J. Appl. Prob. **50,** 918–930 (2013)
Printed in England

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 issue 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 the case of nonuniqueness, a procedure is introduced to identify a transition matrix that takes into account the specificity of the concrete context. In the case of nonexistence of a probability root, the concept of an 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

2010 Mathematics Subject Classification: Primary 15A23; 15B51; 60J10

1. Introduction

A discrete-time Markov chain is considered for which the tth 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 $\boldsymbol{P} = (p_{ij})$ 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 $i, j \in \{1, \ldots, k\}$) from state i to state j in one period of time can be estimated by the maximum likelihood estimator ([1, p. 113]):

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

Received 3 July 2008; revision received 6 February 2013.
* Postal address: Department of Business Technology and Operations, Vrije Universiteit Brussel, Pleinlaan 2, B-1050 Brussels, Belgium. Email address: marie-anne.guerry@vub.ac.be

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 $t = 0, \ldots, T-1$).

The estimations for the transition probabilities according to (1.1) are based on data for the observable variables regarding time intervals that correspond to the time unit of the Markov chain. There is considered to be a lack of data in the case where there are no observations available for some variables or where there are no observations available regarding the time unit of the Markov chain. In previous work, the problem of a lack of observations for some variables is dealt with by building a hidden Markov model or a Markov switching model (see [5], [6], and [10]) 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 have estimations for the transition probabilities for intervals with length 0.5. In that situation there is a lack of data of the type which is dealt with 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 in more detail for discrete-time Markov chains.

An answer to problems of this kind can be useful; among other fields, one example is manpower planning. A Markov model for manpower planning takes into account internal transitions (e.g. promotions), outgoing flows (i.e. wastage) as well as 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. If 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), then the methodology presented in this paper provides some interesting insights. Namely for this company the promotion probabilities for a semester can be quantified, and 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 the case where 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 there exists more than one probability square root for a transition matrix. In answer to the identification problem, a desirable probability square root can be selected based on presented criteria. Section 6 provides 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 $\boldsymbol{P}(1) = (p_{ij}(1))$ and the matrix of the transition probabilities, for a time interval with length 0.5, by $\boldsymbol{P}(0.5) = (p_{ij}(0.5))$. Under the assumption of time homogeneity, the following

https://doi.org/10.1239/jap/1389370090 Published online by Cambridge University Press

relation holds

$$p_{ij}(1) = \sum_m p_{im}(0.5)p_{mj}(0.5) \quad \text{for all } i, j \in \{1, \ldots, k\}.$$

Or in matrix notation

$$\boldsymbol{P}(1) = \boldsymbol{P}(0.5) \times \boldsymbol{P}(0.5).$$

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

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

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

$$\Pi = \left\{ \boldsymbol{M} = (m_{ij}) \in \mathbb{R}^{k \times k} \;\middle|\; \sum_{j=1}^{k} m_{ij} = 1 \text{ and } m_{ij} \geq 0 \text{ for all } i, j \in \{1, \ldots, k\} \right\}.$$

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 [11] and [12]. Deciding whether the observed matrix $\hat{\boldsymbol{P}}(1)$ can be represented in the form $\boldsymbol{A}^2$ for some probability matrix $\boldsymbol{A}$ 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 $\hat{\boldsymbol{P}}(1)$. When 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 respectively called the embedding, the inverse, and the identification problem [11].

3. Probability square roots

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

Definition 3.1. For a probability matrix $\boldsymbol{P}$, a probability square root $\boldsymbol{A}$ of $\boldsymbol{P}$ is a probability matrix $\boldsymbol{A}$ that is a square root of $\boldsymbol{P}$.

For a (2×2) probability matrix

$$\boldsymbol{P} = \begin{pmatrix} c & 1-c \\ d & 1-d \end{pmatrix},$$

the probability matrix $\boldsymbol{A} = (a_{ij})$ is a probability square root of $\boldsymbol{P}$ if and only if

$$\begin{aligned} a_{11}^2 + (1 - a_{11})a_{21} &= c, \\ a_{11}a_{21} + (1 - a_{21})a_{21} &= d. \end{aligned} \tag{3.1}$$

https://doi.org/10.1239/jap/1389370090 Published online by Cambridge University Press

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

Theorem 3.1. *Let $\boldsymbol{P}$ be a (2×2) probability matrix,*

$$\boldsymbol{P} = \begin{pmatrix} c & 1-c \\ d & 1-d \end{pmatrix}.$$

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

$$a_{11} = \frac{\sqrt{c-d}(1-c)+d}{1-c+d}, \qquad a_{21} = d\frac{1-\sqrt{c-d}}{1-c+d}.$$

Moreover, in the case where

$$a_{11} = \frac{\sqrt{c-d}(c-1)+d}{1-c+d}, \qquad a_{21} = d\frac{1+\sqrt{c-d}}{1-c+d}$$

are both elements of $[0, 1]$, there exists a second probability square root $\boldsymbol{A} = (a_{ij})$ of $\boldsymbol{P}$.

- *If $c > d$ and $1 - c + d = 0$, both*

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \textit{and} \quad \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

are probability square roots of $\boldsymbol{P}$.

For a proof of Theorem 3.1, see Appendix A.

Note that three possible situations effectively occur: a (2×2) probability matrix can have no, exactly one, or two, probability square roots. For example, for the case $c > d$,

$$\boldsymbol{P} = \begin{pmatrix} \frac{5}{8} & \frac{3}{8} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

has two probability square roots and

$$\boldsymbol{P} = \begin{pmatrix} \frac{11}{27} & \frac{16}{27} \\ \frac{1}{9} & \frac{8}{9} \end{pmatrix}$$

has one probability square root.

Theorem 3.1 implicitly provides, for a (2×2) probability matrix $\boldsymbol{P}$, a necessary and sufficient condition for there to be a probability square root: for a (2×2) probability matrix

$$\boldsymbol{P} = \begin{pmatrix} c & 1-c \\ d & 1-d \end{pmatrix},$$

https://doi.org/10.1239/jap/1389370090 Published online by Cambridge University Press

there exists a probability square root of $\boldsymbol{P}$ if and only if $c \geq d$. This condition can be formulated equivalently in terms of the trace of $\boldsymbol{P} = \boldsymbol{A} \times \boldsymbol{A}$: $\mathrm{Tr}(\boldsymbol{P}) = \mathrm{Tr}(\boldsymbol{A} \times \boldsymbol{A}) \geq 1$.

In the study of necessary conditions for a $(k \times k)$ probability matrix to have a probability square root $\boldsymbol{A}$, the following theorem can be useful since it formulates some properties of $\mathrm{Tr}(\boldsymbol{A} \times \boldsymbol{A})$.

Theorem 3.2. *Within the set of* $(k \times k)$ *probability matrices* $\boldsymbol{A}$*, for a critical point the equality* $\mathrm{Tr}(\boldsymbol{A} \times \boldsymbol{A}) = 1$ *holds.*

Proof. For a $(k \times k)$ probability matrix $\boldsymbol{A} = (a_{ij})$, we have

$$\begin{aligned}\mathrm{Tr}(\boldsymbol{A} \times \boldsymbol{A}) &= \sum_l (\boldsymbol{A} \times \boldsymbol{A})_{ll} \\ &= \sum_l \left[(a_{ll})^2 + \sum_{i \neq l} a_{li} a_{il} \right] \\ &= \sum_l \left[\left(1 - \sum_{i \neq l} a_{li} \right)^2 + \sum_{i \neq l} a_{li} a_{il} \right].\end{aligned}$$

In this way, $\mathrm{Tr}(\boldsymbol{A} \times \boldsymbol{A})$ is expressed as a function f of the $(k-1)^2$ elements a_{rs}, with $r \neq s \in \{1, \ldots, k\}$, of the probability matrix $\boldsymbol{A}$. Whereby a critical point of f satisfies, for all $s \neq r \in \{1, \ldots, k\}$,

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

Therefore, for a critical point of f, we have $\sum_{i \neq r} a_{ri} + a_{sr} = 1$ for all $s \neq r \in \{1, \ldots, k\}$. Consequently, for $r \in \{1, \ldots, k\}$ and $s \neq r$, all the elements a_{sr} are equal to $a_{sr} = 1 - \sum_{i \neq r} a_{ri} = a_{rr}$. This results in the fact that the critical point of f corresponds with a matrix $\boldsymbol{A}$ having all the elements of the rth column equal, let us say equal to a_{1r}. This property gives rise to

$$\mathrm{Tr}(\boldsymbol{A} \times \boldsymbol{A}) = \sum_l \left[(a_{1l})^2 + \sum_{i \neq l} a_{1i} a_{1l} \right] = \sum_l a_{1l} \left[a_{1l} + \sum_{i \neq l} a_{1i} \right] = \sum_l a_{1l} = 1,$$

which proves the theorem.

As specified in [7], for the specific class of (3×3) and (4×4) state-wise monotone probability matrices $\boldsymbol{A}$, $\mathrm{Tr}(\boldsymbol{A} \times \boldsymbol{A}) \geq 1$ holds, resulting in the necessary condition $\mathrm{Tr}(\boldsymbol{P}) \geq 1$ for a probability matrix $\boldsymbol{P}$ of order (3×3) or (4×4) to have a state-wise monotone probability square root. However, without requiring any restriction on the probability root(s), $\mathrm{Tr}(\boldsymbol{P}) \geq 1$ is not a necessary condition for $\boldsymbol{P}$ to have a probability root: $\boldsymbol{P}$ can have a probability root $\boldsymbol{A}$, while $\mathrm{Tr}(\boldsymbol{P}) < 1$, as is the case for

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

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

https://doi.org/10.1239/jap/1389370090 Published online by Cambridge University Press

restrictive set

$$\Gamma = \left\{ \boldsymbol{M} = (m_{ij}) \in \mathbb{R}^{k\times k} \;\middle|\; \sum_{j=1}^{k} m_{ij} = 1 \text{ for all } i \in \{1,\ldots,k\} \right\} \supset \Pi.$$

Definition 3.2. 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 0,

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

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

$$\boldsymbol{u} \times (\boldsymbol{P} - \lambda \boldsymbol{I})^n \in H_0 \quad \Rightarrow \quad \boldsymbol{u} \in H_0$$

holds.

For a proof of Lemma 3.1, see Appendix A.

In what follows the terminology 'eigenvector' refers to a left eigenvector: the $(1 \times k)$ vector $\boldsymbol{E}$ is an eigenvector of the $(k \times k)$ matrix $\boldsymbol{P}$ associated with the eigenvalue λ if and only if $\boldsymbol{E} \times \boldsymbol{P} = \lambda \boldsymbol{P}$.

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 $\boldsymbol{P}$, a (generalized) eigenvector $\boldsymbol{E}$ associated with the eigenvalue λ satisfies $\boldsymbol{E} \times (\boldsymbol{P} - \lambda \boldsymbol{I})^n = 0$ for some $n \in \{1, 2, 3, \ldots\}$. Therefore, $\boldsymbol{E} \times (\boldsymbol{P} - \lambda \boldsymbol{I})^n$ is an element of H_0 and according to Lemma 3.1 the (generalized) eigenvector $\boldsymbol{E}$ itself is an element of H_0. This proves the corollary.

As examined in previous work (e.g. [4]), it has been proved, under some conditions, that for a $(k \times k)$ row-normalized matrix $\boldsymbol{P} = \boldsymbol{T}^{-1} \times \boldsymbol{J} \times \boldsymbol{T}$ with Jordan matrix $\boldsymbol{J}$, the transformations $f_m : \Gamma \to \Gamma : \boldsymbol{T}^{-1} \times \boldsymbol{J} \times \boldsymbol{T} \to \boldsymbol{T}^{-1} \times \boldsymbol{J}^m \times \boldsymbol{T}$ (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 $\boldsymbol{P} = \boldsymbol{T}^{-1} \times \boldsymbol{J} \times \boldsymbol{T}$ with Jordan matrix $\boldsymbol{J}$, the row sums of $\boldsymbol{T}^{-1} \times \boldsymbol{J}^m \times \boldsymbol{T}$ (as far as this matrix is defined) do not depend on the value of $m \in \mathbb{R}$.*

For a proof of Lemma 3.2, see Appendix A.

Lemma 3.3. *For a $(k \times k)$ row-normalized matrix $\boldsymbol{P} = \boldsymbol{T}^{-1} \times \boldsymbol{J} \times \boldsymbol{T}$ with Jordan matrix $\boldsymbol{J}$, $\boldsymbol{A} = \boldsymbol{T}^{-1} \times \boldsymbol{J}^{1/2} \times \boldsymbol{T}$ (as far as this matrix is a defined element of $\mathbb{R}^{k\times k}$) is a row-normalized matrix for which $\boldsymbol{P} = \boldsymbol{A} \times \boldsymbol{A}$.*

Proof. It is clear that for $\boldsymbol{A} = \boldsymbol{T}^{-1} \times \boldsymbol{J}^{1/2} \times \boldsymbol{T}$ we have $\boldsymbol{P} = \boldsymbol{A} \times \boldsymbol{A}$.

Moreover, since $\boldsymbol{P}$ is a row-normalized matrix, the row sums of $\boldsymbol{T}^{-1} \times \boldsymbol{J} \times \boldsymbol{T}$ are equal to 1. According to Lemma 3.2, the row sums of $\boldsymbol{T}^{-1} \times \boldsymbol{J} \times \boldsymbol{T}$ and $\boldsymbol{T}^{-1} \times \boldsymbol{J}^{1/2} \times \boldsymbol{T}$ are equal. Therefore, the matrix $\boldsymbol{T}^{-1} \times \boldsymbol{J}^{1/2} \times \boldsymbol{T}$ is row-normalized as well.

https://doi.org/10.1239/jap/1389370090 Published online by Cambridge University Press

Corollary 3.2. *For a $(k \times k)$ diagonalizable row-normalized matrix $\boldsymbol{P} = \boldsymbol{T}^{-1} \times \boldsymbol{D} \times \boldsymbol{T}$, with all eigenvalues positive and real, $\boldsymbol{A} = \boldsymbol{T}^{-1} \times \boldsymbol{D}^{1/2} \times \boldsymbol{T}$ is a row-normalized matrix for which $\boldsymbol{P} = \boldsymbol{A} \times \boldsymbol{A}$.*

Proof. Since, for a diagonalizable matrix $\boldsymbol{P}$, the Jordan matrix is a diagonal matrix $\boldsymbol{D}$ with diagonal elements that are the eigenvalues of $\boldsymbol{P}$, Lemma 3.3 is applicable for $\boldsymbol{J} = \boldsymbol{D}$ and, therefore, this proves the corollary.

In what follows the more restrictive set Π of probability matrices is examined. In particular, properties of square roots of probability matrices are discussed.

Lemma 3.4. *Each (2×2) probability matrix $\boldsymbol{P}$ is diagonalizable.*

For a proof of Lemma 3.4, see Appendix A.

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

- *In the case where $k = 2$, the matrix $\boldsymbol{T}^{-1} \times \boldsymbol{J}^{1/2} \times \boldsymbol{T}$ is nonnegative.*
- *In the case where $k > 2$, the matrix $\boldsymbol{T}^{-1} \times \boldsymbol{J}^{1/2} \times \boldsymbol{T}$ is not necessarily nonnegative.*

Proof. For the case where $k = 2$, according to Lemma 3.4, the Jordan matrix of the (2×2) probability matrix $\boldsymbol{P}$ is a diagonal matrix $\boldsymbol{D}$. In the situation that $\lambda_1 = \lambda_2 = 1$, the diagonal matrix $\boldsymbol{D}$ equals the identity matrix $\boldsymbol{I}_{2\times 2}$ and, consequently, $\boldsymbol{T}^{-1} \times \boldsymbol{D}^{1/2} \times \boldsymbol{T} = \boldsymbol{T}^{-1} \times \boldsymbol{T} = \boldsymbol{I}_{2\times 2}$ is a nonnegative matrix.

In the situation where $\lambda_1 = 1$ and $\lambda_2 \neq 1$, let us denote by $\boldsymbol{E}_1 = (a \;\; b)$ an eigenvector of $\boldsymbol{P}$ associated with the eigenvalue $\lambda_1 = 1$. According to Corollary 3.1 it is known that $\boldsymbol{E}_2 = (1 \;\; -1)$ is an eigenvector of $\boldsymbol{P}$ associated with the eigenvalue $\lambda_2 \neq 1$. Consequently, $\boldsymbol{P} = \boldsymbol{T}^{-1} \times \boldsymbol{D} \times \boldsymbol{T}$ with

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

This results in

$$\boldsymbol{P} = \boldsymbol{T}^{-1} \times \boldsymbol{D} \times \boldsymbol{T} = \frac{1}{a+b} \begin{pmatrix} a + b\lambda_2 & b - b\lambda_2 \\ a - a\lambda_2 & b + a\lambda_2 \end{pmatrix},$$
$$\boldsymbol{T}^{-1} \times \boldsymbol{D}^{1/2} \times \boldsymbol{T} = \frac{1}{a+b} \begin{pmatrix} a + b\sqrt{\lambda_2} & b - b\sqrt{\lambda_2} \\ a - a\sqrt{\lambda_2} & b + a\sqrt{\lambda_2} \end{pmatrix}.$$

Since the eigenvalues of the probability matrix $\boldsymbol{P} = (p_{ij})$ are assumed to be real, according to the Perron–Frobenius theorem, the eigenvalue λ_2 is less than 1. Therefore, $p_{12} \geq 0$ implies that $b/(a+b) \geq 0$ and from $p_{21} \geq 0$ it is known that $a/(a+b) \geq 0$. Consequently, $(\boldsymbol{T}^{-1} \times \boldsymbol{D}^{1/2} \times \boldsymbol{T})_{12} \geq 0$ and $(\boldsymbol{T}^{-1} \times \boldsymbol{D}^{1/2} \times \boldsymbol{T})_{21} \geq 0$. Moreover, $\sqrt{\lambda_2} \geq \lambda_2$ and therefore we have

$$(\boldsymbol{T}^{-1} \times \boldsymbol{D}^{1/2} \times \boldsymbol{T})_{11} \geq (\boldsymbol{T}^{-1} \times \boldsymbol{D} \times \boldsymbol{T})_{11} = p_{11} \geq 0,$$
$$(\boldsymbol{T}^{-1} \times \boldsymbol{D}^{1/2} \times \boldsymbol{T})_{22} \geq (\boldsymbol{T}^{-1} \times \boldsymbol{D} \times \boldsymbol{T})_{22} = p_{22} \geq 0,$$

which proves that the elements of $\boldsymbol{T}^{-1} \times \boldsymbol{D}^{1/2} \times \boldsymbol{T}$ are all nonnegative.

https://doi.org/10.1239/jap/1389370090 Published online by Cambridge University Press

In the case where $k > 2$, for the diagonalizable (3×3) probability matrix

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

with eigenvalues 1, $\frac{1}{4}$, and $\frac{1}{36}$ all positive and real, it holds that the matrix

$$\boldsymbol{T}^{-1} \times \boldsymbol{D}^{1/2} \times \boldsymbol{T} = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ -\frac{1}{6} & 1 & \frac{1}{6} \end{pmatrix}$$

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

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

- *In the case where $k = 2$, the matrix $\boldsymbol{T}^{-1} \times \boldsymbol{D}^{1/2} \times \boldsymbol{T}$ is a probability square root of $\boldsymbol{P}$.*
- *In the case where $k > 2$, the matrix $\boldsymbol{T}^{-1} \times \boldsymbol{D}^{1/2} \times \boldsymbol{T}$ is a row-normalized square root of $\boldsymbol{P}$.*

Proof. According to Corollary 3.2, for a diagonalizable $(k \times k)$ probability matrix $\boldsymbol{P} = \boldsymbol{T}^{-1} \times \boldsymbol{D} \times \boldsymbol{T}$, with all eigenvalues positive and real, $\boldsymbol{T}^{-1} \times \boldsymbol{D}^{1/2} \times \boldsymbol{T}$ is a row-normalized square root of $\boldsymbol{P}$.

Moreover, in the case where $k = 2$, according to Theorem 3.3, $\boldsymbol{T}^{-1} \times \boldsymbol{D}^{1/2} \times \boldsymbol{T}$ is a nonnegative matrix. Therefore, $\boldsymbol{T}^{-1} \times \boldsymbol{D}^{1/2} \times \boldsymbol{T}$ is a probability square root of $\boldsymbol{P}$, 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 $\boldsymbol{P}$, there can either exist no, exactly one, or more than one, probability square root $\boldsymbol{A}$ of $\boldsymbol{P}$.

In the case where no probability square root exists, we may ask the following question. Based on what method can a probability matrix $\boldsymbol{B}$ be determined such that $\boldsymbol{B} \times \boldsymbol{B}$ 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 factorisation deals in general with the following type of problem: given a nonnegative matrix $\boldsymbol{U}$, find nonnegative factors $\boldsymbol{V}$ and $\boldsymbol{W}$ such that $\boldsymbol{U} \approx \boldsymbol{V} \times \boldsymbol{W}$ (see, [3] and [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

https://doi.org/10.1239/jap/1389370090 Published online by Cambridge University Press

the Euclidean distance between the matrices $\boldsymbol{R} = (r_{ij})$ and $\boldsymbol{S} = (s_{ij})$ is given by

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

In the more specific context of finding approximate probability square roots, this approach results in the following formulation: given a probability matrix $\boldsymbol{P}$ of order $(k \times k)$, find a probability matrix $\boldsymbol{B}$ of order $(k \times k)$ such that $\boldsymbol{P} \approx \boldsymbol{B} \times \boldsymbol{B}$. That is, find a $(k \times k)$ matrix $\boldsymbol{B} = (b_{ij})$ solution of the following optimization problem

$$\|\boldsymbol{P} - \boldsymbol{B} \times \boldsymbol{B}\|^2 = \min \|\boldsymbol{P} - \boldsymbol{A} \times \boldsymbol{A}\|^2_{\boldsymbol{A} \in \Pi}$$

subject to the constraints

$$\sum_j b_{ij} = 1 \quad \text{and} \quad b_{ij} \geq 0, \qquad \text{for all } i, j.$$

This formulation gives rise to an optimization problem of the Karush–Kuhn–Tucker type, [2]. A solution $\boldsymbol{B}$ of this optimization problem results in an approximate probability square root of $\boldsymbol{P}$.

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 $\hat{\boldsymbol{P}}(1)$, estimations for the number of objects in each state after each time interval of length 0.5. In the case where $\hat{\boldsymbol{P}}(1)$ has more than one (approximate) probability square root, the observations are consistent with more than one discrete-time Markov chain with time unit 0.5. The problem then is how to select from within this set of alternatives. Singer and Spilerman have dealt with this identification problem in general, [11].

By Theorem 3.1 it is proved that there exist (2×2) probability matrices that have more than one probability square root. The nonuniqueness of probability square roots is a more general fact affecting certain probability matrices of any order $(k \times k)$, $k \geq 2$: for $\boldsymbol{Q}$ a (2×2) probability matrix having two probability square roots $\hat{\boldsymbol{A}} \neq \tilde{\boldsymbol{A}}$ and for $\boldsymbol{I}$ the identity matrix of order $((k-2) \times (k-2))$, the block diagonal matrix $\boldsymbol{P} = \text{diag}(\boldsymbol{Q}, \boldsymbol{I})$ is a $(k \times k)$ probability matrix having at least two probability roots, namely $\text{diag}(\hat{\boldsymbol{A}}, \boldsymbol{I})$ and $\text{diag}(\tilde{\boldsymbol{A}}, \boldsymbol{I})$. Since, for each $k \geq 2$, there exist $(k \times k)$ probability matrices with more than one probability square root, it is worth dealing with the identification problem, [11].

In the context of this paper, in the case where 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 each of the states. The criterion for selecting the most desirable probability square root can vary depending on the phenomenon that is modelled by the Markov chain. We could prefer to have an evolution in time of the vector $\boldsymbol{n}(t) = (n_i(t))$ such that, for each state i, there is the smallest possible fluctuation of $n_i(t)$, or, moreover, the evolution of $n_i(t)$ decreases or increases monotonically. It could be preferable that the expected number of objects after half a time period, $\boldsymbol{n}(t) \times \hat{\boldsymbol{P}}(0.5)$, is 'somewhere in between' the starting vector $\boldsymbol{n}(t)$ and the expected vector after one period of time $\boldsymbol{n}(t) \times \hat{\boldsymbol{P}}(1)$. This condition can be expressed in terms of the intervals

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

https://doi.org/10.1239/jap/1389370090 Published online by Cambridge University Press

as follows:

$$(\boldsymbol{n}(t) \times \hat{\boldsymbol{P}}(0.5))_i \in I_i \quad \text{for all } i \in \{1, \ldots, k\}.$$

Such a formulated condition can be, for example, useful in modelling 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 monotonically.

In the case where there does not exist a (approximate) probability square root for which this property holds, the criterion to select $\hat{\boldsymbol{P}}(0.5)$ can then be formulated in terms of an optimization problem, namely: $\hat{\boldsymbol{P}}(0.5)$ is a (approximate) probability square root of $\hat{\boldsymbol{P}}(1)$ minimizing the maximum value, over the different states $i \in \{1, \ldots, k\}$, of the discrepancy between the expected number of members $(\boldsymbol{n}(t) \times \hat{\boldsymbol{P}}(0.5))_i$ and the elements of the interval I_i, i.e. $\hat{\boldsymbol{P}}(0.5)$ is a (approximate) probability square root minimizing

$$\max_{i \in \{1,\ldots,k\}} \min_{x \in I_i} |(\boldsymbol{n}(t) \times \hat{\boldsymbol{P}}(0.5))_i - x|.$$

6. Illustration

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

$$\hat{\boldsymbol{P}}(1) = \begin{pmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

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

$$\boldsymbol{A}_1 = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad \boldsymbol{A}_2 = \begin{pmatrix} \frac{5}{6} & \frac{1}{6} \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix}.$$

For the probability square roots $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ the corresponding evolution of the expected number of members is as follows:

$$\begin{aligned} &\boldsymbol{n}(0) = (60 \;\; 42), \quad \boldsymbol{n}(0) \times \boldsymbol{A}_1 = (72 \;\; 30), \quad \boldsymbol{n}(1) = \boldsymbol{n}(0) \times \boldsymbol{A}_1^2 = (66 \;\; 36), \cdots \\ &\boldsymbol{n}(0) = (60 \;\; 42), \quad \boldsymbol{n}(0) \times \boldsymbol{A}_2 = (64 \;\; 38), \quad \boldsymbol{n}(1) = \boldsymbol{n}(0) \times \boldsymbol{A}_2^2 = (66 \;\; 36), \cdots. \end{aligned}$$

The probability square root $\boldsymbol{A}_2$ results in a monotonic 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 unit 1 and transition matrix $\boldsymbol{P}(1)$, the study is focused on finding a Markov chain with time unit 0.5 that is compatible with $\boldsymbol{P}(1)$. The stated problem and the results are therefore formulated in terms of probability square roots $\boldsymbol{P}(0.5)$. 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 $\boldsymbol{P}(1/m)$ for $m \in \mathbb{N}_0$ of a Markov chain with time unit $1/m$ and that is compatible with $\boldsymbol{P}(1)$.

https://doi.org/10.1239/jap/1389370090 Published online by Cambridge University Press

Appendix A.

Proof of Theorem 3.1. In the special case of a probability matrix

$$\boldsymbol{P} = \begin{pmatrix} c & 1-c \\ d & 1-d \end{pmatrix}$$

satisfying $1 - c + d = 0$, the matrix $\boldsymbol{P}$ equals

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

and has two probability square roots, namely

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

In the special case of a probability matrix $\boldsymbol{A} = (a_{ij})$ with $a_{11} = 1$, the matrix $\boldsymbol{A}$ can only be a square root of a probability matrix of the form

$$\boldsymbol{P} = \begin{pmatrix} 1 & 0 \\ d & 1-d \end{pmatrix}.$$

Under these conditions, and according to (3.1), $\boldsymbol{P}$ has exactly one probability square root namely the probability matrix $\boldsymbol{A} = (a_{ij})$ with $a_{11} = 1$ and $a_{21} = 1 - \sqrt{1-d}$.

The further reasoning can be restricted to the situation of a probability matrix

$$\boldsymbol{P} = \begin{pmatrix} c & 1-c \\ d & 1-d \end{pmatrix}$$

satisfying $1 - c + d \neq 0$ and a probability square root $\boldsymbol{A} = (a_{ij})$ with $a_{11} \neq 1$: the system (3.1) results in the following quadratic equation in a_{11}

$$(1-c+d)a_{11}^2 - (2d)a_{11} + c^2 - c + d = 0 \tag{A.1}$$

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

Therefore, in the case where $c < d$, no probability square root exists for the matrix $\boldsymbol{P}$. In the case where $c = d$, the matrix $\boldsymbol{P}$ has exactly one probability square root $\boldsymbol{A}$. According to (3.1), this matrix $\boldsymbol{A}$ satisfies $a_{11} = a_{21} = c = d$. In the case where $c > d$, the equation (A.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 $a_{21} = (c - a_{11}^2)/(1 - a_{11})$ 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 $a_{11} = (\sqrt{c-d}(1-c)+d)/(1-c+d)$ and $a_{21} = d(1-\sqrt{c-d})/(1-c+d)$ are elements of $[0, 1]$ there exists at least one probability square root $\boldsymbol{A} = (a_{ij})$ of $\boldsymbol{P}$, which proves the theorem.

https://doi.org/10.1239/jap/1389370090 Published online by Cambridge University Press

Proof of Lemma 3.1. The property formulated in Lemma 3.1 can be proved by induction. For a row-normalized matrix $\boldsymbol{P}$ and $\boldsymbol{u} \in \mathbb{R}^k$ satisfying $\boldsymbol{u} \times (\boldsymbol{P} - \lambda \boldsymbol{I}) \in H_0$, we have

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

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

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

Proof of Lemma 3.2. For a probability matrix $\boldsymbol{P}$ the eigenvalue $\lambda = 1$ is a semisimple eigenvalue, i.e. the Jordan blocks are of order (1×1) [9]. Therefore, the Jordan blocks corresponding to $\lambda = 1$ result in the $(r \times r)$ identity matrix when 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 $J_{ls} = \delta_{ls}$ (for all $s \in \{1, \ldots, r\}$ and $l \in \{1, \ldots, k\}$), where the notation δ stands for the Kronecker delta. The matrix $\boldsymbol{T}$ is then a matrix with, in the first r rows, eigenvectors associated with $\lambda = 1$ and in each other row a (generalized) eigenvector associated with an eigenvalue different from 1. Therefore, for all $i \in \{1, \ldots, k\}$, we have

$$\begin{aligned}\sum_j (\boldsymbol{T}^{-1} \times \boldsymbol{J}^m \times \boldsymbol{T})_{ij} &= \sum_j \sum_l \sum_s (\boldsymbol{T}^{-1})_{il} (\boldsymbol{J}^m)_{ls} T_{sj} \\ &= \sum_l (\boldsymbol{T}^{-1})_{il} \left[\sum_{s=1}^{s=r} (\boldsymbol{J}^m)_{ls} \sum_j T_{sj} + \sum_{s=r+1}^{s=k} (\boldsymbol{J}^m)_{ls} \sum_j T_{sj} \right] \\ &= \sum_l (\boldsymbol{T}^{-1})_{il} \sum_{s=1}^{s=r} \delta_{ls} \sum_j T_{sj} + 0,\end{aligned}$$

according to Corollary 3.1, which proves the lemma.

Proof of Lemma 3.4. For a (2×2) probability matrix

$$\boldsymbol{P} = \begin{pmatrix} c & 1-c \\ d & 1-d \end{pmatrix},$$

the discriminant of the characteristic equation can be expressed as $\Delta = (1 - c + d)^2$. In the case where $\Delta > 0$, the matrix $\boldsymbol{P}$ has two different eigenvalues $\lambda_1 = 1$ and $\lambda_2 \neq 1$ so that $\boldsymbol{P} = \boldsymbol{T}^{-1} \times \boldsymbol{D} \times \boldsymbol{T}$ with

$$\boldsymbol{D} = \begin{pmatrix} 1 & 0 \\ 0 & \lambda_2 \end{pmatrix}.$$

https://doi.org/10.1239/jap/1389370090 Published online by Cambridge University Press

In the case where $\Delta = 0$, $c - d = 1$ holds. This condition is only fulfilled for the probability matrix

$$\boldsymbol{P} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

which is diagonalizable. This proves the lemma.

Acknowledgement

The author would like to thank the reviewers for their remarks and valuable suggestions.

References

[1] Bartholomew, D. J., Forbes, A. F. and McClean, S. I. (1991). *Statistical Techniques for Manpower Planning*, 2nd edn. John Wiley, Chichester.
[2] Bazaraa, M. S., Sherali, H. D. and Shetty, C. M. (2006). *Nonlinear Programming: Theory and Algorithms*, 3rd edn. John Wiley, Hoboken, NJ.
[3] Berry, M. W. *et al.* (2007). Algorithms and applications for approximate nonnegative matrix factorization. *Comput. Statist. Data Anal.* **52,** 155–173.
[4] Butler, J. C. (1981). Effect of various transformations on the analysis of percentage data. *Math. Geology* **13,** 53–68.
[5] Cappé, O., Moulines, E. and Rydén, T. (2005). *Inference in Hidden Markov Models*. Springer, New York.
[6] Guerry, M.-A. (2011). Hidden heterogeneity in manpower systems: a Markov-switching model approach. *Europ. J. Operat. Res.* **210,** 106–113.
[7] Guerry, M.-A. (2012). Necessary conditions for the embeddability of discrete-time state-wise monotone Markov chains. Working paper MOSI/53.
[8] Lee, D. D. and Seung, H. S. (2001). Algorithms for non-negative matrix factorization. *Adv. Neural Inf. Processing Systems* **13,** 556–562.
[9] Minc, H. (1988). *Nonnegative Matrices*. John Wiley, New York.
[10] Moon, S., Kamakura, W. A. and Ledolter, J. (2007). Estimating promotion response when competitive promotions are unobserved. *J. Marketing Res.* **44,** 503–515.
[11] Singer, B. and Spilerman, S. (1973/1974). Social mobility models for heterogeneous populations. *Sociological Method.* **5,** 356–401.
[12] Singer, B. and Spilerman, S. (1976). The representation of social processes by Markov models. *Amer. J. Sociology* **82,** 1–54.
[13] Ugwuowo, F. I. and McClean, S. (2000). Modelling heterogeneity in a manpower system: a review. *Appl. Stoch. Models Business Industry* **16,** 99–110.