ON THE EMBEDDING PROBLEM FOR DISCRETE-TIME MARKOV CHAINS

MARIE-ANNE GUERRY,∗ Vrije Universiteit Brussel

Abstract

When a discrete-time homogenous Markov chain is observed at time intervals that correspond to its time unit, then the transition probabilities of the chain can be estimated using known maximum likelihood estimators. In this paper we consider a situation when a Markov chain is observed on time intervals with length equal to twice the time unit of the Markov chain. The question then arises of characterizing probability matrices whose square root(s) are also probability matrices. This characterization is referred to in the literature as the embedding problem for discrete time Markov chains. The probability matrix which has probability root(s) is called embeddable.

In this paper for two-state Markov chains, necessary and sufficient conditions for embeddability are formulated and the probability square roots of the transition matrix are presented in analytic form. In finding conditions for the existence of probability square roots for \((k \times k)\) transition matrices, properties of row-normalized matrices are examined. Besides the existence of probability square roots, the uniqueness of these solutions is discussed: In case of nonuniqueness, a procedure is introduced to identify a transition matrix that takes into account the specificity of the concrete context. In case of nonexistence of a probability root, the concept of approximate probability root is introduced as a solution of an optimization problem related to approximate nonnegative matrix factorization.

Keywords: Markov chain; Probability matrix; Root of a matrix; Embedding problem

2000 Mathematics Subject Classification: Primary 15A23; 15A51; 60J10

Secondary

∗ Postal address: MOSI, Vrije Universiteit Brussel, Pleinlaan 2, B-1050 Brussels, Belgium
1. Introduction

A discrete-time Markov chain is considered for which the $t$-th outcome corresponds with the state in which the process is in at time $t$. The length of the time interval between two subsequent time points $t$ and $t+1$ will be referred to as the time unit of the Markov chain. A characterization of the Markov chain is then given by its transition matrix $P = (p_{ij})$ of transition probabilities between the states on a time interval with unit length.

Based on observations regarding the number of objects in each of the states and the number of transitions between the states, the transition matrix of a Markov chain can be estimated. The transition probability $p_{ij}$ (for $i, j \in \{1, \cdots, k\}$) from state $i$ to state $j$ in one period of time can be estimated by the maximum likelihood estimator ([1], p. 113):

$$\hat{p}_{ij} = \frac{\sum_{t=0}^{T-1} n_{ij}(t, t+1)}{\sum_{t=0}^{T-1} n_i(t)} \tag{1.1}$$

The notation $n_i(t)$ refers to the number of objects in state $i$ at time $t$; and $n_{ij}(t, t+1)$ stands for the number of objects in state $i$ at time $t$ that are in state $j$ at time $t + 1$ (for $t = 0, \cdots, T - 1$).

The estimations for the transition probabilities according to equation (1.1) are based on data for the observable variables regarding time intervals that correspond to the time unit of the Markov chain. In case there are no observations available for some variables or in case there are no observations available regarding the time unit of the Markov chain, there is a lack of data. In previous work there is dealt with the problem of a lack of observations for some variables by building a hidden Markov model or a Markov switching model ([5], [10], [6]) that takes into account latent sources of heterogeneity ([13]). A lack of observations regarding the time unit of the Markov chain occurs when, for example, data is only available on time intervals with length 1 and when it is preferable to get for intervals with length 0.5 estimations for the transition probabilities. In that situation there is a lack of data of the type with which there is dealt in the present paper. The approach is related to the embedding problem.
as considered in [11] and for continuous time Markov models in [12]. In the present paper embeddability is examined more in detail for discrete-time Markov chains.

An answer to problems of this kind can be useful, among other fields, in manpower planning. A Markov model for manpower planning takes into account internal transitions (e.g. promotions), outgoing flows (i.e. wastage) as well incoming flows (i.e. recruitments) ([1]). For a company with known recruitment probabilities for each of the personnel categories and with quantified promotion probabilities from one personnel category to another regarding a period of one year, the expected number of members in each category can be computed year after year. In case management is interested in having an idea of the number of members after half a year, taking into account that for motivational reasons it is preferable to have equal promotion probabilities for each semester (i.e. under time homogeneous conditions), the methodology presented in this paper provides in some interesting insights. Namely for this company the promotion probabilities for a semester can be quantified as well as the expected number of members after each 6 months can be computed.

The paper is organized as follows. In section 2 the embedding problem of the discrete-time Markov chain is formulated in terms of probability square roots of its transition matrix. In section 3 for two-state Markov chains, necessary and sufficient conditions for embeddability are formulated and the probability square roots of the transition matrix are presented in analytic form. In finding conditions for the existence of probability square roots for \((k \times k)\) transition matrices, properties of row-normalized matrices are examined. Based on the Jordan matrix of a transition matrix, square roots and their properties are examined. In section 4 the concept of approximate probability square roots is introduced. In case a transition matrix has no probability square roots, approximate probability square roots can be considered that are the solutions of a presented optimization problem. In section 5 the situation is discussed in which for a transition matrix there exists more than one probability square root. In answer to the identification problem a desirable probability square root can be selected based on presented criteria. Section 6 provides in an illustration.
2. Transition probabilities for a time interval with length 0.5

Let us denote the matrix of the transition probabilities for a time interval with length 1 by $P(1) = (p_{ij}(1))$ and the matrix of the transition probabilities for a time interval with length 0.5 by $P(0.5) = (p_{ij}(0.5))$. Under the assumption of time homogeneity, the following relation holds:

$$p_{ij}(1) = \sum_m p_{im}(0.5) p_{mj}(0.5) \quad \forall i,j \in \{1, \cdots, k\}$$

Or in matrix notation:

$$P(1) = P(0.5) \times P(0.5)$$

In case observations are available for time intervals with length 1, the transition probabilities $p_{ij}(1)$ can be estimated by expression (1.1), resulting in the matrix $\hat{P}(1) = (\hat{p}_{ij}(1))$. One could expect that the estimations for $p_{ij}(0.5)$ then satisfy:

$$\hat{P}(1) = \hat{P}(0.5) \times \hat{P}(0.5)$$

So in case no observations are available for a time interval with length 0.5 the problem is to find a transition matrix $\hat{P}(0.5)$, compatible with the matrix $\hat{P}(1)$ that is estimated based on observations for time intervals with length 1. Therefore the goal is to find a probability matrix $A$ satisfying $\hat{P}(1) = A \times A$, i.e. a square root $A$ of the matrix $\hat{P}(1)$ within the set of probability matrices

$$\Pi = \left\{ M = (m_{ij}) \in \mathbb{R}^{k \times k} \left| \sum_{j=1}^{k} m_{ij} = 1 \text{ and } m_{ij} \geq 0 \quad \forall i,j \in \{1, \cdots, k\} \right. \right\}$$

In this way the stated question can be seen as an embedding problem. The embedding problem for continuous time Markov processes is discussed in detail in Singer and Spilerman (1974) and Singer and Spilerman (1976). Deciding whether the observed matrix $\hat{P}(1)$ can be represented in the form $A^2$ for some probability matrix $A$ is a version of the embedding problem for discrete-time Markov chains. In this paper the discussion of the embedding problem is focused on discrete-time Markov chains and square roots of probability matrices. The question is whether a Markov chain with time
unit 0.5 exists that is compatible with $\hat{P}(1)$. In case this is the situation, further goals are to find a procedure to determine solution(s) for the transition matrix with respect to time unit 0.5 and to identify a unique solution taking into account the specificity of the concrete context. Those phases in the approach are called respectively the embedding, the inverse and the identification problem ([11]).

### 3. Probability square roots

A square root $A \in \Pi$ of the probability matrix $\hat{P}(1)$ provides in a Markov chain with time unit 0.5 that is compatible with $\hat{P}(1)$. In what follows, such a matrix $A$ will be called a probability square root of $\hat{P}(1)$.

**Definition 1.** For a probability matrix $P$, a probability square root $A$ of $P$ is a probability matrix $A$ that is a square root of $P$.

For a $(2 \times 2)$ probability matrix $P = \begin{pmatrix} c & 1 - c \\ d & 1 - d \end{pmatrix}$, the probability matrix $A = (a_{ij})$ is a probability square root of $P$ if and only if

\[
\begin{align*}
    a_{11}^2 + (1 - a_{11})a_{21} &= c \\
    a_{11}a_{21} + (1 - a_{21})a_{21} &= d 
\end{align*}
\] (3.1)

In Theorem 3.1 some results are formulated concerning probability square roots for the category of $(2 \times 2)$ probability matrices. Depending on the values of the $(2 \times 2)$ probability matrix $P$, there can exist either no or exact one probability square root, or two probability square roots:

**Theorem 3.1.** For a $(2 \times 2)$ probability matrix $P = \begin{pmatrix} c & 1 - c \\ d & 1 - d \end{pmatrix}$

- If $c < d$ there does not exist a probability square root $A$ of $P$.
- If $c = d$, there does exist exactly one probability square root $A$ of $P$, namely the probability matrix $A = (a_{ij})$ with $a_{11} = a_{21} = c = d$.
- If $c > d$ and $1 - c + d \neq 0$ there does exist at least one probability square root $A$ of $P$, namely the probability matrix $A = (a_{ij})$ with

\[
a_{11} = \frac{\sqrt{c - d}(1 - c) + d}{1 - c + d} \quad \text{and} \quad a_{21} = d \frac{1 - \sqrt{c - d}}{1 - c + d}
\]
Moreover, in case
\[ a_{11} = \frac{\sqrt{c - d} (c - 1) + d}{1 - c + d} \quad \text{and} \quad a_{21} = d \frac{1 + \sqrt{c - d}}{1 - c + d} \]
are both elements of \([0,1]\) there does exist a second probability square root \(A = (a_{ij})\) of \(P\).

- If \(c > d\) and \(1 - c + d = 0\) both
  \[
  \begin{pmatrix}
  1 & 0 \\
  0 & 1
  \end{pmatrix}
  \quad \text{and} \quad
  \begin{pmatrix}
  0 & 1 \\
  1 & 0
  \end{pmatrix}
  \]
  are probability square roots of \(P\).

**Proof.** See proof of Theorem 3.1 in Appendix.

There can be remarked that the three possible situations effectively occur: a \((2 \times 2)\) probability matrix can have no, exactly one or two probability square roots. So for example for the case \(c > d\), \(P = \begin{pmatrix} \frac{5}{8} & \frac{3}{8} \\ \frac{3}{8} & \frac{5}{8} \end{pmatrix}\) has two probability square roots and \(P = \begin{pmatrix} 11/27 & 16/27 \\ 1/9 & 8/9 \end{pmatrix}\) has one probability square root.

Theorem 3.1 provides implicitly, for a \((2 \times 2)\) probability matrix \(P\), in a necessary and sufficient condition to have a probability square root:

For a \((2 \times 2)\) probability matrix \(P = \begin{pmatrix} c & 1 - c \\ d & 1 - d \end{pmatrix}\), there does exist a probability square root of \(P\) if and only if \(c \geq d\). This condition can be formulated equivalently in terms of the trace of \(P = A \times A\): \(tr(P) = tr(A \times A) \geq 1\).

In the study of necessary conditions for a \((k \times k)\) probability matrix to have a probability square root \(A\), the following Theorem can be useful since it formulates some properties of \(tr(A \times A)\).

**Theorem 3.2.** Within the set of \((k \times k)\) probability matrices \(A\), for a critical point holds that: \(tr(A \times A) = 1\).

**Proof.** For a \((k \times k)\) probability matrix \(A = (a_{ij})\), holds:

\[
tr(A \times A) = \sum_i (A \times A)_{ii} = \sum_i \left[ (a_{ii})^2 + \sum_{i \neq j} a_{ij}a_{ji} \right] = \sum_i \left[ (1 - \sum_{i \neq j} a_{ij})^2 + \sum_{i \neq j} a_{ij}a_{ji} \right]
\]

In this way, \(tr(A \times A)\) is expressed as a function \(f\) of the \((k - 1)^2\) elements \(a_{rs}\), with
On the embedding problem for discrete-time Markov chains

$r \neq s \in \{1, \ldots, k\}$, of the probability matrix $A$. Whereby a critical point of $f$ satisfies $
exists s \neq r \in \{1, \ldots, k\}$:

$$\frac{\partial f}{a_{rs}} = -2 \left( 1 - \sum_{i \neq r} a_{ri} \right) + a_{sr} + a_{sr} = 0$$

Therefore for a critical point of $f$ holds $\nexists s \neq r \in \{1, \ldots, k\}$: $\sum_{i \neq r} a_{ri} + a_{sr} = 1$.

Consequently for $r \in \{1, \ldots, k\}$ and $s \neq r$ all the elements $a_{sr}$ are equal to $a_{sr} = 1 - \sum_{i \neq r} a_{ri} = a_{rr}$. Resulting in the fact that the critical point of $f$ corresponds with a matrix $A$ having all the elements of the $r$-th column equal, let us say equal to $a_{1r}$.

This property gives rise to:

$$tr(A \times A) = \sum_{l} \left[ (a_{1l})^2 + \sum_{i \neq l} a_{1i} a_{1l} \right] = \sum_{l} a_{1l} \left[ a_{1l} + \sum_{i \neq l} a_{1i} \right] = \sum_{l} a_{1l} = 1.$$ 

Which proves the Theorem.

As specified in [7], for the specific class of $(3 \times 3)$ and $(4 \times 4)$ state-wise monotone probability matrices $A$ holds that $tr(A \times A) \geq 1$, resulting in the necessary condition $tr(P) \geq 1$ for a probability matrix $P$ of order $(3 \times 3)$ or $(4 \times 4)$ to have a state-wise monotone probability square root. However without requiring any restriction on the probability root(s), $tr(P) \geq 1$ is no necessary condition on $P$ to have a probability root: $P$ has a probability root $A$, while $tr(P) < 1$ for example for

$$P = \begin{pmatrix} 0.17 & 0.66 & 0.17 \\ 0.17 & 0.17 & 0.66 \\ 0.66 & 0.17 & 0.17 \end{pmatrix} \quad \text{and} \quad A = \begin{pmatrix} 0.1 & 0.1 & 0.8 \\ 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \end{pmatrix}.$$ 

Before examining sufficient conditions for the existence of probability square roots for a $(k \times k)$ probability matrix, some properties are discussed for matrices of the following less restrictive set

$$\Gamma = \left\{ M = (m_{ij}) \in \mathbb{R}^{k \times k} \middle| \sum_{j=1}^{k} m_{ij} = 1 \quad \forall i \in \{1, \ldots, k\} \right\} \supset \Pi$$

**Definition 2.** A row-normalized matrix is a matrix with elements in each row adding up to one.
In what follows the notation $H_0$ refers to the subset of vectors of $\mathbb{R}^k$ with elements summing up to zero:

$$H_0 = \left\{ v = (v_i) \in \mathbb{R}^k \left| \sum_{i=1}^k v_i = 0 \right. \right\}$$

**Lemma 3.1.** For $\lambda \neq 1$, $u \in \mathbb{R}^k$, $n \in \{1, 2, 3, \ldots\}$ and $P$ a $(k \times k)$ row-normalized matrix holds

$$u \times (P - \lambda I)^n \in H_0 \Rightarrow u \in H_0$$

**Proof.** See proof of Lemma 3.1 in Appendix.

In what follows the terminology eigenvector refers to a left eigenvector: the $(1 \times k)$ vector $E$ is an eigenvector of the $(k \times k)$ matrix $P$ associated with the eigenvalue $\lambda$ iff $E \times P = \lambda P$.

**Corollary 3.1.** For a row-normalized matrix, all eigenvectors and generalized eigenvectors associated with an eigenvalue $\lambda \neq 1$ are elements of $H_0$.

**Proof.** For the row-normalized matrix $P$, a (generalized) eigenvector $E$ associated with the eigenvalue $\lambda$ satisfies $E \times (P - \lambda I)^n = 0$ for some $n \in \{1, 2, 3, \ldots\}$. Therefore $E \times (P - \lambda I)^n$ is an element of $H_0$ and according to Lemma 3.1 the (generalized) eigenvector $E$ itself is an element of $H_0$. Which proves the Corollary.

As there is examined in previous work (e.g. [4]) the effect of some particular transformations on the analyses of percentage data, there is proved in the following lemma, under some conditions, that for a $(k \times k)$ row-normalized matrix $P = T^{-1} \times J \times T$ with Jordan matrix $J$, the transformations $f_m : \Gamma \rightarrow \Gamma : T^{-1} \times J \times T \rightarrow T^{-1} \times J^m \times T$ (for $m \in \mathbb{R}$) do not affect the row sums of the matrix.

**Lemma 3.2.** For a $(k \times k)$ row-normalized matrix $P = T^{-1} \times J \times T$ with Jordan matrix $J$, the row sums of $T^{-1} \times J^m \times T$ (as far as this matrix is defined) do not depend on the value of $m \in \mathbb{R}$.

**Proof.** See proof of Lemma 3.2 in Appendix.
Lemma 3.3. For a \((k \times k)\) row-normalized matrix \(P = T^{-1} \times J \times T\) with Jordan matrix \(J\), \(A = T^{-1} \times J^5 \times T\) (as far as this matrix is a defined element of \(\mathbb{R}^{k \times k}\)) is a row-normalized matrix for which \(P = A \times A\).

Proof. It is clear that for \(A = T^{-1} \times J^5 \times T\): \(P = A \times A\).

Moreover since \(P\) is a row-normalized matrix, the row-sums of \(T^{-1} \times J \times T\) are equal to 1. According to Lemma 3.2, the row-sums of \(T^{-1} \times J \times T\) and \(T^{-1} \times J^5 \times T\) are equal. Therefore the matrix \(T^{-1} \times J^5 \times T\) is row-normalized as well.

Corollary 3.2. For a \((k \times k)\) diagonalizable row-normalized matrix \(P = T^{-1} \times D \times T\), with all eigenvalues positive and real, \(A = T^{-1} \times D^5 \times T\) is a row-normalized matrix for which \(P = A \times A\).

Proof. Since for a diagonalizable matrix \(P\) the Jordan matrix is a diagonal matrix \(D\) with diagonal elements eigenvalues of \(P\), Lemma 3.3 is applicable for \(J = D\) and proves the Corollary.

In what follows the more restrictive set \(\Pi\) of probability matrices is examined. In particular properties of square roots of probability matrices are discussed.

Lemma 3.4. Each \((2 \times 2)\) probability matrix \(P\) is diagonalizable.

Proof. See proof of Lemma 3.4 in Appendix.

Theorem 3.3. For a \((k \times k)\) probability matrix \(P = T^{-1} \times J \times T\), with Jordan matrix \(J\) and with all eigenvalues positive and real holds:

- In case \(k = 2\): \(T^{-1} \times J^5 \times T\) is a nonnegative matrix.
- In case \(k > 2\): \(T^{-1} \times J^5 \times T\) is not necessarily a nonnegative matrix.

Proof. In case \(k = 2\):

According to Lemma 3.4 the Jordan matrix of the \((2 \times 2)\) probability matrix \(P\) is a diagonal matrix \(D\). In the situation that \(\lambda_1 = \lambda_2 = 1\), the diagonal matrix \(D\) equals the identity matrix \(I_{2 \times 2} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\) and consequently \(T^{-1} \times D^5 \times T = T^{-1} \times T = I_{2 \times 2}\) is a nonnegative matrix.
In the situation that $\lambda_1 = 1$ and $\lambda_2 \neq 1$: Let us denote by $E_1 = (a \ b)$ an eigenvector of $P$ associated with the eigenvalue $\lambda_1 = 1$. According to Corollary 3.1 it is known that $E_2 = (1 \ -1)$ is an eigenvector of $P$ associated with the eigenvalue $\lambda_2 \neq 1$.

Consequently $P = T^{-1} \times D \times T$ with

$$T = \begin{pmatrix} a & b \\ 1 & -1 \end{pmatrix}, \quad T^{-1} = \frac{1}{a+b} \begin{pmatrix} 1 & b \\ 1 & -a \end{pmatrix}, \quad D = \begin{pmatrix} 1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$

Resulting in

$$P = T^{-1} \times D \times T = \frac{1}{a+b} \begin{pmatrix} a + b\lambda_2 & b - b\lambda_2 \\ a - a\lambda_2 & b + a\lambda_2 \end{pmatrix}$$

and

$$T^{-1} \times D^5 \times T = \frac{1}{a+b} \begin{pmatrix} a + b\sqrt{\lambda_2} & b - b\sqrt{\lambda_2} \\ a - a\sqrt{\lambda_2} & b + a\sqrt{\lambda_2} \end{pmatrix}$$

Since the eigenvalues of the probability matrix $P = (p_{ij})$ are assumed to be real, according to the Perron-Frobenius Theorem the eigenvalue $\lambda_2$ is less than 1. Therefore $p_{12} \geq 0$ implies that $\frac{b}{a+b} \geq 0$ and from $p_{21} \geq 0$ it is known that $\frac{a}{a+b} \geq 0$. Consequently $(T^{-1} \times D^5 \times T)_{12} \geq 0$ and $(T^{-1} \times D^5 \times T)_{21} \geq 0$.

Moreover $\sqrt{\lambda_2} \geq \lambda_2$ and therefore $(T^{-1} \times D^5 \times T)_{11} \geq (T^{-1} \times D \times T)_{11} = p_{11} \geq 0$ and $(T^{-1} \times D^5 \times T)_{22} \geq (T^{-1} \times D \times T)_{22} = p_{22} \geq 0$.

Which proves that the elements of $T^{-1} \times D^5 \times T$ are all nonnegative.

In case $k > 2$:

For the diagonalizable $(3 \times 3)$ probability matrix

$$P = T^{-1} \times D \times T = \begin{pmatrix} 1 & 0 & 0 \\ 1 & -1 & 0 \\ 1 & -3 & 1 \end{pmatrix} \times \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{36} \end{pmatrix} \times \begin{pmatrix} 1 & 0 & 0 \\ 1 & -1 & 0 \\ 2 & -3 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ \frac{11}{36} & \frac{24}{36} & \frac{1}{36} \\ \frac{11}{36} & \frac{24}{36} & \frac{1}{36} \end{pmatrix}$$

with eigenvalues $1, \frac{1}{4}$ and $\frac{1}{36}$ all positive and real, holds that
\[ T^{-1} \times D^5 \times T = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ -\frac{1}{5} & 1 & \frac{1}{5} \end{pmatrix} \]

is a matrix that is not nonnegative.

Which proves that for \( k > 2 \), there do exist \((k \times k)\) probability matrices \( P = T^{-1} \times J \times T \), with all eigenvalues positive and real and with \( T^{-1} \times J^5 \times T \) a matrix that is not nonnegative.

**Theorem 3.4.** For a \((k \times k)\) diagonalizable probability matrix \( P = T^{-1} \times D \times T \), with all eigenvalues positive and real holds:

- In case \( k = 2 \): \( T^{-1} \times D^5 \times T \) is a probability square root of \( P \).
- In case \( k > 2 \): \( T^{-1} \times D^5 \times T \) is a row-normalized square root of \( P \).

**Proof.** According to Corollary 3.2, for a diagonalizable \((k \times k)\) probability matrix \( P = T^{-1} \times D \times T \), with all eigenvalues positive and real, \( T^{-1} \times D^5 \times T \) is a row-normalized square root of \( P \).

Moreover in case \( k = 2 \), according to Theorem 3.3, \( T^{-1} \times D^5 \times T \) is a nonnegative matrix. And therefore \( T^{-1} \times D^5 \times T \) is a probability square root of \( P \).

Which proves the Theorem.

### 4. Approximate probability square roots

From the previous results it is clear that, depending on the elements of the probability matrix \( P \), there exists no, exactly one or more than one probability square root \( A \) of \( P \).

In case no probability square root exists, the question that can be stated is based on what method a probability matrix \( B \) can be determined such that \( B \times B \) results in a 'good' approximation of \( P \).

In this paper, this problem will be referred to as the problem of finding approximate probability square roots and will be tackled by an approach inspired by nonnegative matrix factorization.
Nonnegative matrix factorization deals in general with the following type of problem: Given a nonnegative matrix $U$, find nonnegative factors $V$ and $W$ such that $U \approx V \times W$ ([3], [8]).

The quality of the approximation is hereby expressed in terms of a cost function, constructed by using a measure of distance between two matrices. For example the square of the Euclidean distance, resulting for the matrices $R = (r_{ij})$ and $S = (s_{ij})$ in:

$$\|R - S\|^2 = \sum_{i,j} (r_{ij} - s_{ij})^2$$

In the more specific context of finding approximate probability square roots, this approach results in the following formulation: Given a probability matrix $P$ of order $(k \times k)$, find a probability matrix $B$ of order $(k \times k)$ such that $P \approx B \times B$.

I.e. find a $(k \times k)$ matrix $B = (b_{ij})$ solution of the following optimization problem

$$\text{Min} \|P - B \times B\|^2$$

subject to the constraints

$$\sum_j b_{ij} = 1 \quad \text{and} \quad b_{ij} \geq 0 \ \forall i,j$$

This formulation gives rise to an optimization problem of the Karush-Kuhn-Tucker type [2]. And a solution $B$ of this optimization problem results in an approximate probability square root of $P$.

5. Selecting a desirable (approximate) probability square root

The concepts of probability square root and approximate probability square root are introduced in order to find, based on the matrix $\hat{P}(1)$, estimations for the number of objects in the states after each time interval with length 0.5. In case for $\hat{P}(1)$ more than one (approximate) probability square root exists, the observations are consistent with more than one discrete-time Markov chain with time unit 0.5. The problem is then of how to select within these set of alternatives. Singer and Spilerman have dealt with this identification problem in general [11].
By Theorem 3.1 it is proved that there do exist \((2 \times 2)\) probability matrices that have more than one probability square root. The nonuniqueness of probability square roots is a fact, more in general, for probability matrices of any order \((k \times k), k \geq 2\): For \(Q\) a \((2 \times 2)\) probability matrix having two probability square roots \(\hat{A} \neq \bar{A}\) and for \(I\) the identity matrix of order \(((k - 2) \times (k - 2))\), the block diagonal matrix \(P = \text{diag}(Q, I)\) is a \((k \times k)\) probability matrix having at least 2 probability roots, namely \(\text{diag}(\hat{A}, I)\) and \(\text{diag}(\bar{A}, I)\). Since for each \(k \geq 2\) there do exist \((k \times k)\) probability matrices with more than one probability square root, it is worth to deal with the identification problem [11].

In the context of this paper, in case more than one (approximate) probability square root exists, the question is whether all the square roots result in desirable evolutions of the expected number of objects in the states. The criterion to handle in selecting the most desirable probability square root can vary depending on the phenomenon that is modeled by the Markov chain. One could prefer to have an evolution in time of the vector \(n(t) = (n_i(t))\) such that for each state \(i\) there is as less as possible fluctuation of \(n_i(t)\), with in extremis an evolution of \(n_i(t)\) that is monotonous (increasing or decreasing). It could be preferable that the expected number of objects after half a period of time \(n(t) \times \hat{P}(0.5)\) is ‘somewhere in between’ the starting vector \(n(t)\) and the expected vector after one period of time \(n(t) \times \hat{P}(1)\). This condition can be expressed in terms of the intervals

\[
I_i = \left[ \min \left\{ (n(t))_i, \left( n(t) \times \hat{P}(1) \right)_i \right\}, \max \left\{ (n(t))_i, \left( n(t) \times \hat{P}(1) \right)_i \right\} \right]
\]

as follows:

\[
\forall i \in \{1, \cdots, k\} : \left( n(t) \times \hat{P}(0.5) \right)_i \in I_i
\]

Such a formulated condition can be for example useful in modeling a manpower system by a Markov chain with states corresponding to homogeneous grades in the company. It could be preferable for each of the grades in the organization that the fluctuations in the number of members are limited, or moreover that the number of members increases/decreases monotonously.
In case there does not exist a (approximate) probability square root for which this property holds, the criterion to select \( \hat{P}(0.5) \) can then be formulated in terms of an optimization problem, namely:

\[ \hat{P}(0.5) \text{ is a (approximate) probability square root of } \hat{P}(1) \text{ minimizing the maximum value over the different states } i \in \{1, \ldots, k\} \text{ of the discrepancy of the expected number of members } \left( n(t) \times \hat{P}(0.5) \right)_i \text{ and the elements of the interval } I_i, \text{ i.e. } \hat{P}(0.5) \text{ is a (approximate) probability square root minimizing } \max_{i \in \{1, \ldots, k\}} \min_{x \in I_i} \left| \left( n(t) \times \hat{P}(0.5) \right)_i - x \right|. \]

6. Illustration

For a two-state manpower system with at time \( t = 0 \) the number of members given by \( n(0) = (60 \ 42) \) and with the number of transitions from \( t = 0 \) to \( t = 1 \) equal to \( n_{11} = 45, n_{12} = 15, n_{21} = 21, n_{22} = 21 \), the estimated transition matrix for a time interval with length one is

\[ \hat{P}(1) = \begin{pmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} \]

According to Theorem 3.1 this transition matrix \( \hat{P}(1) \) has two probability square roots, namely

\[ A_1 = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix} \text{ and } A_2 = \begin{pmatrix} \frac{5}{6} & \frac{1}{6} \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix} \]

For the probability square roots \( A_1 \) and \( A_2 \) the corresponding evolution of the expected number of members is as follows:

\[
\begin{align*}
\text{n}(0) &= (60 \ 42) & \text{n}(0) \times A_1 &= (72 \ 30) & \text{n}(1) &= n(0) \times A_1^2 = (66 \ 36) & \cdots \\
\text{n}(0) &= (60 \ 42) & \text{n}(0) \times A_2 &= (64 \ 38) & \text{n}(1) &= n(0) \times A_2^2 = (66 \ 36) & \cdots
\end{align*}
\]

The probability square root \( A_2 \) results in a monotonous evolution of the expected number of members that is, as discussed in Section 5, probably desirable in a manpower system context.

7. Generalizations

In this paper the discussion of the embedding problem for discrete-time Markov chains is restricted to time intervals with length 0.5. For a Markov chain with time
On the embedding problem for discrete-time Markov chains

unit 1 and transition matrix $P(1)$, the study is focused on finding a Markov chain with
time unit 0.5 that is compatible with $P(1)$. The stated problem and the results are
therefore formulated in terms of probability square roots $P(0.5)$. The idea of finding a
transition matrix for a Markov chain with time unit equal to half the length of the time
intervals on which there are observations available, can be generalized to the problem
of finding a transition matrix $P \left( \frac{1}{m} \right)$ for $m \in \mathbb{N}_0$ of a Markov chain with time unit $\frac{1}{m}$
and that is compatible with $P(1)$.

8. Appendix

Proof of Theorem 3.1.

In the special case of a probability matrix $P = \begin{pmatrix} c & 1-c \\ d & 1-d \end{pmatrix}$ satisfying $1-c+d = 0$,
the matrix $P$ equals $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and has 2 probability square roots, namely $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$
and $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$.

In the special case of a probability matrix $A = (a_{ij})$ with $a_{11} = 1$, the matrix $A$
can only be a square root of a probability matrix of the form $P = \begin{pmatrix} 1 & 0 \\ d & 1-d \end{pmatrix}$.
Under these conditions and according to (3.1), $P$ has exactly one probability square
root namely the probability matrix $A = (a_{ij})$ with $a_{11} = 1$ and $a_{21} = 1 - \sqrt{1-d}$.

The further reasoning can be restricted to the situation of a probability matrix
$P = \begin{pmatrix} c & 1-c \\ d & 1-d \end{pmatrix}$ satisfying $1-c+d \neq 0$ and a probability square root $A = (a_{ij})$
with $a_{11} \neq 1$: The system (3.1) results in the following quadratic equation in $a_{11}$

$$(1-c+d)a_{11}^2 - (2d)a_{11} + c^2 - c + d = 0 \quad (8.1)$$

with discriminant $D = 4(1-c)^2.(c-d)$.

Therefore, in case $c < d$, no probability square root exists for the matrix $P$. In case
$c = d$, the matrix $P$ has exactly one probability square root $A$. According to (3.1) this
matrix $A$ satisfies $a_{11} = a_{21} = c = d$. In case $c > d$, the equation (8.1) results in the
following solutions for $a_{11}$:

$$a_{11} = \frac{\sqrt{c - d}(1 - c) + d}{1 - c + d} \quad \text{and} \quad a_{11} = \frac{\sqrt{c - d}(c - 1) + d}{1 - c + d}$$

According to (3.1), $a_{21}$ can be expressed as $a_{21} = \frac{c - a_{11}^2}{1 - a_{11}}$, resulting respectively in:

$$a_{21} = d \frac{1 - \sqrt{c - d}}{1 - c + d} \quad \text{and} \quad a_{21} = d \frac{1 + \sqrt{c - d}}{1 - c + d}$$

Since both $a_{11} = \frac{\sqrt{c - d}(1-c) + d}{1 - c + d}$ and $a_{21} = \frac{c-a_{11}^2}{1-a_{11}}$ are elements of $[0,1]$ there does exist at least one probability square root $A = (a_{ij})$ of $P$.

Which proves the Theorem.

**Proof of Lemma 3.1.**

The property formulated in Lemma 3.1 can be proved by induction.

For a row-normalized matrix $P$ and $u \in \mathbb{R}^k$ satisfying $u \times (P - \lambda I) \in H_0$ holds

$$\sum_j [u \times (P - \lambda I)]_j = \sum_j \sum_i u_i p_{ij} - \lambda \sum_j u_j = \sum_i u_i \sum_j p_{ij} - \lambda \sum_i u_i = \sum_i u_i (1 - \lambda) = 0$$

Which implies that $\sum_i u_i = 0$ in case $\lambda \neq 1$ and proves the property for $n = 1$.

Let us now assume that the property holds for $n - 1$. If $u \times (P - \lambda I)^n \in H_0$ then also $u \times (P - \lambda I)^{n-1} \times (P - \lambda I) \in H_0$. Consequently $u \times (P - \lambda I)^{n-1} \in H_0$ (since the property holds for $n = 1$) and therefore $u \in H_0$ (since the property holds for $n - 1$).

Which proves the Lemma.

**Proof of Lemma 3.2.**

For a probability matrix $P$ the eigenvalue $\lambda = 1$ is a semisimple eigenvalue, i.e. the Jordan blocks are of order $(1 \times 1)$ ([9]). Therefore the Jordan blocks corresponding to $\lambda = 1$ result in the $(r \times r)$ identity matrix, in case the algebraic multiplicity of $\lambda = 1$ equals $r$. Let us consider the order of the Jordan blocks such that the first $r$ blocks correspond to $\lambda = 1$. In this way the Jordan matrix, as well as its powers, satisfies $J_{is} = \delta_{is}$ ($\forall s \in \{1,\ldots,r\}$ $\forall l \in \{1,\ldots,k\}$) in which the notation $\delta$ stands for Kronecker delta. The matrix $T$ is then a matrix with in the first $r$ rows eigenvectors associated with $\lambda = 1$ and in each other row a (generalized) eigenvector associated
On the embedding problem for discrete-time Markov chains

with an eigenvalue different from 1. Therefore \( \forall i \in \{1, ..., k\} \) holds:

\[
\sum_j (T^{-1} \times J^m \times T)_{ij} = \sum_j \sum_l \sum_s \left( T^{-1} \right)_{il} (J^m)_{ls} T_{sj}
\]

\[
= \sum_l \left( T^{-1} \right)_{il} \left[ \sum_{s=r}^k (J^m)_{ls} \sum_j T_{sj} + \sum_{s=r+1}^k (J^m)_{ls} \sum_j T_{sj} \right]
\]

\[
= \sum_l \left( T^{-1} \right)_{il} \sum_{s=r}^k \delta_{ls} \sum_j T_{sj} + 0 \quad \text{(according to Corollary 3.1)}
\]

Which proves the lemma.

Proof of Lemma 3.4.

For a \((2 \times 2)\) probability matrix \( P = \begin{pmatrix} c & 1-c \\ d & 1-d \end{pmatrix} \), the discriminant of the characteristic equation can be expressed as \( \Delta = (1 - c + d)^2 \).

In case \( \Delta > 0 \), the matrix \( P \) has two different eigenvalues \( \lambda_1 = 1 \) and \( \lambda_2 \neq 1 \) so that \( P = T^{-1} \times D \times T \) with \( D = \begin{pmatrix} 1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \).

In case \( \Delta = 0 \) holds \( c - d = 1 \). This condition is only fulfilled for the probability matrix \( P = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \), that is diagonalizable.

Which proves the Lemma.

Acknowledgements

The author thanks the reviewers for their remarks and valuable suggestions.

References


