Symbolic Markov Chains

Fraiture, Monique Noirhomme; Cuvelier, Etienne

Published in:
Selected Contributions in Data Analysis and Classification

Publication date:
2007

Document Version
Early version, also known as pre-print

Link to publication

Citation for pulished version (HARVARD):
Fraiture, MN & Cuvelier, E 2007, Symbolic Markov Chains. in P Brito, P Bertr, G Cucumel & F de (eds), *Selected Contributions in Data Analysis and Classification.* pp. 103-111.

Symbolic Markov Chains

Monique Noirhomme-Fraiture[1] and Etienne Cuvelier[2]

[1] Institut d'Informatique, Université de Namur
rue Grandgagnage, 21, B-5000 Namur, Belgium
mno@info.fundp.ac.be
[2] *ecu@info.fundp.ac.be*

Abstract. Stochastic processes have, since a long time, large applications in quite different domains. The standard theory considers discrete or continuous state space. We consider here the concept of Stochastic Process associated to all the cases of symbolic variables: quantitative, categorical single and multiple, interval, modal. More particularly, we adapt the definition of Markov Chain and give the equivalent of the Chapman-Kolmogorov theorem in all cases.

1 Introduction

Frequently, we have to consider systems which develop in time or space in accordance with probabilistic laws. The study of such systems is called the theory of Stochastic Processes. More precisely, a Stochastic Process is a random variable which depends on time or space.

The aim of this paper is to propose theoretical bases for generalisation of Stochastic Processes to Symbolic variables.

Indeed, Stochastic Processes are defined for variables for which the state space (or values) is a countable or finite set or the real line $(-\infty, \infty)$. In the first case, the process is called a Chain. Here, we want to extend this concept to variables which can be multivalued, interval or even modal.

This problem is practically meaningful. For example, let us consider the evolution of the value of stock. Usually, each day, the stock has several values: open, close, mean, maximum, minimum. The stock value can thus be characterised by an interval of values and not by a unique number.

If we consider daily audience of a TV channel, the audience for a family is given by the percentage of time spent at watching different broadcasts and not by a single category. In this case, the variable is modal.

This paper does not deal with the statistical analysis of symbolic data from Stochastic Processes. We try only to modelise the problem from a probabilistic point of view.

We will concentrate our study to a special case of Stochastic Process which is Markov Chains. We will first recall the definition and principal characteristics of Markov Chain in the case of categorical and continuous variables and we will extend them to the case of multivalued categorical, interval and modal variables.

To simplify the presentation, we will speak only about time and not space. This choice is motivated by the fact that it concerns the more frequent applications.

We have also chosen to present here only the case of discrete time. But continuous time could also be considered, in an extended paper.

Numerous books have been written about Stochastic Processes. From others, let us quote Cox and Miller (1965), Bartlett (1978), Prabhu (1965), Karlin (1966), Neveu (1964), Feller (1968), Bailey (1964) and more recently, Stierzaker (2005), Lawler (2006), Beichelt (2006), Meyn & Tweedie (1993), Girkhman and Skorokhod (2004). On the other hand, in Symbolic Data Analysis, very few has been done in Stochastic Processes. Diday et al. (2004) and De Carvalho et al. (2004) have studied linear symbolic regression. Prudencio et al. (2004) have considered time series. We can also mention the work of Soule et al. (2004) in flow classification.

2 Definitions

Let us consider $(\Omega, \mathcal{A}, Pr)$ a probability space and $\{\underline{X}_t, t \in T\}$ a Stochastic Process defined on this space, i.e. a random variable depending upon the parameter t, considered as the time.

We will consider the particular case where the time is discrete, with values represented by the positive integers. In this case, the Stochastic Process is often written $\{\underline{X}_n, n \in \mathbb{N}\}$.

The set of values of $\underline{X}_t$ is the state space. In the standard theory, it can be continuous or discrete. The study of a Stochastic Process is very complex except if we make hypothesis on the behavior of the process.

One common hypothesis is the Markovian one. A Markov process is a process with the property that, given the value $\underline{X}_t$, the values of $\underline{X}_s$, $s > t$, do not depend on the values of $\underline{X}_u$, $u < t$.

In formal terms, a process is said to be Markovian if

$$\begin{aligned} Pr[a < \underline{X}_t \le b \mid \underline{X}_{t_1} = x_1, \underline{X}_{t_2} = x_2, \dots, \underline{X}_{t_n} = x_n] \\ = Pr[a < \underline{X}_t \le b \mid \underline{X}_{t_n} = x_n] \end{aligned}$$

whenever $t_1 < t_2 < \cdots < t_n < t$.

The function

$$Pr[\underline{X}_t \in A \mid \underline{X}_s = x] \,, \qquad t > s$$

is called the transition probability function and is basic to the study of the structure of Markov processes.

A Markov process is said to have **stationary transition probabilities** if the transition probabilities are function only of $t - s$ and not s. We say also “homogeneous in time”. A Stochastic Process $\underline{X}_t$ for t in T is said to be **stationary** if the joint distribution function of the families of random variables $(X_{t_1+h}, X_{t_2+h}, \dots, X_{t_n+h})$ and $(X_{t_1}, X_{t_2}, \dots, X_{t_n})$ are the same for all $h > 0$ and arbitrary selections $t_1, t_2, \ldots, t_n$ of T. This property means that the particular times at which we examine the process are of no relevance. In particular, the distribution of $\underline{X}_t$ is the same for each t. Let us note that there is no reason to expect that a Markov process with stationary probabilities is a stationary process (Karlin (1966), p 204).

3 Single valued categorical variables

Let us consider the case where $\underline{X}_n$ means belonging to one category among s at time n.

We can modelise this case in writing $\underline{X}_n = k$, $1 \leq k \leq s$. The process $\{\underline{X}_n, n \in T\}$ is thus a classical Stochastic Process whose state space is the finite set $(1, 2, \ldots, s)$. We will suppose that the process is Markovian, with stationary transition probabilities.

The stationary transition probabilities are defined by:

$$P_{ij}(n) = Pr[\underline{X}_{m+n} = j \mid \underline{X}_m = i] \ .$$

For such probabilities, it can be shown easily the Chapman-Kolmogorov property:

$$P_{ij}(m+n) = \sum_k P_{ik}(n) \ P_{kj}(m) \ , \qquad \forall \ i, j$$

or

$$P(m+n) = P(m) \ P(n)$$

with $P(n)$ the matrix with element $P_{ij}(n)$ and

$$P \equiv P(1) \ .$$

¿From this, we have

$$P(n) = P^n$$

which allows the computation of the matrix $P(n)$ when n is small.

With some properties on transition probabilities, it is possible to show that the Markov Chain is stationary and to compute easily $\lim_{n \to \infty} P(n)$ (Cox and Miller (1965)), (Prabhu (1965)).

4 Multivalued categorical variables

In this case, the variable $\overrightarrow{X_t}$ indicates belonging to several categories, among s $(C_1, \ldots, C_s)$.

We can modelise this case in considering the multidimensional variable $\underline{\overrightarrow{X_t}}$ with state $\overrightarrow{\jmath} = (j_1, \ldots, j_s)$ where

$$j_k = \begin{cases} 1 & \text{if the category } C_k \text{ is present,} \\ 0 & \text{elsewhere.} \end{cases}$$

$\overrightarrow{\underline{X}_t}$ is here a s-vector process.

Such a process is a Markov process if

$$Pr[\overrightarrow{\underline{X}_{t_{n+1}}} = \overrightarrow{a_{n+1}} \mid \overrightarrow{\underline{X}_{t_1}} = \overrightarrow{a_1}, \ldots, \overrightarrow{\underline{X}_{t_n}} = \overrightarrow{a_n}] = Pr[\overrightarrow{X}_{t_{n+1}} = \overrightarrow{a_{n+1}} \mid \overrightarrow{\underline{X}_{t_n}} = \overrightarrow{a_n}]$$
$$for\ all\ t_1 < t_2 < \cdots < t_n < t_{n+1}\ .$$

If we suppose that the transition probabilities are stationary, let us define:

$$P_{\overrightarrow{\imath}\ \overrightarrow{\jmath}}(n) = Pr[\overrightarrow{\underline{X}_{t+n}} = \overrightarrow{\jmath} \mid \underline{X}_t = \overrightarrow{\imath}]\ .$$

The Chapman-Kolmogorov property is still valid :

$$P_{\overrightarrow{\imath}\ \overrightarrow{\jmath}}(n+m) = \sum_{\overrightarrow{k}} P_{\overrightarrow{\imath}\ \overrightarrow{k}}(n)\ P_{\overrightarrow{k}\ \overrightarrow{\jmath}}(m)$$

which allows to compute $P_{\overrightarrow{\imath}\ \overrightarrow{\jmath}}(n)$ from $P_{\overrightarrow{\imath}\ \overrightarrow{\jmath}}(1)$.

5 Single quantitative variable

In this case, the state space of the Markov Chain $\underline{X}_n$ is $(-\infty, +\infty)$. As previously, we restrict to chains with stationary transition probabilities.

$$P_n(x;y) = Pr[\underline{X}_{m+n} \leq y \mid \underline{X}_m = x] \tag{1}$$

defines the n-th order transition distribution function.

In particular, let

$$P_1(x;y) \equiv P(x;y) = Pr[\underline{X}_{m+1} \leq y \mid \underline{X}_m = x]\ .$$

The Chapman-Kolmogorov equation can be written:

$$P_{m+n}(x;y) = \int_{-\infty}^{+\infty} d_z\ P[X_m \leq z \mid X_0 = x]\ Pr[X_{m+n} \leq y \mid X_m = z] \tag{2}$$

or

$$P_{m+n}(x;y) = \int_{-\infty}^{+\infty} d_z\ P_m(x;z)\ P_n(z;y)\ . \tag{3}$$

If $p_m(x;y)$ denotes the probability densities, if they exist, this relation can be written

$$P_{m+n}(x;y) = \int_{-\infty}^{y} p_{m+n}(x;u)\ du = \int_{-\infty}^{+\infty} p_m(x;z) \int_{-\infty}^{y} p_n(z;u)\ du\ dz$$

and thus, it can be proven that (Cox and Miller (1965), p 134) :

$$p_{m+n}(x;u) = \int_{-\infty}^{+\infty} p_m(x;z)\ p_n(z;u)\ dz\ . \tag{4}$$

A Markov process is specified by giving the initial distribution and transition probabilities $P(x;y)$.

The use of Kolmogorov equation gives all the other transition probabilities $P_n(x;y)$ and the state distributions.

An alternative approach is given by the use of Copulas (Nelsen (1999)). A Copula function C is a multivariate uniform distribution (a multivariate distribution with uniform margins).

It can be shown, from Sklar's theorem, that if F is a N-dimensional distribution function with continuous margins $F_1, \ldots, F_N$, then F has a unique Copula representation

$$F(x_1, \ldots, x_N) = C(F_1(x_1), \ldots, F_N(x_N)) \ .$$

The product of Copulas is defined by

$$C_1 \star C_2(u,v) = \int_0^1 \tfrac{\partial}{\partial v}\ C_1(u,z)\ \tfrac{\partial}{\partial u}\ C_2(z,v)\ dz \ .$$

Darsow et al. (1992) prove that if $\underline{X}_t$ is a Markov process and let $C_{m,n}$ denote the Copula of the random variables X_m and X_n, then the Chapman-Kolmogorov equation is equivalent to

$$C_{t,t+m+n} = C_{t,t+m} \star C_{t+m,t+m+n} \tag{5}$$

where $\star$ denotes the product of Copulas.

With this approach, a Markov process is specified by giving all the marginal distributions and a family of 2-Copulas satisfying (5) (Joe (1997)).

6 Interval Symbolic variable

Let us suppose that, at each time, the variable is known only by its belonging to an interval of the real line.

It means that we are here interested by the transition probabilities

$$Pr[a_2 \le \underline{X}_{m+n} \le b_2 \mid a_1 \le \underline{X}_m \le b_1]$$

which we will write

$$Pr[\underline{X}_{m+n} \in A_2 \mid \underline{X}_m \in A_1] = P_n(A_1;A_2) \tag{6}$$

if A_1 and A_2 are intervals of $]-\infty,+\infty[$ and if this probability does not depend on m.

We will define an Interval Markov Chain, a chain such that

$$\begin{aligned} Pr[\underline{X}_{t_{n+1}} \in A_{n+1} \mid \underline{X}_{t_1} \in A_1, \ldots, \underline{X}_{t_n} \in A_n] \\ = Pr[\underline{X}_{t_{n+1}} \in A_{n+1} \mid \underline{X}_{t_n} \in A_n] \end{aligned} \tag{7}$$

where $A_j = [a_j, b_j]$.

Let us note that we have a particular case

$$Pr[\underline{X}_{t_{n+1}} \in A_{n+1} \mid \underline{X}_{t_1} = a_1, \ldots, \underline{X}_{t_n} = a_n] = Pr[\underline{X}_{t_{n+1}} \in A_{n+1} \mid \underline{X}_{t_n} = a_n]$$

when $a_j = b_j$.

Let $A_1 = [a_1, b_1]$ and

$$P_m(A_1; z) = Pr[\underline{X}_{t+m} \leq z \mid \underline{X}_t \in A_1]$$

which is supposed not depending on t.

Then,

$$P_m(A_1; A_2) = P_m(A_1; b_2) - P_m(A_1; a_2)$$

for an interval $A_2 = [a_2, b_2]$ and continuous P_m function.

If the derivative of $P_m(A_1; z)$ exists, we will note

$$p_m(A_1; u) = \frac{\partial}{\partial u} \; P_m(A_1; u) \; .$$

Theorem: *For an Interval Markov Chain with stationary transition probabilities, we have the relation*

$$P_{m+n}(A_1; A_2) = \int_{-\infty}^{\infty} d_z \; P_m(A_1; z) \; P_n(z; A_2) \tag{8}$$

and, if the probability density exists,

$$p_{m+n}(A_1; u) = \int_{-\infty}^{+\infty} p_m(A_1; z) \; p_n(z; u) \; dz \; . \tag{9}$$

Proof. From conditional probability property, we know that

$$\begin{aligned} &Pr[\underline{X}_{t+m+n} \in A_2 \mid \underline{X}_t \in A_1] \\ &= \int_{-\infty}^{\infty} d_z \, Pr[\underline{X}_{t+m} \leq z \mid \underline{X}_t \in A_1] \, Pr[\underline{X}_{t+m+n} \in A_2 \mid \underline{X}_t \in A_1, \underline{X}_{t+m} = z] \; . \end{aligned}$$

Using the Markovian property (7) we have

$$\begin{aligned} &Pr[\underline{X}_{t+m+n} \in A_2 \mid \underline{X}_t \in A_1] \\ &\qquad = \int_{-\infty}^{\infty} d_z \; Pr[\underline{X}_{t+m} \leq z \mid \underline{X}_t \in A_1] \;\; \mathrm{Pr}[\underline{X}_{t+m+n} \in A_2 \mid \underline{X}_{t+m} = z] \end{aligned}$$

and using the fact that the transition probabilities do not depend on time and notations (6)

$$P_{m+n}(A_1; A_2) = \int_{-\infty}^{+\infty} d_z \; P_m(A_1; z) \; P_n(z; A_2) \; .$$

If the densities $p_m(A_1;u)$ and $p_n(z;u)$ exist, then

$$P_{m+n}(A_1;A_2) = \int_{a_2}^{b_2} p_{m+n}(A_1;u)\;du = \int_{-\infty}^{+\infty} p_m(A_1;z)\;\int_{a_2}^{b_2} p_n(z;u)\;du\;dz\;.$$

Thus

$$p_{m+n}(A_1;u) = \int_{-\infty}^{+\infty} p_m(A_1;z)\;p_n(z;u)\;dz\;. \qquad \blacksquare$$

Remark: It is possible to modelise an interval by two values : its center and its half-length. In this case, $\underline{X}_t$ is in fact a two dimensions continuous variable.

7 Modal variable

A Modal variable is known by the belonging probability to classes $C_1, \ldots, C_s$ (Bock and Diday (2000)).

For a Modal Stochastic Process, it means that the variable $\overrightarrow{\underline{X}_t}$ is defined by $\Pi_1(t), \ldots, \Pi_s(t)$ with

$$\Pi_1(t) + \cdots + \Pi_s(t) = 1\;, \qquad 0 \le \Pi_j(t) \le 1\;, \quad \forall\; j\;.$$

$\overrightarrow{\underline{X}_t}$ is thus in fact a multidimensional continuous process whose value will be written $\overrightarrow{\Pi}_t$ and whose state space is the hypercube $[0,1] \times \cdots \times [0,1]$ with constraint $\sum\limits_{j=1}^{s-1} \Pi_j \le 1$.

The Markov hypothesis is still

$$\begin{aligned} Pr[\{\overrightarrow{\underline{X}_{t_{n+1}}} \le \overrightarrow{\Pi}(n+1) \mid \overrightarrow{\underline{X}_{t_1}} = \overrightarrow{\Pi}(1), \overrightarrow{\underline{X}_{t_2}} = \overrightarrow{\Pi}(2), \ldots \overrightarrow{\underline{X}_{t_n}} = \overrightarrow{\Pi}(n)] \\ = Pr[\overrightarrow{\underline{X}_{t_{n+1}}} \le \overrightarrow{\Pi}(n+1) \mid \overrightarrow{\underline{X}_{t_n}} = \overrightarrow{\Pi}(n)]\;. \end{aligned}$$

If the process is homogeneous in time (has stationary transition probabilities), using a multidimensional analog of (1) and (2), we have

$$P_n(\overrightarrow{\Pi};\overrightarrow{y}) = Pr[\overrightarrow{\underline{X}_{m+n}} \le \overrightarrow{y} \mid \overrightarrow{\underline{X}_m} = \overrightarrow{\Pi}] \qquad \text{with} \quad y_j \le 1$$

$$P_n(\overrightarrow{\Pi};\overrightarrow{y}) = 0 \qquad \text{if} \quad \sum_{j=1}^{s-1} \Pi_j > 1$$

Let

$$p_n(\overrightarrow{\Pi};\overrightarrow{y}) = \frac{d}{d\overrightarrow{y}}\;P_n(\overrightarrow{\Pi};\overrightarrow{y})\;.$$

It can be proved that

$$p_{m+n}(\overrightarrow{\Pi};\overrightarrow{y}) = \int p_m(\overrightarrow{\Pi};\overrightarrow{z})\;p_n(\overrightarrow{z};\overrightarrow{y})\;d\overrightarrow{z}$$

where the integral is an $s-1$ multiple integral on the space $[0,1] \times [0,1] \times \cdots \times [0,1]$.

Let us notice that for two categories, as $\Pi_1(t) + \Pi_2(t) = 1$, $\underline{X}(t)$ is a one-dimensional process, so that the problem is a particular case of §5 where the state space is $[0,1]$ and not $]-\infty,+\infty[$.

Conclusion

In this paper, we have defined Symbolic Markov Chain for all the cases of Symbolic variables: quantitative, categorical single and multiple, interval, modal. We have also given the equivalent of Chapman-Kolmogorov equations in all cases. This property is the bases of the theoretical study of Markov Chains. We intend to continue this work in giving the more interesting results which give the knowledge of the state probabilities in interval and modal cases.

Let us note that in the case of continuous state space, we get interesting results with continuous time. In particular, the Kolmogorov equations are then known as the Fokker-Planck diffusion equations.

Acknowledgement

Our thanks to Yves Lechevallier and Paula Brito for their useful suggestions.

References

AFONSO, S., BILLARD, D. and DIDAY, E. (2004): Régression linéaire symbolique avec variables taxonomiques. *Revue RNTI.* G. Hébrail et al. Eds., Vol. 1, 205–210, Cépadues.

BAILEY, T.J.(1964) : *The elements of stochastic processes with applications to the natural sciences.* Wiley, New York.

BARTLETT, M.S. (1978): *An introduction to Stochastic processes.* Cambridge University Press, 3rd ed., Cambridge.

BEICHELT, F. (2006): *Stochastic processes in science, engineering and finance.* Chapman & Hall.

BOCH, H.H. and DIDAY, E. (2000): *Analysis of Symbolic Data.* Springer-Verlag, Berlin.

COX, D.R. and MILLER, H.D. (1965): *The theory of Stochastic processes.* Methuen & Co., London.

DARSOW, W.F., NGUYEN, B. and OLSEN, E.T. (1992): Copulas and Markov processes. *Illinois J. Math. 36, 600–642.*

De CARVALHO, F.A.T., NETO Eufrazi de A. Lima and TENERO C.P. (2004): A new method to fit a linear regression model for interval-valued data. In: S. Brundo, T. Fruckrirth and G. Palm (Eds.): *Advances in Artificial Intelligence: Proceedings of the Twenty Seventh German Conference on Artificial Intelligence.* Springer-Verlag, Berlin, 295–306.

FELLER W. (1968): *An introduction to probability theory and its applications* Wiley, New York.

GIRKHMAN, I.I. and SKOROKHOD, A.V. (2004): The theory of Stochastic processes. *Classics in Mathematics.* Springer-Verlag, Berlin.

JOE, H. (1997): *Multivariate models and dependence concepts.* Chapman & Hall, London.

KARLIN, S. (1966): *A first course in Stochastic processes.* Academic Press, New York.

LAWLER, G.F. (2006): *Introduction to Stochastic processes, 2nd Ed.* Chapman & Hall.

MEYN, S.P. and TWEEDIE R.L. (1993): Markov chains and Stochastic stability. *Communications & Control.* Springer-Verlag, New York.

NELSEN, R.B. (1999): An introduction to Copulas. *Lecture Notes in Statistics.* Springer, New York.

NEVEU, J. (1964): *Bases mathématiques du calcul des probabilités.* Masson, Paris.

PRABHU, N.V. (1965): *Stochastic Processes.* MacMillan, New York.

PRUDENCIO, R.B.C., LUDERNIR, T., and De CARVELHO F.A.T. (2004): A model sympolic classifier for selecting time series models. *Pattern Recognition Letters, 25 (8), 911–921.*

SOULE, A., SLAMETIAN, K., TAFT, N. and EMILION, R. (2004): Flow classification by histograms; *ACM Sigmetrics.* New York, http://ps.lip6.fr/˜soule/SiteWeb/Publication.php.

STIRZAKER, D. (2005): *Stochastic processes and models.* Oxford University Press, Oxford.