Using Phase-type Models to Monitor and Predict Process Target Compliance

Sally I. McClean¹, David A. Stanford², Lalit Garg³ and Naveed Khan⁴

¹School of Computing, Ulster University, Coleraine, BT52 1SA, U.K.
²Department of Statistical & Actuarial Sciences, Western University, London, ON N6A 5B7, Canada
³Faculty of Information & Communication Technology, University of Malta, Msida, Malta
⁴School of Computing, Ulster University, Jordanstown, BT37 0QB, U.K.

Keywords: Process Mining, Process Modelling, Phase-Type Models, Targets.

Abstract: Processes are ubiquitous, spanning diverse areas such as business, production, telecommunications and healthcare. They have been studied and modelled for many years in an attempt to increase understanding, improve efficiency and predict future pathways, events and outcomes. More recently, process mining has emerged with the intention of discovering, monitoring, and improving processes, typically using data extracted from event logs. This may include discovering the tasks within the overall processes, predicting future trajectories, or identifying anomalous tasks. We focus on using phase-type process modelling to measure compliance with known targets and, inversely, determine suitable targets given a threshold percentage required for satisfactory performance. We illustrate the ideas with an application to a stroke patient care process, where there are multiple outcomes for patients, namely discharge to normal residence, nursing home, or death. Various scenarios are explored, with a focus on determining compliance with given targets: such KPIs are commonly used in Healthcare as well as for Business and Industrial processes. We believe that this approach has considerable potential to be extended to include more detailed and explicit models that allow us to assess complex scenarios. Phase-type models have an important role in this work.

1 INTRODUCTION

Processes are ubiquitous, spanning diverse areas such as business, production, telecommunications and healthcare, and have been studied and modelled for many years in an attempt to increase understanding, improve efficiency and predict future pathways, events and outcomes. With the ever-increasing capability of computer systems to collect, process, store and exchange data and the advent of the Big Data era, the concept of Process Mining has emerged to form a bridge between data mining and process modelling (Van Der Aalst, 2012). Process Mining provides a framework for service design, an underpinning for process improvement and a scientific basis for decision making. For example, a process can be concerned with how an execution of process instances should occur, where each execution instance of a business process is identified as one process instance. A process instance contains information on the tasks executed and the attributes observed during execution. A process instance execution of the tasks is performed according to the structure and definitions in the real-world process. The process attributes typically consist of information such as start time, end time, customer name etc. and are stored in log files. Hence, a log file typically provides an automatically produced and time-stamped documentation of events relevant to a particular system.

In general, process mining aims to discover, monitor, and improve processes by extracting data from event logs. This may include discovering the tasks within the overall processes, predicting future process trajectories, or identifying anomalous tasks and task sequences. Such process mining activities can build on standard approaches to data mining problems such as classification, clustering, regression, association rule learning, and sequence mining or more recent approaches for Big Data, such as deep learning. However, if the structure of the process is known, model-based approaches can also be useful for incorporating structural process knowledge into the analysis and simplifying the problem. Thus process mining can be use in various applications such as manufac-
Using Phase-type Models to Monitor and Predict Process Target Compliance

2 BACKGROUND

Markov models have proved to be a useful representation of process behaviour in many contexts, including call centres (Dudin et al., 2016), sensor networks (Dudin and Lee, 2016), telecommunications (Vishnevskii and Dudin, 2017), production modelling...
Phase-type models are a type of Markov model with a number of transient states (or phases) and a single absorbing state. Such models can be used to predict individual behaviour or to assess future resource needs and costs. They are intuitively appealing as they conceptualise process progression, for instance, through acute care, into treatment, and on to rehabilitation. A phase type distribution (PHD) describes a non-negative random variable (generally a duration) generated by a Markov model where the PHD represents the duration from admission to the transient states of the Markov process until absorption in one of the recurrent states. In particular, Coxian phase type distributions (C-PHDs) are a special case of PHDs in which a process always starts in the first transient state, and only sequential transitions are allowed between transient states; transition from any state to the absorbing state is also possible (Figure 1). PHDs provide a simple description of a variable such as length of stay (LOS) in hospital, duration of a particular activity of daily living, or duration from order placement to completion in a business process. PHDs also typically have the advantage of ease of parameter estimation (Garg et al., 2012). In particular, Coxian phase-type models (C-PHDs) work well for a range of settings and scales, including hospitals (Fackrell, 2009), (Tang et al., 2012), (Marshall and Zenga, 2012), and (Griffiths et al., 2013), community care (Xie et al., 2006), emergency services (Knight and Harper, 2012) and patient activity recognition (Duong et al., 2009). They are also intuitively appealing as; for example, we can think of a patient as progressing through various phases of hospital, social or community care such as acute care, treatment, rehabilitation and long stay (Figure 1).

From the technical point of view, the advantages of using the PHDs are (i) their mathematical tractability; (ii) parsimonious parameterization, in the case of C-PHDs, - a general phase-type representation requires a large number of parameters, with associated difficulties in estimation; (iii) flexibility - any distribution can be approximated by a phase-type distribution with an appropriate number of parameters; and (iv) the ease with which such representations can be migrated to a more complex setting.

Our previous work (Garg et al., 2012) developed a framework that classifies the patient stays based on identifying homogeneous groups, or classes, in terms of their LOS distributions; where different admission probabilities pertain to different classes. Classes are characterized using appropriate covariates, which in our case were gender, age, diagnosis and outcome. Patients in the various classes follow separate pathways, with correspondingly different admission probabilities for each class. Another feature of our framework is that, unlike earlier work, we allow for a number of absorbing states - for example, these might be the patient’s normal residence, a private nursing home (PNH), or death. Such an approach allows us to extend phase-type models to describe community as well as hospital states thus modelling an integrated system of stroke patient care, rather than sub-systems of the overall care process. Generally speaking, we can think of this model as a mixture of parallel C-PHDs, with multiple absorbing states.

Hidden Markov models (HMMs) are similar to phase-type models in that the system being modelled is also assumed to be Markovian, with unobserved (hidden) states. However, HMMs are generally more focussed on the pathway through the Markov system while phase-type models have a particular relevance to duration in the transient states and have their origins in queueing theory where performance, in terms of timeliness, is central. HMMs are especially known for their application to temporal pattern recognition such as speech processing (Rabiner, 1989). They have also been extended to include parallel and hierarchical structures, for example the hierarchical hidden Markov model (HHMM) (Fine et al., 1998). Such approaches have already been used in healthcare settings, in particular for patient temporal pattern discovery from Australian medical claims data (Tsoi et al., 2005). The HHMM here profiles the patients into sub-behavioural groups based on similar temporal profiles and medical behaviours. Another important healthcare use of HHMMs has been for activity recognition, specifically aimed at developing automated reminder systems for patients with dementia (Youngblood and Cook, 2007). Here, repeated behaviours in sensorised smart homes are observed and categorized into patterns that represent the inhabitant behaviour. As for patients moving through a health system, home-based activities also move through a process, comprising a sequence of smaller steps, or phases, such as treatment or rehabilitation in the former case or instrumental activities of daily living (IADLs), such as taking pills or making a cup of tea, in the latter case.

In this paper, based on our previous work (Faddy and McClean, 2005), (Jones et al., 2018), we incorporate the covariates into the model by allowing the tran-
sition rates between states of the underlying Markov model to depend explicitly on appropriate covariates. Based on process data, the model can then be pruned by eliminating those covariates which are not statistically significant, on the basis of likelihood-based statistical tests. The specific functional form of this covariate model will be described in the next section.

We note that the approach proposed here is to use a PHD, which may, or may not be, Coxian (C-PHD). However in what follows we illustrate the ideas with a specific healthcare case study where the patients progress through successive states of health, so we assume that they follow a C-PHD. Also, in general we propose to incorporate covariates by modelling them through an explicit functional dependency for the input, exit and transition probabilities. This is also illustrated in the healthcare case study.

3 PHASE-TYPE MODELS

3.1 The Basic Phase-type Model

As in (McClean et al., 2011) we employ a phase-type model for process planning, based on Markov phase-type models. Thus the phase-type model can allow us to easily implement and quickly evaluate changes in process circumstances. Frequently a C-PHD is used and is intuitively appealing as we can think of the process as progressing through sequential phases, without loops, before completion (Figure 1). As discussed, C-PHDs provide a simple description of a variable such as length of stay in hospital or duration of an activity of daily living and also have advantages over other types of PHDs, such as ease of parameter estimation.

We begin by considering a basic C-PHD and then extend the model to a general PHD with entry to any transient state, and \( k \) absorbing states. We thus initially consider a system of \( k+1 \) states (or phases) and a Markov stochastic process defined according to the transition probabilities defined for \( i = 1, 2, \ldots, k-1 \)

\[
\begin{align*}
P(X(t + \delta t) = i + 1|X(t) = i) &= \frac{\lambda_i \delta t + o(\delta t)}{\lambda_i + \mu_i}, \\
P(X(t + \delta t) = k + 1|X(t) = i) &= \frac{\mu_i \delta t + o(\delta t)}{\lambda_i + \mu_i}.
\end{align*}
\]

(The latter of these equations applies as well for \( i = k \).) Here the rates \( \lambda_1, \lambda_2, \ldots, \lambda_{k-1} \) describe sequential transitions between phase \( S_i \) and the subsequent phase \( S_{i+1} \) while \( \mu_1, \mu_2, \ldots, \mu_k \) describe transitions from phase \( S_i \) to the absorbing state \( S_{k+1} \) for \( i = 1, 2, \ldots, k \) (see Figure 1). If \( \mu_i > 0 \) for \( i = 1, 2, \ldots, k \) then phases \( S_1, S_2, \ldots, S_k \) are transient while phase \( S_{k+1} \) is the unique absorbing state. Writing the admission vector as

\[
\alpha' = (\alpha_1, \alpha_2, \ldots, \alpha_k)
\]

where \( \alpha_i \) denotes the probability of admission to phase \( S_i; i = 1, 2, \ldots, k \), we obtain the probability density function (p.d.f.) for the distribution of time until absorption as

\[
f(x) = \alpha' \exp(Tx)t_0.
\]

In the foregoing formula, \( t_0 = -T1 \) represents the absorption rate vector from the various transient states. For a C-PHD with \( k \) transient phases, the infinitesimal generator (rate) matrix \( T \) is null except for the main diagonal and prime super-diagonal, which are defined for \( i = 1, 2, \ldots, k-1 \) by

\[
T_{ii} = -(\lambda_i + \mu_i); \quad T_{i,i+1} = \lambda_i.
\]

Furthermore, \( T_{ik} = -\mu_k \). Integrating the p.d.f., one obtains the cumulative distribution function

\[
F_X(y; \alpha, T) = 1 - \alpha' \exp(Ty)1; y \geq 0.
\]

which represents the probability of meeting a given time target \( y \). Conversely, the probability of missing a time target \( y \) is given by

\[
\tilde{F}_X(y; \alpha, T) = \alpha' \exp(Ty)1; y < 0.
\]

In this way, it is possible for us not only to quantify the relative likelihoods of compliance and exceedance, but also to formulate sub-tending network paths for the corresponding actions that would arise in each of these cases. Using the same approach as (Jones et al., 2018), we can find the conditional probability of meeting (or alternatively, missing) a given target at time \( y \) given that a known amount of time (say, \( d \)) has already elapsed. The chance of the former is given by

\[
F_X[y > d; y; \alpha, T] = 1 - \frac{\alpha' \exp(Ty)}{\alpha' \exp(Td)}1; y \geq d.
\]

while the chance of the latter is given by

\[
\tilde{F}_X[y > d; y; \alpha, T] = \frac{\alpha' \exp(Ty)}{\alpha' \exp(Td)}1; y \geq d.
\]

In like fashion, conditional means can be computed by integrating the corresponding conditional densities over the appropriate regions.


4 CASE STUDY: STROKE CARE

In many cases there are several possible absorbing states for termination of a process, corresponding to different targets. In previous work we have extended the basic phase-type models to accommodate such multiple absorbing states (McClean et al., 2011), (Jones et al., 2018). This model pertains to care of stroke patients using data collected over a period of 5 years. In this case we have identified 4 transient states of the phase-type model which relate to different types of stroke with differing severity and corresponding absorption probabilities for decreasing severity of stroke. The model allows for three different types of stroke: haemorrhagic (the most severe, caused by ruptured blood vessels that cause brain bleeding), cerebral infarction (less severe, caused by blood clots) and transient ischaemic attack or TIA (the least severe, a mini-stroke caused by a temporary blood clot). In this setting, there are three possible ways in which the hospital stay can conclude: 1) upon the patient’s demise, 2) with a transfer to a nursing home, and 3) with a return to the patient’s usual residence. These differing possibilities can be handled readily by replacing the absorption rate vector \( t_0 \) by a matrix of distinct absorption rates \( t_i \) defined as follows:

\[
    t_i = \begin{pmatrix}
        \mu_1 & \nu_1 & \rho_1 \\
        \mu_2 & \nu_2 & \rho_2 \\
        \mu_3 & \nu_3 & \rho_3 \\
        \mu_4 & \nu_4 & \rho_4
    \end{pmatrix}
\]  

(9)

In (9), each column of the matrix corresponds to a distinct concluding event for the hospital stay (demise, nursing home, and usual residence, respectively) while each row refers to the particular transient phase from which the admission occurred. Each such phase corresponds to a particular recovery stage from which absorption occurred, with the first corresponding to the sickest individuals, and the last to the least sick. The interested reader is directed to (Jones et al., 2018) where the rational behind the four transient states and other details are fully described.

The univariate density of time to absorption may still be written as before, or alternatively as

\[
    f_x(y | \alpha, T) = \alpha' \exp(Ty) t_i e_j, \quad y \geq 0
\]  

(10)

which is equivalent. If one wishes instead to determine the joint density of absorption together with a particular concluding event (say, the \( j \)th possible cause), the corresponding formula is given by

\[
    f_x(y | \alpha, T) = \alpha' \exp(Ty) t_i e_j, \quad y \geq 0
\]  

(11)

where \( e_j \) denotes a column vector with unity in the \( j \)th position, and zeroes elsewhere.

In this case study data were collected over a 5 year period, on admission date, length of stay in hospital, diagnosis and discharge destination, and other covariates, such as age on admission and gender. The transition rates of the model comprise both those that depend upon the age and stroke type of the patient, and those which do not depend on age; these dependencies were previously established using statistical analysis.

For \( i = 1, 2 \), let \( \lambda_i(x) \) be the transition intensity from phase \( i \) to phase \( i + 1 \) for a patient who is age \( x \), where \( \lambda_i(x) = \exp(p_i + \beta_i x) \). Also, let \( p(x) \) represent the probability that a TIA stroke patient age \( x \) is in recovery phase 4 upon admission to hospital (representing the less severe TIAs). Consequently, a TIA patient starts in phase 3 with probability 1 − \( p(x) \). We assume that \( p(x) = \exp(- \exp(\theta_0 + \theta_1 x)) \). The exponential functions used in modelling \( \lambda_i(x) \) and \( p(x) \) are fairly standard, and ensure that their values are constrained to the required ranges. These functions arise when using the log link and complementary log-log link functions in generalized linear models (Dobson and Barnett, 2008). As indicated in Figure 1, it is assumed that \( \mu_4 = \nu_4 = 0 \).

We note that the routes missing from the diagram in Figure 1, and corresponding zero parameters, are determined by statistical testing based on Likelihoods; for further details see (Jones et al., 2018).

4.1 Interventions to Increase Compliance Likelihood

Another benefit of the phase-type approach is that it allows the user to explore the probabilistic consequences of interventions to enhance the chances of compliance, provided that the sequence of tasks involved in the intervention can likewise be described by a C-PHD process. Using such an approach, in Section 3.1 we have provided formulae to predict the probability of compliance with a given completion target, assuming that current patterns (i.e. parameter estimates) continue. However, an important follow-up question is clearly: can we intervene to modify or “restart” the process at the current point in time and can we then predict the new probability of meeting the target.

Following the logic behind equations (7) and (8), the vector for the conditional probability of being in the various transient states at the intervention instant \( d \) is readily seen to be given by \( \alpha' = \alpha' \exp(Td) / [\alpha' \exp(Td)|1] \).

Now let \( S \) denote the part of the infinitesimal generator restricted to the transient states for the enhanced C-PHD process including the intervention
Figure 2: Stroke care transition diagram.

which need not entail the same number of transient states as the original process). The probability vector \( \beta \) for starting the enhanced process in the various possible states \( \beta \) denote the vector of starting the enhanced process at time \( d \) in the various transient states will be a direct function \( \beta = f(\omega) \), such as perhaps merely repositioning the components of \( \omega \) in some enlarged state space.

The corresponding probability of compliance by time \( y \) given an intervention time \( d \) is then given by

\[
F_{X|X>d}(y;\beta,S) = 1 - \beta' \exp(S(y-d))1; \quad y \geq d, \quad (12)
\]

The other extensions can be found in like fashion. We note here that the inverse problem of determining a suitable target value, given a desired service level can also be derived from this equation, by solving to find the target \( y \) for a given value of \( F \). Here \( F \) can be thought of corresponding to a service level agreement, where, for example, we may require 95% of jobs to be completed within a given time. Although we cannot solve equation (5) explicitly for \( y \), we can instead use a numerical approach, such as Newton-Raphson where the estimate of \( y \) is given at the \( n+1 \)st iteration \( y_{n+1} \) is given by:

\[
y_{n+1} = y_n - F(y_n)/F'(y_n) \quad \text{where} \quad F'(y) = \alpha' \exp(Ty)T1; y \geq 0, \quad (13)
\]

4.2 Estimating the Model Parameters

As discussed, a phase-type model typically contains a number of parameters, representing the initial entrance probabilities to each state, the transition rates between transient states and exit transition rates from each state to the absorbing state. In addition there are often covariates, both static (process covariates or features) and associated dynamic covariates with individual log files. In process mining, these covariates are extracted from the logs and can be used to improve the model. Two common strategies are (1) to use the covariates to cluster the process instances and then use a cluster specific model to predict outcomes and (2) to explicitly model the parameters as functions of the covariates, thus facilitating more parsimonious models and less data-consuming estimation (McClean et al., 2011).

In our case study we employ the latter strategy; this is partly due to inherently limited data in such healthcare applications (Jones et al., 2018). The model which we eventually decided upon represents the smallest one to allow for sufficient distinction of the various types of stroke. It also has the desirable statistical property of being more parsimonious in terms of the number of parameters to be estimated than a larger model. Its state transition diagram is shown in Figure 2.

4.3 Results and Discussion

Fig. 3, 4 and 5 present the choice of target as a function of the probability of compliance for different types of stroke and each of the modes of discharge: death, nursing home, and usual residence. In the case of haemorrhagic strokes, we see that, for a given probability of compliance, the target should be lowest, with TIAs highest, representing the fact that haemorrhagic strokes are most severe and TIAs least severe.

Figure 3: Number of days to achieve the target compliance for patients with Haemorrhagic stroke.
Similarly, for discharge destination, for a given target, the probability of compliance should be lowest for discharge to the usual residence, as an extended period is required for such patients to pass through the corresponding recovery phases before discharge. This is reflected in the observation that for a given probability of compliance, the target of such patients is lowest. In the figure we have highlighted the targets for 95% compliance as this is a typical value.

Overall, modelling can be used to characterise the whole system of stroke patient care and the associated clinical pathways, integrating hospital and community services to provide tools for describing current services, assessing the impact of proposed changes, and predicting resource requirements in future scenarios. Our previous paper (Jones et al., 2018) focussed on developing models that use routinely available hospital discharge data to describe patient admissions, movements through hospital, and discharge modes. Such models can be used to facilitate performance modelling, bed occupancy analysis, capacity planning, and prediction of patient numbers in different components of the overall care system. By using such a model to quantify resource consumption, and costs of such proposed interventions, we can compare different solutions and determine optimal strategies. Stroke patient care thus provides an important paradigm example for healthcare process modelling, as there are many possible interventions which straddle hospital and community services. However, for such models to be effective, a robust estimation process and thorough evaluation is essential. Overall length of stay in hospital and compliance with related targets are key performance indicators for hospital services, and it is therefore useful to assess the impact of key interventions, in terms of their impact on the achievement of length of stay targets.

5 SUMMARY AND FURTHER WORK

This paper has focussed on using process mining to extract data on processes and learn appropriate parameters for phase-type models where we focus on using such models to measure compliance with known targets or determine suitable targets given a threshold percentage required for satisfactory performance. We have described an application of such phase-type models to stroke patient care, where there are multiple outcomes for patients, such as discharge to normal residence, nursing home, or death. Based on these data, various scenarios have been explored, with a focus on determining compliance with given
targets; such KPIs are commonly used in Healthcare as well as for Business and Industrial processes.

Our current framework represents initial work towards developing integrated models for processes, with the aim of supporting cohesive management and planning. However, we believe that it also has considerable potential to be extended to include more detailed and explicit models that allow us to assess complex scenarios involving interactions between processes. Also, our current analytic model has the advantage that the results are based on routinely available data. Another important aspect of extending our current framework is to consider the distributions and moments of numbers of process instances complying with targets for multiple absorbing states, including processes using Poisson arrivals to describe streamed data of independent instances; costs can also be associated with various options within the model. We plan to explore such options in further work.

The experience gained and techniques learned are likely to be relevant to business processes in general. Phase-type models have an important role in this work.

ACKNOWLEDGEMENT

This research is partly supported by BTIC (BT Ireland Innovation Centre), funded by BT and Invest Northern Ireland and by the Natural Sciences and Engineering Research Council of Canada (NSERC).

REFERENCES


Tang, X., Luo, Z., and Gardiner, J. C. (2012). Modeling hospital length of stay by coxian phase-type re-