. *Theor. Popul. Biol.*, 70:498–510, 2006.

R.Y. Rubinstein and D.P. Kroese. *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation, and Machine Learning*. Springer-Verlag, New York, 2004.

R.B. Sidje. EXPOKIT. A software package for computing matrix exponentials. *ACM Trans. Math. Software*, 24:130–156, 1998.