

UNIVERSITY OF WESTMINSTER



WestminsterResearch

<http://www.wmin.ac.uk/westminsterresearch>

Markov model-based clustering for efficient patient care.

Sally McClean¹
Malcolm Faddy²
Peter Millard³

¹ University of Ulster, Northern Ireland

² Queensland University of Technology, Brisbane

³ School of Informatics, University of Westminster

Copyright © [2006] IEEE. Reprinted from the of the 18th IEEE Symposium on Computer-Based Medical Systems, IEEE CBMS 2005, Dublin, Ireland, 23-24 June 2005. IEEE, Los Alamitos, USA, pp. 467-472. ISBN 0769523552.

This material is posted here with permission of the IEEE. Such permission of the IEEE does not in any way imply IEEE endorsement of any of the University of Westminster's products or services. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE. By choosing to view this document, you agree to all provisions of the copyright laws protecting it.

The WestminsterResearch online digital archive at the University of Westminster aims to make the research output of the University available to a wider audience. Copyright and Moral Rights remain with the authors and/or copyright owners. Users are permitted to download and/or print one copy for non-commercial private study or research. Further distribution and any use of material from within this archive for profit-making enterprises or for commercial gain is strictly forbidden.

Whilst further distribution of specific materials from within this archive is forbidden, you may freely distribute the URL of the University of Westminster Eprints (<http://www.wmin.ac.uk/westminsterresearch>).

In case of abuse or copyright appearing without permission e-mail wattsn@wmin.ac.uk.

Markov Model-Based Clustering for Efficient Patient Care

Sally McClean¹, Malcolm Faddy², and Peter Millard³

¹ *University of Ulster, Northern Ireland*

² *Queensland University of Technology, Brisbane, Australia*

³ *University of Westminster, London, UK*

si.mcclean@ulster.ac.uk, m.faddy@qut.edu.au, phmillard@tiscali.co.uk

Abstract

Phase-type distributions were used to carry out model-based clustering of patients using the time spent by the patients in hospital, with maximum likelihood estimation of the model parameters. These parameters were allowed to vary with covariates so that the probability of cluster membership was dependent on these covariates. Expressions for the cluster membership probabilities and corresponding distributions of length of stay in care were found where the membership probabilities can be updated to take account of length of stay to date. The approach was applied to data on geriatric patients from an administrative database of a London hospital. The age of the patients at admission to care and the year of admission were included as covariates. Differential effects of these covariates on the various parameters of the fitted model were demonstrated, and interpretations of these effects made. The clusters here corresponded to patient pathways, with different length of stay distributions, varying care needs and different associated costs. By using the membership probabilities to assign patients to such clusters, care may thus be suited to their predicted pathway. Such an approach might be used in association with healthcare process improvement technologies, such as Lean Thinking or Six Sigma.

1. Introduction

Hospital patients may be thought of as progressing through phases such as acute care, assessment, diagnosis, rehabilitation and long-stay care. Most hospital patients are eventually rehabilitated and discharged. Those who become long-stay may remain in hospital for months, or even years. These patients are very consuming of resources thereby distorting performance statistics and costs¹.

Such processes may be modelled using continuous time Markov chains, where there is a single absorbing state and the process starts in one of the transient states. The time to absorption is then described by a phase-type distribution². Such phase-type distributions have considerable generality, and include exponential, Erlang and mixed exponential distributions. In fact, the distribution of any non-negative random variable can be arbitrarily closely approximated by one of phase-type form. In addition, covariates may be incorporated into the parameterisation, thus enabling the modelling to describe complex processes.

However, this generality can lead to difficulties in estimating parameters defining the transitions between the states of the Markov chain, due to identifiability problems. Such difficulties can to some extent be overcome by using the Coxian sub-class^{3,4} with n transient states (or phases) and the process starting in the first of these, then movement through them sequentially with different probabilities of absorption from each transient state. In our application these transient states are phases of hospital care, and absorption

represents discharge from hospital. In this paper, we use the model to cluster patients into classes on the basis of the number of phases involved. Such clusters may be regarded as different patient pathways.

The data used to illustrate this approach are described in McClean and Millard⁵ and refer to 2090 male geriatric patients at St George's Hospital, London, over the period 1969–85. Durations of hospital treatment were available for these patients, along with two covariates: age at admission and year of admission. The analysis was concerned with clustering the patients into different pathways, based on the covariates and their current length of stay in hospital. Cluster membership probabilities can be further updated as the patients' treatment progresses

This work extends that of Faddy and McClean^{3,4} to provide an explicit identification of clusters thus facilitating the use of healthcare process improvement technologies, such as Lean Thinking or Six Sigma. In recent years there has been considerable interest in the possibility of using such ideas from manufacturing and engineering to improve healthcare, where a key concept is the clustering of patients into more homogenous classes followed by improvement within clusters to increase efficiency. Our current approach can be used alongside such developments.

2. Clustering

We consider a system of $n+1$ states (or phases) and a Markov stochastic process to describe the patients' stay in hospital, as illustrated in Figure 1.

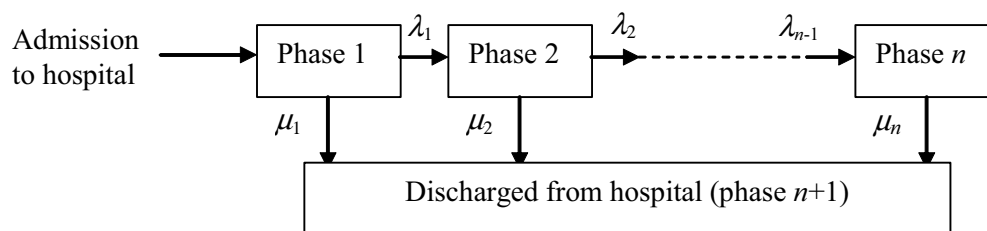


Figure 1. The Coxian phase-type model

Here the parameters $\lambda_1, \lambda_2, \dots, \lambda_{n-1}$ describe sequential transitions between phases $1, 2, \dots, n$ and $\mu_1, \mu_2, \dots, \mu_n$ describe transitions from phases $1, 2, \dots, n$ to phase $n+1$. If $\lambda_1, \lambda_2, \dots, \lambda_{n-1}$ and (at least) μ_n are all positive then phases $1, 2, \dots, n$ are transient and phase $n+1$ is absorbing with absorption occurring with probability one.

Now $p_i = \frac{\lambda_1}{\lambda_1 + \mu_1} \frac{\lambda_2}{\lambda_2 + \mu_2} \dots \frac{\lambda_{i-1}}{\lambda_{i-1} + \mu_{i-1}} \frac{\mu_i}{\lambda_i + \mu_i}$ is the probability of moving from phase 1 to phase 2, then from phase 2 to phase 3, and so on until phase i is reached, and then absorption occurs from this phase; *i.e.*, p_i is the probability that the time in care ends from phase i . And the probability density function of the time of absorption, given that absorption occurs from this phase i , will be:

$$f(t | i) = \mathbf{p}_i \exp(\mathbf{Q}_i t) \mathbf{q}_i \quad (1)$$

where:

$$\mathbf{p}_i = (1 \ 0 \ 0 \ \dots \ 0), \quad (2)$$

$$\mathbf{Q}_i = \begin{pmatrix} -(\lambda_1 + \mu_1) & (\lambda_1 + \mu_1) & 0 & \cdots & 0 \\ 0 & -(\lambda_2 + \mu_2) & (\lambda_2 + \mu_2) & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -(\lambda_i + \mu_i) \end{pmatrix} \quad (3)$$

and

$$\mathbf{q}_i = [0 \ 0 \ 0 \ \cdots \ (\lambda_i + \mu_i)]^T. \quad (4)$$

[Here, λ_n is taken to be zero in determining p_n and $f(t|n)$.]

So, for example, when $n=2$:

$$p_1 = \frac{\mu_1}{\lambda_1 + \mu_1}, \quad f(t|1) = (\lambda_1 + \mu_1)e^{-(\lambda_1 + \mu_1)t} \quad \text{and}$$

$$p_2 = \frac{\lambda_1}{\lambda_1 + \mu_1}, \quad f(t|2) = \frac{(\lambda_1 + \mu_1)\mu_2}{\lambda_1 + \mu_1 - \mu_2} (e^{-\mu_2 t} - e^{-(\lambda_1 + \mu_1)t}).$$

The distribution of length of stay in care is then a mixture distribution, with mixing probabilities p_i and corresponding component densities $f(t|i)$, $i=1,2,\dots,n$, so that its probability density function is:

$$f(t) = \sum_{i=1}^n p_i f(t|i). \quad (5)$$

Thus there are different classes of patient corresponding to the components in this mixture distribution: class i includes patients who will eventually be discharged from phase i . This approach may be regarded as a type of model-based clustering⁶ where the component length of stay distributions are generalized Erlang, equations (1) – (4), rather than Gaussian. In addition, we may adapt the class probabilities with increasing length of stay, since the conditional probability:

$$\Pr(\text{class } i | \text{still in care after time } t) = \frac{p_i P_i(t)}{\sum_{j=1}^n p_j P_j(t)} = c_i(t), \quad (6)$$

where:

$$P_i(t) = \Pr(\text{in care at time } t | \text{class } i) = \int_t^{\infty} f(x|i) dx = \mathbf{p}_i \exp\{\mathbf{Q}_i t\} \mathbf{e}_i \quad (7)$$

and $\mathbf{e}_i = (1 \ 1 \ \cdots \ 1)^T$, from equations (1) – (4).

So, for example, when $n=2$:

$$P_1(t) = e^{-(\lambda_1 + \mu_1)t} \quad \text{and}$$

$$P_2(t) = \frac{(\lambda_1 + \mu_1)\mu_2}{(\lambda_1 + \mu_1 - \mu_2)} \left(\frac{e^{-\mu_2 t}}{\mu_2} - \frac{e^{-(\lambda_1 + \mu_1)t}}{(\lambda_1 + \mu_1)} \right).$$

We may therefore estimate using equations (6) and (7) the most likely class i , given time spent in hospital to date is t , such that $c_i(t) > c_j(t)$ for all $j \neq i$.

Dependence on covariates $\mathbf{x} = (x_1 \ x_2 \ \dots \ x_m)^\top$ can be incorporated into the model by having the λ_i and μ_i transition rates in equations (3) and (4) take the form:

$$\exp(a + \mathbf{b}^\top \mathbf{x}) \quad (8)$$

with coefficient parameters a and \mathbf{b} estimated by maximum likelihood for each of these transition rates. Starting with $n = 1$ phase, n can be increased until a distribution that adequately describes the variation in the observed data is obtained. Determination of standard errors of parameter estimates and assessment of significance of covariate effects can be carried out using asymptotic likelihood theory.

3. Results

For the data analysed here, there were two covariates: x_1 = patient's age at admission and x_2 = year of admission. In Faddy and McClean^{3,4} maximum likelihood estimates of the parameters of a four phase fit to the data on durations of hospital treatment were presented. The covariate effects remaining after backwards elimination, based on the criterion estimate/standard error < 1, were such that, for age in years and year of admission measured from an origin of 1900, the estimated parameters (with standard errors in brackets) were:

$$\begin{aligned} \hat{\mu}_1 &= \exp\{-5.82(1.29) + 0.027(0.017) \times year\} \\ \hat{\lambda}_1 &= \exp\{-4.49(1.10) + 0.024(0.015) \times year\} \\ \hat{\mu}_2 &= \exp\{-3.64(0.62) + 0.012(0.008) \times year\} \\ \hat{\lambda}_2 &= \hat{\mu}_1 + \hat{\lambda}_1 - \hat{\mu}_2 \\ \hat{\mu}_3 &= \exp\{-6.25(0.74) + 0.027(0.009) \times age\} \\ \hat{\lambda}_3 &= \exp\{0.41(3.27) - 0.098(0.043) \times age\} \\ \hat{\mu}_4 &= \exp\{-6.14(0.29)\}. \end{aligned}$$

Four phases were chosen here using penalised maximum likelihood with a penalty against multi-modal distributions⁷. The probabilities of each class for different lengths of stay and several scenarios along with the allocated classes are given in Table 1 and Figure 2.

We have previously interpreted the four classes as follows⁴. Patients in class 1 either had very little wrong with them, typically younger patients who were subsequently discharged, or were very seriously ill, typically older patients who subsequently died. The majority of patients were well enough to be discharged after some treatment (*i.e.*, they were in class 2, leaving after completing phase 2). Those patients who were in the later two classes (3 and 4) are the most interesting from a management point of view, since these patients offer most opportunities for rehabilitation and stay longer in hospital, thereby incurring greater costs. Older patients were less likely to become long-stay (class 4) and, not surprisingly, usually died if they did.

As Table 1 and Figure 2 show, 60 and 80 year old patients have a similar pattern, with most patients initially being expected to be in class 2. However, in each case, once the patient has been in hospital for a length of time (29 and 24 days respectively) it becomes more likely that they will be in class 3, and once the patient has been in hospital for a greater length of time (87 and 236 days respectively) it becomes more likely that they will be in class 4. The differences

Table 1. Class probabilities and most likely class as a function of length of stay to date

Age=60 Year=1976	t (days)	0	10	50	100	300	1000
	$c_1(t)$	0.25	0.14	0.01	0.00	0.00	0.00
	$c_2(t)$	0.53	0.56	0.14	0.00	0.00	0.00
	$c_3(t)$	0.15	0.21	0.52	0.46	0.07	0.00
	$c_4(t)$	0.07	0.09	0.33	0.54	0.93	1.00
	class	2	2	3	4	4	4
Age=80 Year=1976	t (days)	0	10	50	100	300	1000
	$c_1(t)$	0.25	0.14	0.01	0.00	0.00	0.00
	$c_2(t)$	0.53	0.56	0.17	0.01	0.00	0.00
	$c_3(t)$	0.21	0.29	0.77	0.88	0.27	0.00
	$c_4(t)$	0.01	0.01	0.04	0.11	0.73	1.00
	class	2	2	3	3	4	4

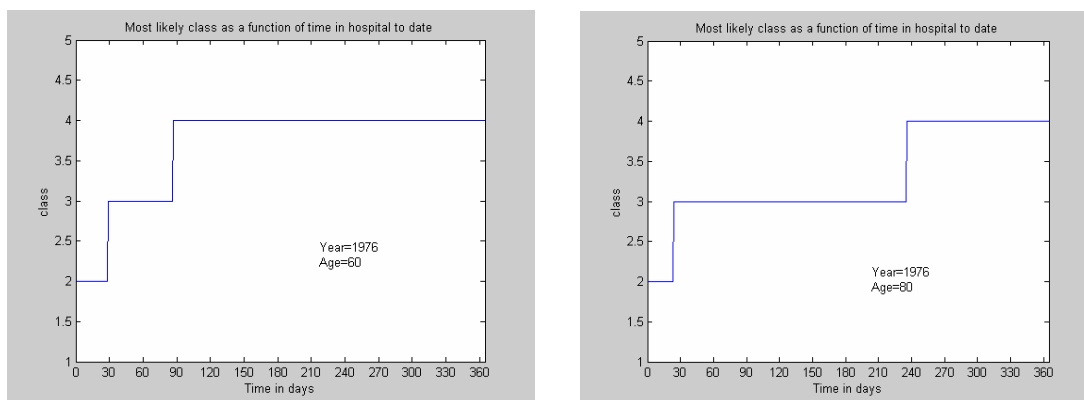


Figure 2. Most likely class for ages 60 and 80 as a function of length of stay to date

here reflect the initial expectation that a younger patient will be discharged earlier but, as their stay in hospital increases, our expectation changes to one of their being discharged from a later phase. Younger patients move through the phases faster than older patients and so we decide that they are likely to be discharged from a later phase at an earlier time.

4. Conclusions

We have described a methodology whereby patients may be assigned to a class, depending on how long they might spend in hospital and what phases of care they are likely to pass through. This assignment used covariate dependence where age and calendar year were the only ones available here, but many other covariates, such as clinical variables, dependency levels, gender, *etc.* could be used if available. The class assignment was also updated over time to take account of a patient's current length of stay in hospital.

This approach can complement the use of process improvement technologies, such as Lean Thinking or Six Sigma, within a healthcare setting⁸. Such methodologies identify patient pathways, or value streams, that can be thought of as mapping onto our clusters. A particular problem for the development of such a framework is the lack of clearly identified value streams within the healthcare system and the lack of suitable patient pathway data. This problem is exacerbated by the fact that different parts of the healthcare system often use different information systems. Such difficulties are not a

problem for traditional users of the Lean approach, such as manufacturing industry, where the value streams are clearly identified products and the corresponding data are easily obtainable. Within the healthcare context, the problem of value stream identification can be thought of as that of identifying homogeneous clusters of patients. This can be achieved using a model-based clustering approach, as described in this paper. We may thus characterise patients, in advance, as belonging to a particular cluster, or value stream, using covariates, so that we can treat these patients appropriately. Within the Lean paradigm, such streaming serves to eliminate waste by removing redundancy and other unnecessary delay. By providing a means of identifying groups of patients that will have similar lengths of stay within the appropriate healthcare domain, we may rationalise the care process thus reducing waste, in terms of unnecessary or inappropriate treatment, and avoiding delay, often the result of batch and queue processes, in a similar fashion to that adopted for industrial processes. In addition, we have provided a means of updating the cluster assignment probabilities thus facilitating re-assignment of patients between value streams as their case histories evolves. This is necessary because, unlike in the industrial analogy where the product specification is clearly identified, the appropriate value stream for a patient may only become clear after some time in care has elapsed.

Describing such heterogeneity by identifying different groups, or clusters, has already been done on an *ad hoc* basis by several authors^{9,10}. Such studies produce clear evidence that efficiency gains can be obtained by the streaming of patients into such groups. These groups can then be processed in different ways that vary according to the specific processes and procedures required. Our current approach provides a structured methodology for the identification, characterisation and assignment of patients to clusters, thus facilitating intelligent patient management using hospital administrative data to identify major patient pathways and value streams.

5. References

- [1] Millard, P.H. (1991) Throughput in a department of geriatric medicine: a problem of time, space and behaviour. *Health Trends*, **24**, 20–24.
- [2] Neuts, M.F. (1981) *Matrix Geometric Solutions in Stochastic Models*. Johns Hopkins University Press, Baltimore, Maryland.
- [3] Faddy, M.J. and McClean, S.I. (1999) Analysing data on lengths of stay of hospital patients using phase-type distributions. *Appl. Stochastic Models Bus. Ind.*, **15**, 311–317.
- [4] Faddy, M.J. and McClean, S.I. (2005) Markov chain modelling for geriatric patient care, *Methods of Information in Medicine*, in press.
- [5] McClean, S.I. and Millard, P.H. (1993) Patterns of length of stay after admission in geriatric medicine: an event history approach. *The Statistician*, **42**, 263–274.
- [6] Fraley, C. and Raftery, A.E. (2002) Model-based clustering, discriminant analysis, and density estimation, *Journal of the American Statistical Association*, **97**, 611–631.
- [7] Faddy, M.J. (2002) Penalised maximum likelihood estimation of the parameters in a Coxian phase-type distribution, in *Matrix-Analytic Methods Theory and Applications* (eds. G. Latouche and P. Taylor), World Scientific, Singapore, pp 107–114.
- [8] Young, T., Brailsford, S., Connell, C., Davies, R., Harper, P. and Klein, J.H. (2004) Using industrial processes to improve patient care, *British Medical Journal*, **328**, 162–164.
- [9] Brailsford, S.C., Lattimer, V.A., Tarnaras, P. and Turnbull, J.C. (2004) Emergency and on-demand health care: modelling a large complex system, *Journal of the Operational Research Society*, **55**, 34–42.
- [10] Lehaney, B., Clarke, S.A., and Paul, R.J. (1999) A case of an intervention in an outpatients department, *Journal of the Operational Research Society*, **50**, 877–891.