Journal of Modern Applied Statistical Methods

Volume 1 | Issue 1 Article 17

5-1-2002

Two Methods To Estimate Homogenous Markov Processes

Ricardo Ocaña-Rilola
Escuela Andaluza de Salud Pública, Granada, Spain, ricardo@easp.es

Journal Of Modern Applied Statistical Methods
Winter 2002, Vol. 1, No. 1, 131-138

1538 - 9472/02/$30.00

Two Methods To Estimate Homogenous Markov Processes

Ricardo Ocaña-Rilola
Escuela Andaluza de Salud Pública
Granada (Spain)

Multi-state Markov processes have been introduced recently in Health Sciences in order to study disease history events. This sort of model have some advantages respect to traditional survival analysis, therefore they are an important line of research into stochastic processes applied to Epidemiology. However these types of models increase the complexity of analysis, even for simpler processes, and standard software is limited. In this paper, two methods for fitting homogeneous Markov models are proposed and compared.

Keywords: Stochastic processes, Markov processes, Breast cancer data

Introduction

Statistical models such Kaplan-Meier[1], proportional hazards[2], or competing risks[3] are used in survival analysis to study the distribution of the time elapsed between two events. An example would be the diagnosis date of a disease and the date of death of the patient. More recently, stochastic processes, and in particular Markov processes,[4] have been introduced to analyze problems such as this.[5] With this sort of model it is possible to analyze jointly the evolution of the patients through different states of their disease, obtaining a complete and detailed study on disease history. They are designated multistate models or multistate stochastic processes.

Aalen[5] and Andersen and Borgan [6 7] showed that survival analysis can be treated as a particular type of two-state Markov process where the transition intensity from the transient state *live* to the absorbent *dead* is the hazard function of the variable survival time. Similarly, competing risks models can be described as a stochastic process with a transient state live and several absorbing states, corresponding to the different causes of death. The advantage of including traditional survival analysis under the framework of Markov processes resides in the possibility of studying the detailed evolution of the patients through different states or stages before death, even when the exact transition time occurs is not known.[8,9]

Markov processes have been used in to study AIDS,[10,11,12] use of contraceptives[13] and cancer.[9] These studies employed different methodologies depending on the particular conditions of each case. In practice, it is often useful to use a homogeneous Markov process in order to model disease history data because generally they are easier to interpret. Moreover, the assumption that the process is homogeneous simplifies the methods used to fit the model.

In this context, the main objective of this paper is to propose and to compare two methods, both of which are computationally tractable, to estimate homogeneous Markov models in continuous time. To illustrate the methods described in the following sections, breast cancer data from Granada (South of Spain), are considered. These data are continuous with one absorbing state.

Breast Cancer Data

Breast cancer in Granada is the most frequent cancer in woman.[14] It represents 19% of diagnosed cases of cancer, and 17% of the cancer deaths in women in the Granada province. The data originated from the Granada Cancer Registry population base, developed by the Andalusian School of Public Health and integrated in the European Network of Cancer Registries.

The study was carried out with 241 women with breast cancer diagnosed in 1985/86, who received radical treatment and had a period free of symptoms. The follow-up ended on 31 of December 1990. Approximately 5% of the cases were lost in follow-up. The 5-years Kaplan-Meier survival rate was 58% for these patients, with the 95% confidence interval between 52% and 65%.

The data are represented by a three-state Markov model with two transient states and one absorbing. These states are state 1 = "With Symptoms", state 2 = "Without Symptoms", and state 3 = "Death". The transitions are represented in Figure 1.

Each patient was observed on the date of the diagnosis, the date of the last treatment, the date of the recurrence, where the date of the recurrence was defined as "the first attendance with evidence of relapse"[15], and date of death. Thus, it is considered that at the start of follow-up all the patients are with symptoms (state 1), when the last treatment has been completed the patient is without

Ricardo Ocaña-Riola. Escuela Andaluza de Salud Pública, Campus Universitario de Cartuja, Apdo de Correos 2070, 18080 Granada, SPAIN, E-mail: ricardo@easp.es. This research was developed at the *Escuela Andaluza de Salud Pública* and financed by grant number IN92-D24255738 from the *Programa Nacional de Formación de Personal Investigador en España* of the *Ministerio de Educación y Ciencia*. The author thanks Jacques Estève for comments and suggestions.

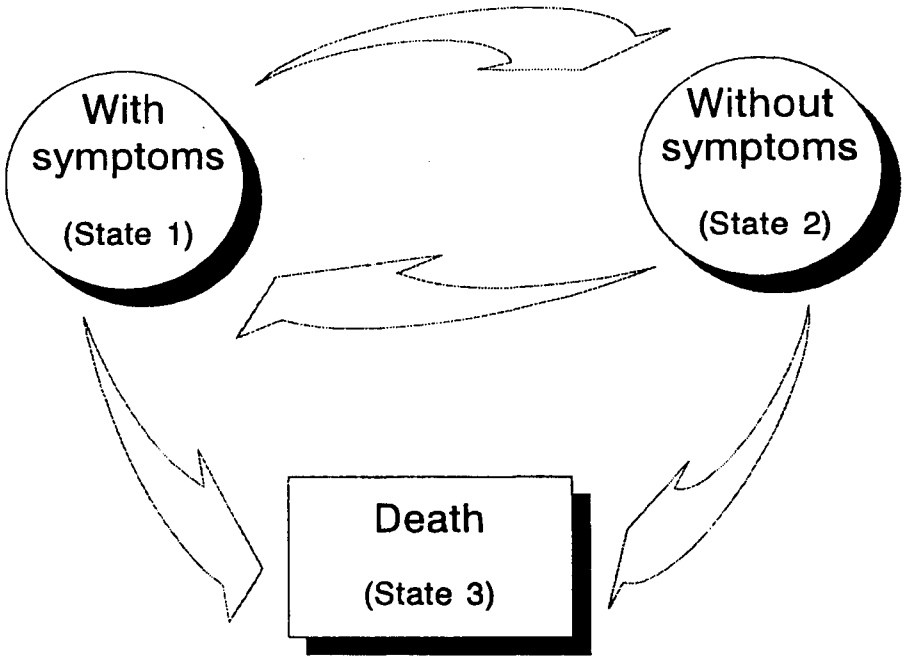

Figure 1. A three-state Markov process for analysing breast cancer

symptoms (state 2), on the date of the recurrence the patient is again considered with symptoms (state 1) and so on. Finally the patients who die are considered to be in state 3 at that date. Using these definitions, the observation is continuous (i.e., information on exact transition times between transient states is available).

Probabilities and Transition Intensities in a Markov Process

Let $E = \{1,...,m\}$ be a state space consisting of *m* disease states in which patients make independent transitions. Let *X(t)* be the state occupied by an individual at time *t* for $0 \le t$ and $\{X(t): t \ge 0\}$ a collection of random variables that define a stochastic process. In this context, if the process is Markov, we denote $p_{ij}(s,t)$ the additional probability that the process shall be in state *j* at time *t* given that it was at time *s* in state *i*, for $0 \le s \le t$.

Homogeneous processes are the simplest form of Markov models. In this case, assuming a stationary process, the transition probabilities satisfy

$p_{ij}(s, s+u) = p_{ij}(0,u) = p_{ij}(u)$ for $u, s \ge 0$, with

$P(u) = (p_{ij}(u))$ the *mxm* transition probability matrix.

The transition intensities for the process, q_{ij}, are defined by

$$p_{ij}(h) = q_{ij}h + o(h) \qquad i \ne j \;\; i,j \in E \qquad (1)$$

$$1 - p_{ii}(h) = q_i h + o(h) \qquad i \in E \qquad (2)$$

where $o(h)$ is an infinitesimal that satisfies $\frac{o(h)}{h} \to 0$ when $h \to 0$ and $Q = (q_{ij})$ is the $m \times m$ transition intensity matrix satisfying $q_i = -q_{ii} = \sum_{j \ne i} q_{ij}$ for $i, j \in E$.

The relationship between $P(u)$ and Q is

$$P(u) = \exp(Qu) = \sum_{S=0}^{\infty} Q^S \frac{u^S}{S!} \qquad (3)$$

Let T_i be the random variable "time elapsed in the state $i \in E$". It is known that this variable is distributed according to an exponential law with parameter q_i and average $E[T_i] = \frac{1}{q_i}$.

Likelihood Function

Suppose that for each subject c in a cohort *(c=1,*

..., *N)*, m_c transitions between states has been observed at times

$$w_{c0} < w_{c1} < ... < w_{cm_c}$$

If m_c transitions has been observed for each subject *(c=1,* ..., *N)*, then there is a set of $N(m_c+1)$ times in which there was a jump from a state to another. These times can be ordered to form a partition of the time given by $N(m_c+1)$ cut points in which at least a transition between states has been observed. Let $\wp$ be such a partition with $\wp = \{s_1,...,s_{N(m_c+1)}\}$. To simplify the notation we denote $r = N(m_c+1)-1$, so $\wp = \{s_1,...,s_{r+1}\}$.

There are $r + 1$ cut points that form r intervals with length $u_k = S_{k+1} - S_k$, $k = 1,...,r$. If n_{ijk} denotes the number of patients who were in the state i in the time S_k and are in the state j at the instant S_{k+1}, the likelihood function is

$$L = \prod_{k=1}^{r}\left(\prod_{i=1}^{m}\prod_{j=1}^{m} p_{ij}(u_k)^{n_{ijk}}\right) \quad (4)$$

Two Methods to Estimate Intensity Matrix

This section shows how matrix Q can be estimated when homogeneous Markov processes in continuous time are used. Assume that transition intensities depend on a beta vector of unknown parameters with components $\beta_1,...,\beta_b$. That is, $q_{ij} = q_{ij}(\beta)$, $i, j \in E$, but to simplify the notation write $p_{ij}(u)$ and q_{ij} instead of $p_{ij}(u,\beta)$ and $q_{ij}(\beta)$.

Kalbfleisch and Lawless Method

By using $\wp$ partition, the likelihood function in (4) can be seen as an extension of the methodology of Kalbfleisch and Lawless[8] when observations are continuous and transitions between states are observed at different times. In order to obtain the Maximum Likelihood Estimate (MLE) for β, the first and second partial derivatives of *ln(L)* are required. The Kalfleisch and Lawless method uses a quasi-Newton algorithm for this purpose in which it is necessary to compute complex calculations for every interval of the partition. (See Appendix.)

Approximate Method

To estimate matrix Q, an algorithm is proposed that consists of modeling transition probabilities with a simple function that depends on the transition intensities. From expressions (1) and (2), in a homogeneous Markov process the transition probabilities can be approximated by $p_{ij}(h) = q_{ij}h$ and $p_{ii}(h) = 1 - q_i h$ when $h \to 0$. If there is a sufficiently fine $\wp$ partition composed of very small intervals, u_k, whose lengths tend to be zero, the approximation

$$p_{ij}(u_k) = q_{ij}u_k \quad i, j \in E \quad (5)$$

$$p_{ii}(u_k) = 1 - q_i u_k \quad k \in E \quad (6)$$

can be used for all $k=1,...,r$. Usually, $q_{ij} = \exp(\beta_{ij})$ can be taken in order to avoid problems in the range of the estimates upon implementing the algorithm in a computer. The likelihood function given in (4) simplifies to

$$L = \prod_{k=1}^{r}\left(\prod_{i=1}^{m}\left(1 - u_k\sum_{v\neq i}^{m}\exp(\beta_{iv})\right)^{n_{iik}}\prod_{j\neq i}^{m}\left(u_k\exp(\beta_{ij})\right)^{n_{ijk}}\right)$$

using this approximation.

The MLE for β can be obtained using traditional partial derivatives of *ln(L)*. In this case, these partial derivatives are calculated easily. The algorithm for fitting the model is iterative and it will asymptotically approach the best fitting values of the parameters. The convergence criterion, $\in$, is the maximum relative change in all the parameters. The estimated transition probability matrix is $\hat{p}(u) = \exp(\hat{Q}u)$ where $\hat{Q} = (\hat{q}_{ij}) = (\exp(\hat{\beta}_{ij}))$.

Selection of an Initial Estimate

As mentioned above, the algorithms for fitting the model are iterative, and will asymptotically approach the best solution. However, a serious problem in any algorithm is the selection of an appropriate initial value $\beta^{(0)}$. The calculations used to obtain an estimate of β are complex, particularly in the Kalbfleisch and Lawless method, and the time invested in each iteration is high. A poor choice of the initial value will increase enormously the time necessary to obtain a final estimate and even could cause the algorithm does not to converge.

An initial estimate of the transition intensities based on the observed data can be obtained as follows: From expression (1), in each interval u_k it is obtained an approximation of the form

$$q_{ij}(u_k) = \frac{p_{ij}(u_k)}{u_k} \quad i \neq j,\ i, j \in E$$

If the value $p_{ij}(u_k)$ is estimated $\frac{n_{ijk}}{n_{i.k}}$, an initial estimate will be obtained for the transition intensity in each one of the intervals given by

$$q_{ij}(u_k) = \frac{n_{ijk}}{n_{i.k}u_k} \quad i \neq j,\ i, j \in E \quad (6)$$

where $k = 1,...,r$. Because the transition intensities are independent of the time, consider the means

$$q_{ij}^{(0)} = \frac{\sum_{k=1}^{r} q_{ij}(\upsilon_k)}{r} \qquad i \neq j$$

and the initial vector $\beta^{(0)} = \left(\log\left(q_{ij}^{(0)}\right)\right)_{i,j}, \quad i \neq j.$

Figure 2 shows this method based on the breast cancer data mentioned above, where the dashed line represents the initial values, $q_{12}^{(0)}$, $q_{13}^{(0)}$, $q_{21}^{(0)}$, $q_{23}^{(0)}$, given respectively by 0.37, 0.24, 0.10, 0.02. The circles represent the values given by (6), where υ_k units are years.

Two Methods for Breast Cancer Analysis

To analyze the breast cancer data, consider the transition intensity matrix

$$Q = \begin{bmatrix} -(q_{12}+q_{13}) & q_{12} & q_{13} \\ q_{21} & -(q_{21}+q_{23}) & q_{23} \\ 0 & 0 & 0 \end{bmatrix}$$

The transition 2-3 corresponds to death from other causes which can not be controlled. Thus, the entry q_{23} is included in the intensity matrix.

After building the $\wp$ partition, 35 different intervals were found between 0.002 and 0.260 years. The mean was 0.071 years. The 25th, 50th, and 75th, percentiles were 0.026, 0.049 and 0.075, respectively. These results show a sufficiently fine partition where most intervals (75%) extended less than 0.075 years.

Table 1 shows a comparative analysis of the Kalbfleisch-Lawless vs the approximate method. The estimates obtained are very similar. The criterion of convergence was $\epsilon = 0.00001$ and the implementation in a computer was accomplished through functions written for *S-PLUS*.[16]

Figure 3 shows the difference between estimated transition probability matrices from each method for each interval of length u. These differences have been calculated using $Error = \left\|\hat{P}_I(u) - \hat{P}_{II}(u)\right\|_2$, where $\hat{P}_I(u)$ and $\hat{P}_{II}(u)$ are the estimated transition probability matrices from Kalbfleisch-Lawless and Approximate method respectively.

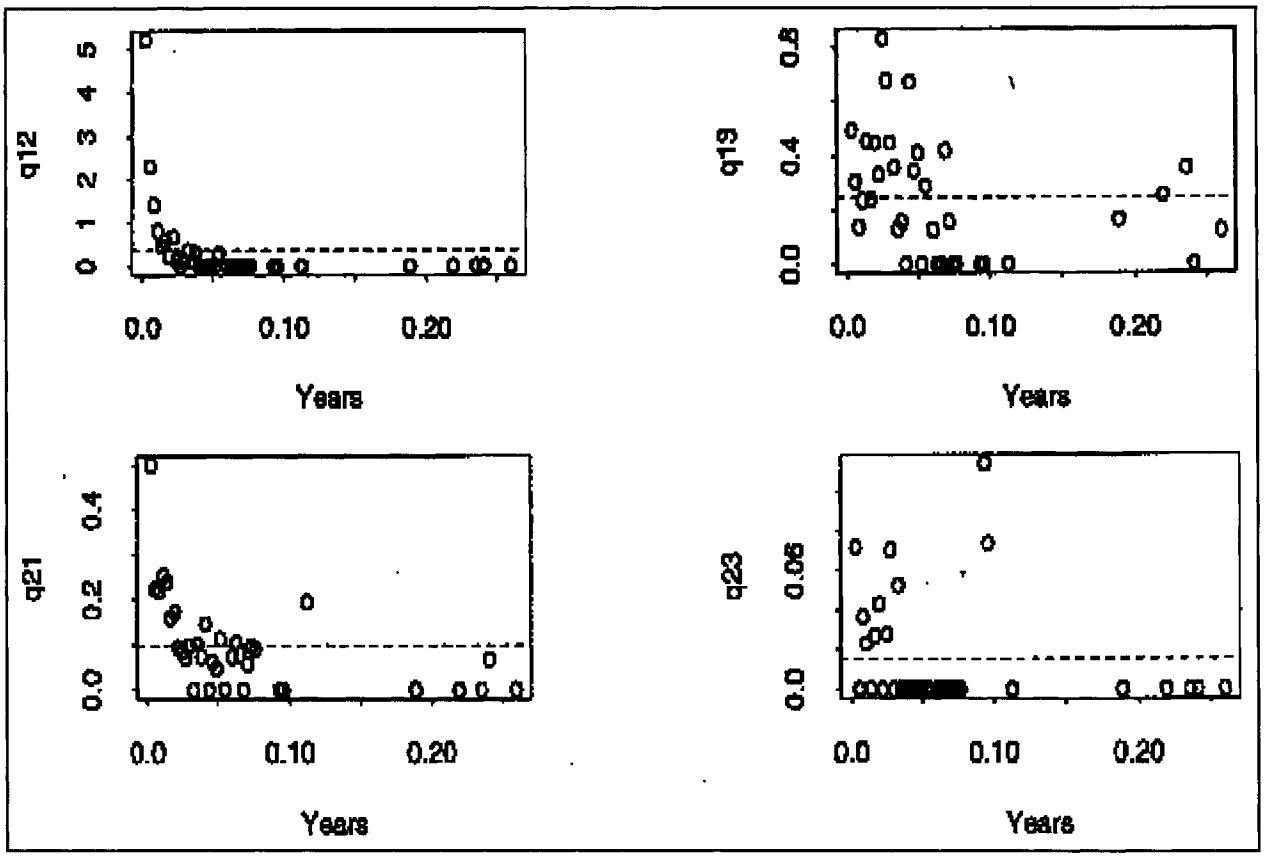

Figure 2. Initial estimation for transition intensities. Dashed lines is the mean of all transition intensities, used as initial value in iterative methods

Table 1. Homogeneous Markov Models for Breast Cancer Data.

	Kalbfleisch-Lawless		Approximate	
Transition intensities	Estimate	Standard error	Estimate	Standard error
q_{12}	1.1653	0.0049	1.0738	0.0044
q_{13}	0.2929	0.0024	0.2729	0.0023
q_{21}	0.1235	0.0008	0.1175	0.0008
q_{23}	0.0150	0.0003	0.0153	0.0003

If in the breast cancer study, the approximate algorithm is used instead of the method of Kalbfleisch and Lawless, the maximum error is 0.036. The estimated intensity matrix is similar by using either algorithm. However the approximate method is less costly computationally.

Mean sojourn times in each state i = 1,2 were obtained from $1/\hat{q}_i$, that is, 0.69 years (8.2 months) and 7.22 years, respectively, for the mean sojourn time with and without symptoms.

Figures 4 shows the estimated transition probabilities between states, obtained from $\hat{P}(u) = exp(\hat{Q}u)$. The 5-years probability of death from with symptoms and without symptoms state was 0.30 and 0.15 respectively. The 5-years survival probability was 0.55. Note that the 5-years Kaplan-Meier survival rate was 58% for these patients.

Conclusion

Multistate Markov models offer some advantages on traditional survival models for studying disease history events. Using this sort of model, it is possible to estimate the proportions of patients who shall be in each disease state in the future. Therefore, highly relevant information for health planning services can be obtained.

In this article, the partition of time used was for

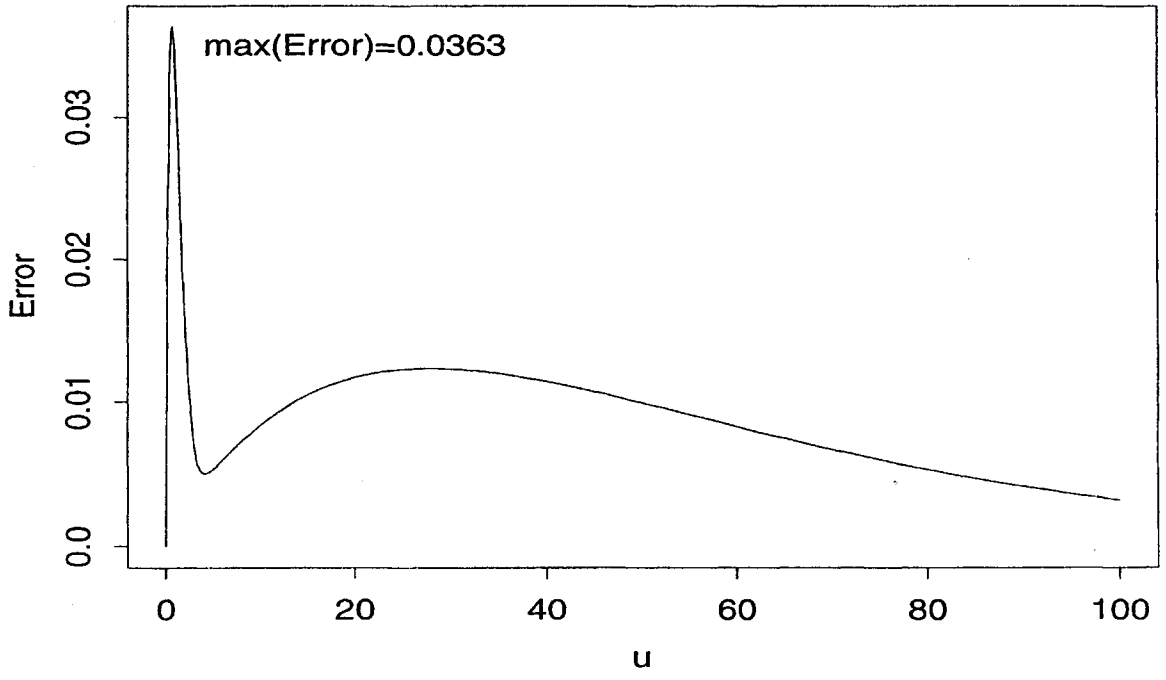

Figure 3. Differences among the transiton probability matrices obtained from Kalbfleisch-Lawless algorithm and Approximation method. These differences have been calculated using $Error = \left\| \hat{P}_I(u) - \hat{P}_{II}(u) \right\|_2$, $0<u<100$

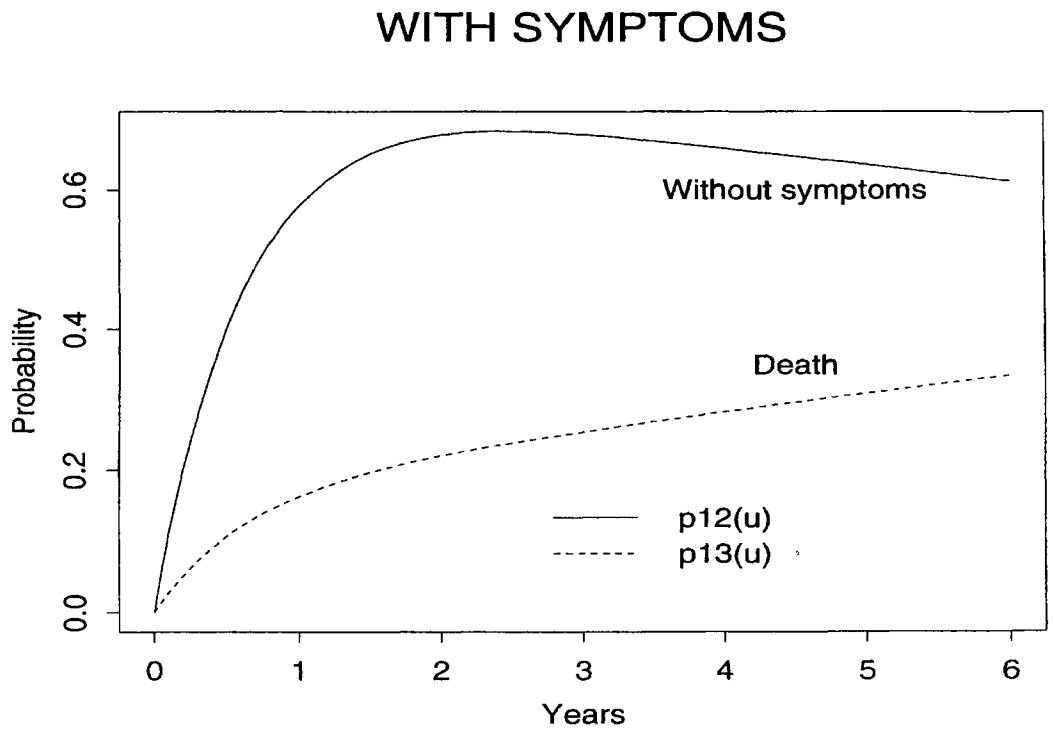

Figure 4. Estimated transition probabilities from *with symptoms* state to *without symptoms state* and *death* states

continuous observations. In this situation, the methodology proposed by Kalbfleisch and Lawless has been extended, and an approximate method was proposed in order to estimate intensity matrix. If the intervals in which patients are observed are sufficiently small, the Kalbfleisch-Lawless algorithm can be simplified using the approximate method, obtaining similar MLEs.

In survival studies, the use of models that incorporate covariates permits further analysis of patients' survival. When multistate models are used, it is also possible to study the effect of covariates on different transitions between states in patients disease history. Some authors have worked on the introduction of covariates in multistate processes and particularly in homogeneous Markov processes.[8,17] However they mention the increased complexity of analysis in this sort of model, where an additional problem is the shortage of standard software.

In spite of these problems, the introduction of covariates in stochastic processes are required to explain the effect of these factors on disease history events. In this case, a homogeneous Markov model with covariates would be an interesting option. An advantage of the approximate method proposed in this paper is the ease with which covariates can be incorporated in the likelihood function, therefore the study of prognostic factors is not difficult. Interesting results for breast cancer survival and breast cancer recurrence using the approximate method with TNM variables have been found.[18]

Computing approximate MLEs replaces the intensity transition matrix with a first order Taylor series. In this article, an empirical comparison showed that, in some cases, it can be a good approximation. Similar conclusions in other applications, so the question of when the Taylor expansion is liable to produce accurate approximations and on developing a diagnosis for examining the accuracy of the approximation is of interest. The results obtained, therefore, could be an interesting finding in applying Markov processes to Health Sciences and Epidemiology.

References

1. Kaplan, E. L., & Meier, P. (1958). Non parametric estimation from incomplete observations. *J. Am. Stat. Assoc*, *53*, 457-481.
2. Cox, D. R. (1972). Regression models and lifetime (with discussion). *J. Roy. Stat. Soc. B. Met.*, *34*, 187-220.
3. Prentice, R. L., Kalbfleisch, J. D., Peterson Jr., A. V., Flournoy, N. S., Farewell, V. T., & Breslow, N. (1978). The analysis of failure times in the presence of competing risks. *Biometrics*, 34, 541-554.

4. Chiang, C. L. (1968). *An introduction to stochastic processes and their applications*. New York: John Wiley & Sons.
5. Aalen, O. O. (1978). Non-parametric estimation of partial transition probabilities in multiple decrement models. *Ann. Stat., 6,* 534-545.
6. Andersen, P. K. (1988). Multistate models in survival analysis: A study of nephropathy and mortality in diabetes. *Stat. Med., 7,* 661-670.
7. Andersen, P. K, & Borgan, O. (1985). Counting process models for life history date: A review (with discussion). *Scand. J. Stat., 12,* 97-158.
8. Kalbfleisch, J. D., & Lawless, J. F. (1985). The analysis of panel data under a Markov assumption. *J. Am. Stat. Assoc., 80,* 863-871.
9. Kay, R. A. (1986). Markov model for analyzing cancer markers and disease states in survival studies. *Biometrics, 42,* 855-865.
10. De Gruttola, V., & Lagakos, S. W. (1989). Analysis of doubly-censored survival date, with application to AIDS. *Biometrics, 1989, 45,* 1-11.
11. Frydman, H. A. (1992). Nonparametric estimation procedure for a periodically observed three-state Markov process, with application to AIDS. *J. Roy. Stat. Soc. B Met., 54,* 853-866.
12. Mariotto, A. B., Mariotti, S., Pezzotti, P., Rezza, G., & Verdecchia, A. (1992). Estimation of the Acquired Immunodeficiency Syndrome incubation period in intravenous drug users: a comparison with male homosexual. *Am. J. Epidemiol., 135,* 428-437.
13. Islam, M. A. (1994). Multistate survival models for transitions and reverse transitions: an application to contraceptive use date. *J. Roy. Stat. Soc. A Sta., 157,* 441-455.
14. Martínez, C. (1994). *El cáncer en Granada: incidencia y mortalidad 1988-1990*. Granada: EASP.
15. WHO. (1980). *WHO handbook for reporting results of cancer treatment*, WHO offset publication No. 48., Geneva.
16. *S-Plus*. (1998). *S-PLUS 4 programmer's guide*. Seattle: MathSoft, Inc.
17. Pastorello, S. (1993). La mobilità nel mercato del lavoro: un analisi econometrica con osservacioni in tempo discreto. *Statistica, 53,* 185-206.
18. Ocaña-Riola, R., Sánchez-Cantalejo, E., & Martínez-García, C. (2000). Homogeneous Markov processes for breast cancer analysis. Unpublished Technical Report.

Appendix

Summary of the Kalbfleisch and Lawless Method

The Newton-Raphson method is usually applied in order to obtain MLE for β. In this case, first and second partial derivatives of ln(L) with respect to each unknown parameter $\beta_h (h = 1,...,b)$ are required, that is

$$S_h(\beta) = \frac{\partial ln(L)}{\partial \beta_h} = \sum_{k=1}^{r}\sum_{i=1}^{m}\sum_{j=1}^{m} n_{ijk} \frac{\frac{\partial p_{ij}(u_k)}{\partial \beta_h}}{p_{ij}(u_k)}$$

$$J_{hr}(\beta) = \frac{\partial^2 ln(L)}{\partial \beta_h \partial \beta_r} = \sum_{k=1}^{r}\sum_{i=1}^{m}\sum_{j=1}^{m} n_{ijk} \left(\frac{\frac{\partial^2 p_{ij}(u_k)}{\partial \beta_h \partial \beta_r}}{p_{ij}(u_k)} - \frac{\frac{\partial p_{ij}(u_k)}{\partial \beta_h}\frac{\partial p_{ij}(u_k)}{\partial \beta_r}}{p_{ij}(u_k)^2} \right)$$

We obtain then a vector $S(\beta) = (S_h(\beta))_{b\times 1}$ of dimension b and a $b \times b$ squared matrix $J(\beta) = (J_{hr}(\beta))_{b\times b}$. However this method can be simplified using a quasi-Newton method that employs, instead of $J_{hr}(\beta)$, the information matrix $M(\beta) = E[-J(\beta)]$ with elements

$$M_{hr}(\beta) = \sum_{k=1}^{r}\sum_{i=1}^{m}\sum_{j=1}^{m} \frac{\partial p_{ij}(u_k)}{\partial \beta_h}\frac{\partial p_{ij}(u_k)}{\partial \beta_r}\frac{n_{i.k}}{p_{ij}(u_k)}$$

where $n_{i.k} = \sum_{j \in E} n_{ijk}$.

If $\beta^{(0)}$ is an initial estimate of β we can built the sequence

$$\beta^{(n+1)} = \beta^{(n)} + M\left(\beta^{(n)}\right)^{-1} S\left(\beta^{(n)}\right)$$

which converges to the MLE for β, assuming that $M\left(\beta^{(n)}\right)$ is nonsingular in each iteration.