

A Multidimensional approach to Markov Chains Analysis: an application to Reinforcement Learning framework

*Un approccio Multidimensionale per l'Analisi delle Catene di Markov:
un'applicazione al contesto dell'Apprendimento con Rinforzo*

Marilena Fucili*, Danilo Leone**

Università degli Studi di Napoli "Federico II"

Dipartimento di Scienze Statistiche*, Dipartimento di Matematica e Statistica**

fucili@unina.it, danilo.leone@dsaonline.it

Riassunto: Le Catene di Markov sono spesso usate per modellizzare la dinamica di sistemi complessi. In particolare, nel contesto dell'Apprendimento con Rinforzo l'acquisizione di conoscenza si considera risultante dalla congiunta attività di esplorazione e ricezione di ritorni positivi o negativi prodotti in risposta dall'ambiente. In questo lavoro si propone un'originale metodologia di analisi delle strutture dei dati proprie di tale contesto mediante un approccio di analisi fattoriale non simmetrica che consente di indagare e visualizzare le dinamiche strutturali e relativi rinforzi, offrendo una visione della conoscenza appresa da un agente su mappe di apprendimento interpretabili dal punto di vista geometrico.

Keywords: Non Symmetrical Data Analysis, Non Symmetrical Correspondence Analysis, Markov Chains, Reinforcement Learning, Principal Component Analysis

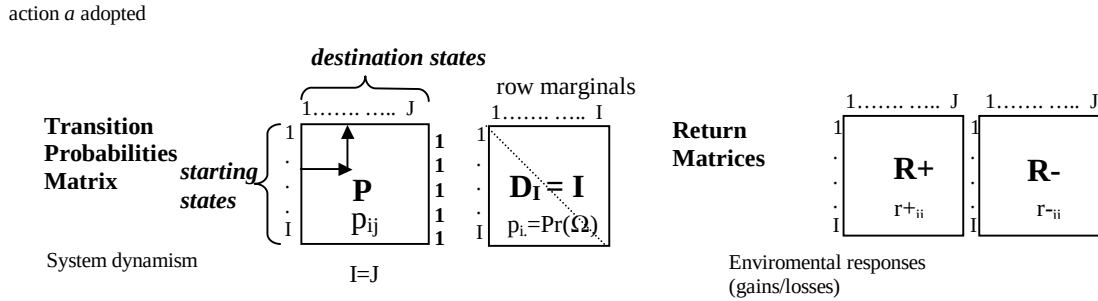
1. Introduction

Reinforcement Learning (Kaelbling et al., 1996) can be used to learn the optimal agent's behaviour by letting it interact with its dynamic environment. During the training the agent can move according to a trial and error approach or conforming to rules. Anyway it collects data related to its exploration and exploitation activities. The learning process provides information about transition probabilities and the reward or the penalty provided by the system each time a transition occurs. Rewards or penalties indicate the desirability of the resulting states. In this work the authors propose an original tool, i.e. the supplementary projection of Returns (positive and negative) points on the directions of maximum inertia based on the Non Symmetrical Data Analysis of a transition probabilities matrix. Such analysis has manifold advantages: it allows to discover and visualize the inner structure of Markov chains, the dynamics of the system and the advancement of the reinforcement learning process. It allows to decompose such structure according to different factorial axes effective as long as the dimension of the transition probability matrix grows up. Furthermore, it uses Non Symmetrical Correspondence Analysis (NSCA) (Lauro, D'Ambra, 1984) for dealing with the Markovian dependence assumption; it allows a direct visualization of the relationships between starting states, destination states and Returns by using graphical displays (Learning Maps) in the style of principal component analysis.

2. The Data Structure

An environment is given by a finite set of system states S , a set of possible actions A of the learner and a set of transition probabilities P that summarizes the inner dynamics occurred in the system when an action is performed. In Reinforcement Learning (RL) field the transitions of the system induce reactions of the environment which improve the learning process. The data structure we refer in this paper can be assumed as the result of the learning process performed by the agent related to the adoption of a specific action $a \in A$. When a relevant number of training cases is performed, it is possible to define a transition probabilities matrix. The Transition Probabilities Matrix $\mathbf{P}$ ($I \times J$) represents the conditional probabilities the system moves to one of the J destination states on the columns starting from one of the I states on the rows (both sets are assumed as levels of the same discrete variable observed at times t and $t+1$). It presents many peculiarities which imply original issues in geometrical interpretation of graphical results. In particular, there is an “asymmetrical” relationship between starting and destination states, and row marginals are equal to 1. That is why row marginal matrix $\mathbf{D_I}$ (which is used in obtaining the NSCA metric for the column space, $\mathbf{D_I^*}$, as we will show in the following paragraph) is an identity matrix. The gains and the losses associated to transitions occurred from each starting state to each destination state can be recorded in the *Matrix of Positive Returns (Rewards)* $\mathbf{R^+}$ ($I \times J$) and in the *Matrix of Negative Returns (Losses)* $\mathbf{R^-}$ ($I \times J$). Such information can be collected in the data structure reported in figure 1.

Figure 1: The data structure for a Reinforcement Learning Experiment



3. The Non Symmetrical Correspondence Analysis for Markov Chains with Reinforcement Returns

The analysis of the above described data structure can be performed by means of a multidimensional approach to Markov Chains analysis (MC-NSCA) (Leone, Fucili, 2003) based on Non Symmetrical Correspondence Analysis (Lauro, D'Ambra, 1984) and enriched by some particular tools to manage the RL framework. In the first step we compute the *Relative Transition Indices Matrix* $\mathbf{P^*}$ whose general element represents the weight of every transition from a state i to a state j (with $i=1, \dots, I$ and $j=1, \dots, J$) with respect to the whole set of transitions in the system. Successively $\mathbf{P^*}$ is centered with respect to the stochastic independence hypothesis of destination states from starting states to obtain the matrix $\bar{\mathbf{T}}$ and the row marginals matrix $\mathbf{D_I^*}$. In the following step a matrix of *Conditional Transition Indices of destination states | starting states* (centered row profiles) is derived as $\mathbf{D_I^{*-1} \bar{T}}$ with element $\bar{t}_{ij}^* = (p_{ij}^* - p_i^* p_{.j}^*) / p_i^*$.

MC-NSCA aims at searching the optimal synthesis of the latent transition structures in the collected data by looking for an orthonormal basis accounting for the largest part of system dynamism predictability. In accordance with the criterion optimized by NSCA, such purpose is reached by performing a PCA on the triplet $(\mathbf{D}_I^* \bar{\mathbf{T}}, \mathbf{I}, \mathbf{D}_I^*)$ and carrying out the eigen-analysis of the matrix $\mathbf{N} = \bar{\mathbf{T}}' \mathbf{D}_I^{*-1} \bar{\mathbf{T}}$. Following Greenacre (Greenacre, 1984) it can be considered as a particular case of General Analysis based on the Generalized Singular Value Decomposition: $\mathbf{D}_I^{*-1} \bar{\mathbf{T}} = \mathbf{U} \mathbf{\Lambda}^{1/2} \mathbf{V}'$ with the orthonormalizing constraints: $\mathbf{U}' \mathbf{D}_I^* \mathbf{U} = \mathbf{V}' \mathbf{V} = \mathbf{I}$, where $\mathbf{\Lambda}^{1/2}$ is the diagonal matrix of the square roots of the eigenvalues λ_α ($\alpha=1, \dots, I-1$) of $\mathbf{N}$ and $\mathbf{U}$ and $\mathbf{V}$ are respectively the left and right (generalized) singular vectors matrices of $\mathbf{D}_I^* \bar{\mathbf{T}}$. The coordinates of the starting states are $\mathbf{S} = \mathbf{D}_I^{*-1} \bar{\mathbf{T}} \mathbf{V}$ and the coordinates of the destination states are $\mathbf{E} = \bar{\mathbf{T}}' \mathbf{U}$. The information about Returns can be represented on factorial planes by means of supplementary points whose coordinates are based on the double centered matrices $\mathbf{R}^+$ and $\mathbf{R}^-$ as follows: *Reward coordinates*: $\mathbf{RC} = \bar{\mathbf{T}} \mathbf{R}^+ \mathbf{U}$, *Losses coordinates*: $\mathbf{LC} = \bar{\mathbf{T}} \mathbf{R}^- \mathbf{U}$.

4. Numerical example

An application on a real data-set will illustrate the usefulness of the methodology. The graphical display for the simple 5x5 Markov Chain of Figure 2 is proposed in Figure 4 with reference to the starting states and the destination states respectively identified by the labels “s” and “d” followed by the numbers of the states. This result is obtained by means of the MC-NSCA analysis.

Figure 2: Transition Probabilities Matrix for an almost stationary system

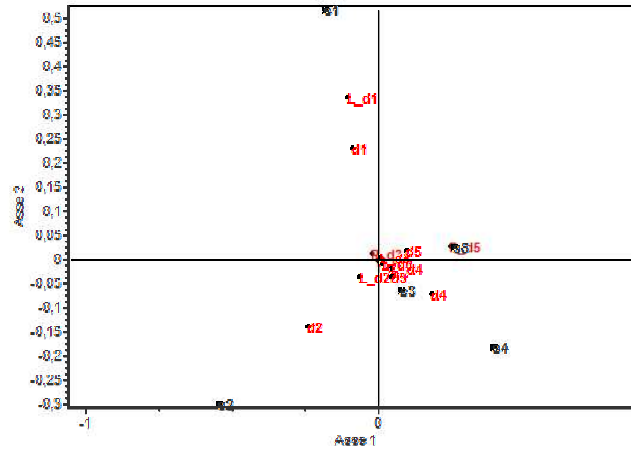
$$R3 = \begin{bmatrix} 0.7 & 0.125 & 0.025 & 0.075 & 0.075 \\ 0.1 & 0.8 & 0.04 & 0.03 & 0.03 \\ 0.1 & 0.15 & 0.5 & 0.15 & 0.1 \\ 0.025 & 0.075 & 0.125 & 0.7 & 0.075 \\ 0.066 & 0.066 & 0.068 & 0.2 & 0.6 \end{bmatrix}$$

A more complete Reinforcement Learning data-set includes the positive and negative Returns. The Return matrix for this example is a 5x5 matrix whose generic element $R(i,j)$ is the return gained by the agent just after having reached the destination state j starting from i . Examples of R matrices are proposed in the following Figure 3. $R1$ and $R2$ are filled with all the same values (zeros or unos) so they provide the agent no information. The graphical display (for both $R1$ and $R2$ cases) has all the rewards and the losses gathered together near the origin of the axes. The graphical display of rewards and losses in the factorial plan for the $R3$ matrix is proposed in the Figure 4.

Figure 3: R matrix of zeros, R matrix of unos and generic R matrix.

$$R1 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad R2 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \quad R3 = \begin{bmatrix} -1.5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 \end{bmatrix}$$

Figure 4: Graphical display of rewards and losses as supplementary points.



The graphical representations are interpretable in according with peculiar rules: the data structure allows to assume an euclidean metric in both the starting states and the destination states spaces (the metrics differs for a factor scale) jointly representable. Moreover a starting state point far from the origin indicates a great impact power on the behaviour of the destination state at time $t+1$ explained by the axe (transition structure). A low distance between two starting points suggests they have a similar impact. The destination states coordinates are measures of expectancy given a starting state and a low distance between two destination points represents a similar way of depending on the time t . Finally, the relative position of a starting state (see **p2**) and a destination state (see **d2**) can be evaluated in terms of squared cosinus of the angle spanned by the relative vectors and can be interpreted in terms of the influence of the first one on the transition of the system to the second one (which is stronger as smaller is such angle). The method allows to read the order of the possible transitions from a starting state by projecting the destination points on its direction. The presence of the negative return (Loss), with reference to the transition from **s1** to **d1**, is clearly visualized by the display in which **L-d1** is aligned with **s1** and **d1**. The presence of the positive return (Reward) with reference to the transition from **s5** to **d5** is clearly visualized by the display in which **R-d5** is aligned with **s5** and **d5**. The presence of the positive return (Reward) with reference to the transition from **s4** to **d3** is not disclosed by the display because of the low probability related to this transition.

References

- Bertsekas D.P. (1987) Dynamic Programming. Deterministic and Stochastic Models. Prentice Hall.
- Greenacre M., (1984) Theory and applications of Correspondence Analysis. Academic Press.
- Lauro C., D'Ambra L. (1984) Non symmetrical correspondence analysis. in: *Data Analysis and Informatics III* (eds E. Diday et al.), pp. 433-446. North Holland.
- Leone D., Fucili M. (2003) Graphical displays of Markov chains by means of nonsymmetrical correspondence analysis. in: *Abstracts of CARME 2003*.
- Kaelbling L., Littman M., Moore A., (1996) Reinforcement learning. A survey. *Journal of Artificial Intelligence Research*, 4, 237-285.