Noname manuscript No.
(will be inserted by the editor)

Markov models for duration-dependent transitions: Selecting the states using duration values or duration intervals?

Received: date / Accepted: 29 March 2022

Abstract In a Markov model the transition probabilities between states do not depend on the time spent in the current state. The present paper explores two ways of selecting the states of a discrete-time Markov model for a system partitioned into categories where the duration of stay in a category affects the probability of transition to another category. For a set of panel data, we compare the likelihood fits of the Markov models with states based on duration intervals and with states defined by duration values. For hierarchical systems, we show that the model with states based on duration values has a better maximum likelihood fit than the baseline Markov model where the states are the categories. We also prove that this is not the case for the duration-interval model, under conditions on the data that seem realistic in practice. Furthermore, we use the Akaike and Bayesian information criteria to compare these alternative Markov models. The theoretical findings are illustrated by an analysis of a real-world personnel data set.

Keywords Markov chain · maximum likelihood · duration of stay · model selection

1 Introduction

Markov models are widely employed to analyze transitions in a system partitioned into categories. For instance, in a credit risk study the system consists of bond issuing companies and the categories represent rating levels given to these companies by rating agencies such as Fitch, Moody's and Standard & Poor's (see D'Amico et al., 2006); in a manpower planning model the system under study is a firm and the categories are the job levels or grades in that firm (see Bartholomew et al., 1991). Systems, for instance manpower systems,

can be of hierarchical nature so that the categories are ordered and transitions within the system only occur from a category to the next level category (Bányai et al., 2018).

Markov chain models are suitable for aggregated analysis at the level of the states and hence assume that the transition probabilities for any given state apply to all system units in that state. In such models the expected state frequencies are calculated using the multinomial distribution where the estimated transition probabilities are parameters. To obtain valid results, it is thus required that all system units in a same state have similar transition propensity. Bartholomew et al. (1991) point out that the state space of the Markov chain should be chosen so as to reflect this homogeneity of system units regarding the transition probabilities. Uche (1990) examines the effect of heterogeneity on Markovian analysis and stresses the need to disaggregate system units into homogeneous subgroups to which the Markovian analysis can be applied. It is therefore necessary to strive, in the state selection phase of a Markov model, for homogeneity of these states regarding the transition probabilities (Bartholomew et al., 1991; De Feyter, 2006; Rombaut and Guerry, 2015).

A common instance where the transition homogeneity of states is not fulfilled is when transitions between states depend on the duration of stay in the originating state. For example, in manpower systems promotion probabilities may depend on length of service in the grade and in a health context some diseases are such that the probability of recovering from these depends on the duration of past episodes of the disease (Patten, 2005). Ugwuowo and McClean (2000) provide an overview of modelling approaches that account for the heterogeneity of personnel in a manpower system. Observable as well as latent sources of heterogeneity are tackled and the duration variable "length of service" is explicitly mentioned as an important observable source of heterogeneity.

It is well known that transitions depending on the duration of stay (henceforth called “DS-transitions”) can be addressed by semi-Markov models that allow for various distributional forms of the time spent in a state prior to a transition to an other state. For more details on semi-Markov models, we refer to Barbu and Limnios (2008). Dewar et al. (2012) study DS-transitions in a hidden semi-Markov model through explicit parameterization and inference of state duration distributions. Cartella et al. (2015) build a hidden semi-Markov model for predicting maintenance and determine the optimal number of hidden states as well as the optimal sojourn time distribution.

However, Markov models are of lesser complexity than semi-Markov models and thus are less data demanding and more attractive to use in the eyes of a decision maker. For example in a manpower planning context they are transparent and easy to understand for practitioners (Parker and Caine, 1996). It thus seems natural and valuable to look for ways to build DS-transitions into the Markov chain framework.

In previous research DS-transitions in Markov models have been approached in a variety of ways. Some empirical studies follow a stratification approach,

where the population is divided into age subgroups and separate Markov models are set up for each of the subgroups (see Jiang and Sinha, 1989; Longini et al., 1991; Duffy et al., 1997). A drawback of these approaches is that they do not allow for long-run analyses for the population as a whole since transitions between the age subgroups are not captured in the model making an analysis across age subgroups impossible. Instead, based on the Markov chains for each of the respective subgroups, the population as a whole can be modeled by a mixed Markov model (Langeheine and Van de Pol, 1994; Frühwirth-Schnatter et al., 2018).

Another approach is the use of Markov chain models of higher order. In their model of U.S. GNP growth, Durland and McCurdy (1994) consider a two-state Markov model where the state transition probabilities are described by logistic functions of state duration.

Other papers deal with DS-transitions through the selection of states by duration values or duration intervals. In a Markov model for major depressive disorder, Patten (2005) uses an ordered set of temporary states in which a patient can spend only one time-step before going to the recovery state or to the next temporary state. These states are called *tunnel states* (Sonnenberg and Beck, 1993) and they account for different probabilities of transition to the recovery state based on depression episode durations. DS-transitions are also incorporated in the so-called Cornell mobility model postulating the *axiom of cumulative inertia*. This axiom expresses that the longer an individual has stayed in a particular status, the more likely he will remain there. In the context of social mobility, see McGinnis (1968), this means that an individual who has already obtained a particular social status for some time has a higher probability to remain in that status for the next time period compared to a relative newcomer. For this reason, McFarland (1970) suggests to define the states of the Markov model based on a social status combined with the duration of stay in that status.

In this paper, we further investigate the technique proposed by Bartholomew et al. (1991) to partition the system's categories into subgroups by duration of stay intervals and then to use these subgroups as the states of a Markov model. First, we aim to address some of the problems with the proposed duration-interval approach. Such a duration-interval model does not fulfill the required homogeneity of states regarding the transition probabilities because, within a category, a system unit can only make a transition (in one step) to the next (age- or seniority) interval if it has reached the endpoint of the previous interval. Second, we prove that the likelihood of a set of panel data given this duration-interval model cannot exceed the likelihood given the baseline Markov model with the categories as the states if certain conditions on the data are met. We propose to remedy these issues by disaggregating the duration of stay intervals into discrete duration of stay point values, thus considering subgroups according to the exact number of periods of time in the category.

The present paper is organised as follows. Section 2 looks at alternative time-homogeneous and discrete Markov models by selecting the states in two different ways: states equal to the categories partitioned in duration intervals

and states equal to the categories divided by duration values. In Section 3, we study the relative quality of the resulting Markov models using the Akaike Information Criterion (AIC; Akaike, 1973) and the Bayesian Information Criterion (BIC; Schwarz, 1978) for a given set of panel data. In Section 4 the findings are illustrated using a real-world data set. We conclude with a discussion in Section 5. Supporting lemmas and their proofs are gathered in Appendix A.

2 Markov models for duration-dependent transitions

In the following, we consider the case of an hierarchical system, i.e. a system that can be divided into s categories $C_1, \ldots, C_s$, in such a way that one-step transitions from one category to another only occur from C_i to C_{i+1} ($s \geq 2$, $1 \leq i \leq s-1$). An example of such a system is a hierarchical organization or firm in which the employees can only promote from one grade to the next.

Choose $i_0 \in \{1, \ldots, s-1\}$ so that C_{i_0} is a category for which the transition probability to the next category C_{i_0+1} depends on the duration of stay in C_{i_0}. For the sake of simplicity, denote $A = C_{i_0}$ and $B = C_{i_0+1}$. Suppose, the duration effect is such that all system units with duration of stay $\leq k$ in A have the same transition probability, say p_1, to B and that those with duration of stay $> k$ in A face the transition probability p_2 ($\neq p_1$) to B. Dichotomous duration effects of this kind appear in career mobility studies that try to reveal for employees with comparable career attainment or career success whether they are characterized by a fast-track or a slow-track career progression (Forbes, 1987; Lyness and Thompson, 2000).

Suppose panel data on the system units' categories are available. Generally, these data are censored on both sides of the observation period. However, since the focus of this paper is on comparing alternative Markov models, we consider, for estimation purposes, the subset of the data with complete information on the observed transitions from A to B. We call such subset *AB-complete*. Formally, an AB-complete data set satisfies the following properties:

- A system unit that moves from A to B during the observation period has also entered A during that period.
- A system unit that visits A during the observation period also makes a transition to B during that period.

In any Markov chain model, the states should satisfy the homogeneity requirement concerning the transition probabilities. In the following sections, we discuss two ways of selecting the states. We compare the maximized likelihoods of the data in view of the associated Markov chain models. The presented approaches are fully described for the transitions from A to B but are also transferable to any pair of consecutive categories C_{i_0} and C_{i_0+1} of the system.

2.1 State selection by duration intervals

As a modeling approach to duration effects, Bartholomew et al. (1991) have suggested to consider the Markov model with the states defined by length-of-stay intervals in a category. We now formalize his idea.

Denote $\mathcal{C} = \{C_1, \ldots, C_s\}$. Let $A = C_{i_0}$ be a category for which a dichotomous duration effect exists in the transition to the next category $B = C_{i_0+1}$ $(1 \leq i_0 \leq s-1)$. Let k be a natural number. Denote by $A_{[0,k]}$ the state of being in category A for at most k periods of time, and by $A_{(k,\infty)}$ the state of being in A for more than k periods of time. We employ the notation "M(k)" for a Markov chain with state space $\mathcal{C}_{i_0} \cup \mathcal{S}$, where

$$\mathcal{C}_{i_0} = \mathcal{C} \setminus \{C_{i_0}, C_{i_0+1}\} = \mathcal{C} \setminus \{A, B\} \tag{1}$$

and

$$S_1 = A_{[0,k]}\ ,\ S_2 = A_{(k,\infty)}\ ,\ S_3 = B\ ,\ \mathcal{S} = \{S_1, S_2, S_3\}, \tag{2}$$

and where the transition probabilities from state S_i to state S_j, denoted π_{ij}, are subject to the constraints

$$\pi_{21} = \pi_{31} = \pi_{32} = 0 \tag{3}$$

due to the hierarchical nature of the considered system.

For $C_i, C_j \in \mathcal{C}_{i_0}$, let p_{ij} denote the transition probability from C_i to C_j. Denote $n(X,Y)$ the number of one-step transitions from state X to state Y and let $n(X) = \sum_Y n(X,Y)$. The maximized likelihood of the panel data given M(k) and given the initial distribution among the states is then

$$\hat{L}_{\mathrm{M}(k)} = \hat{L}_{\mathcal{C}_{i_0}} \cdot \hat{L}_{\mathcal{S}}$$

where

$$\hat{L}_{\mathcal{C}_{i_0}} = \prod_{\substack{(C_i,C_j)\in\mathcal{C}^2 \\ i\neq i_0}} \hat{p}_{ij}^{\,n(C_i,C_j)} \quad \text{and} \quad \hat{L}_{\mathcal{S}} = \prod_{(S_i,S_j)\in\mathcal{S}^2} \hat{\pi}_{ij}^{\,n(S_i,S_j)} \tag{4}$$

and any factors 0^0 are set equal to one. The maximum likelihood estimators of the various transition probabilities are given by

$$\hat{p}_{ij} = \frac{n(C_i,C_j)}{n(C_i)}\ ,\ \hat{\pi}_{ij} = \frac{n(S_i,S_j)}{n(S_i)}. \tag{5}$$

This state selection method seems a natural approach to the assumed duration effects. However, there are some issues with it. First, we shall show in Section 3 that, under certain conditions, the maximized likelihood $\hat{L}_{\mathrm{M}(k)}$ cannot exceed the maximized likelihood of the same data given the baseline Markov chain model with state space $\mathcal{C}$, which we denote "M". This entails that the AIC- and BIC-values of model M(k) cannot be smaller than those of model M (see for more details Section 3). A second issue is that state S_1

in model M(k) is not homogeneous regarding the transition probabilities. Indeed, system units in S_1 can only move to S_2 if they have stayed exactly k periods of time in category A. As such, the transition from S_1 to S_2 still remains duration-dependent. In order to avoid these issues without sacrificing the simplicity of a Markov chain model, we consider an alternative definition of the states as explained in the following section.

2.2 State selection by duration values

Let $A_{[0,k]} = \cup_{t=0}^{k} A_t$, where A_t is the event of being in category A for *exactly* t periods of time. Define $\widetilde{\mathrm{M}}(k)$ to be a Markov chain with state space $\mathcal{C}_{i_0} \cup \tilde{\mathcal{S}}$, where $\tilde{\mathcal{S}} = \{\tilde{S}_1, \ldots, \tilde{S}_{k+3}\}$ and

$$\tilde{S}_1 = A_0 \ , \ \tilde{S}_2 = A_1 \ , \ \ldots \ , \ \tilde{S}_{k+1} = A_k \ , \ \tilde{S}_{k+2} = A_{(k,\infty)} \ , \ \tilde{S}_{k+3} = B, \tag{6}$$

and such that the transition probabilities from states $\tilde{S}_i$ to $\tilde{S}_j$, denoted τ_{ij}, are subject to the constraints

$$\begin{cases} \tau_{ij} = 0, \text{ if } j \leq i \text{ or } i+2 \leq j \leq k+2, \\ \tau_{1,k+3} = \cdots = \tau_{k+1,k+3} \end{cases} \tag{7}$$

by virtue of the hierarchical nature of the considered system.

The maximized likelihood of the data in view of $\widetilde{\mathrm{M}}(k)$ conditional on the initial distribution among the states is

$$\hat{L}_{\widetilde{\mathrm{M}}(k)} = \hat{L}_{\mathcal{C}_{i_0}} \cdot \hat{L}_{\tilde{\mathcal{S}}},$$

where $\hat{L}_{\mathcal{C}_{i_0}}$ is given by (4) and $\hat{L}_{\tilde{\mathcal{S}}} = \prod_{(\tilde{S}_i,\tilde{S}_j)\in\tilde{\mathcal{S}}^2} \hat{\tau}_{ij}^{\,n(\tilde{S}_i,\tilde{S}_j)}$ (and possible factors 0^0 are taken equal to one). For the sake of simplicity, let us denote $\tau_{1,k+3} = \cdots = \tau_{k+1,k+3} = \tau_1$ and $\tau_{k+2,k+3} = \tau_2$. Using the constraints (7) on the transition probabilities, we can rewrite the expression for $\hat{L}_{\tilde{\mathcal{S}}}$ as follows:

$$\hat{L}_{\tilde{\mathcal{S}}} = (1-\hat{\tau}_1)^{\sum_{i=1}^{k+1} n(\tilde{S}_i,\tilde{S}_{i+1})} \hat{\tau}_1^{\sum_{i=1}^{k+1} n(\tilde{S}_i,\tilde{S}_{k+3})} (1-\hat{\tau}_2)^{n(\tilde{S}_{k+2},\tilde{S}_{k+2})} \hat{\tau}_2^{n(\tilde{S}_{k+2},\tilde{S}_{k+3})}, \tag{8}$$

where

$$\hat{\tau}_1 = \frac{\sum_{i=1}^{k+1} n(\tilde{S}_i,\tilde{S}_{k+3})}{\sum_{i=1}^{k+1} n(\tilde{S}_i)} \quad , \quad \hat{\tau}_2 = \frac{n(\tilde{S}_{k+2},\tilde{S}_{k+3})}{n(\tilde{S}_{k+2})}.$$

The state spaces of models M(k) and $\widetilde{\mathrm{M}}(k)$ are connected as follows

$$S_1 = \cup_{i=1}^{k+1} \tilde{S}_i \ , \ S_2 = \tilde{S}_{k+2} \ , \ S_3 = \tilde{S}_{k+3}, \tag{9}$$

hence $\sum_{i=1}^{k+1} n(\tilde{S}_i,\tilde{S}_{i+1}) = n(S_1,S_1)+n(S_1,S_2)$, $\sum_{i=1}^{k+1} n(\tilde{S}_i,\tilde{S}_{k+3}) = n(S_1,S_3)$, $\sum_{i=1}^{k+1} n(\tilde{S}_i) = n(S_1)$. Denoting $n_{ij} = n(S_i,S_j)$ and $n_i = n(S_i)$ for simplicity, (8) can be rewritten as

$$\hat{L}_{\tilde{\mathcal{S}}} = \left(1-\tfrac{n_{13}}{n_1}\right)^{n_{11}+n_{12}} \left(\tfrac{n_{13}}{n_1}\right)^{n_{13}} \left(1-\tfrac{n_{23}}{n_2}\right)^{n_{22}} \left(\tfrac{n_{23}}{n_2}\right)^{n_{23}} \tag{10}$$

In the next section, we show that $\hat{L}_{\widetilde{\mathrm{M}}(k)}$ does always exceed the maximized likelihood of the same data given the baseline Markov chain model with state space $\{C_1, \ldots, C_s\}$, except for the case when $\hat{\tau}_1 = \hat{\tau}_2$.

3 Likelihood fit comparison

In a model selection process using an information criterion, the maximized likelihoods of the data in view of the candidate models are considered. We now compare – conditional on the initial distribution among the states – the likelihoods for the duration-interval model $\mathrm{M}(k)$ and the duration-value model $\widetilde{\mathrm{M}}(k)$ to the baseline Markov chain model M, having the set $\mathcal{C} = \{C_1, \ldots, C_s\}$ of categories as state space. To this end, we adopt the following notation:

$$\Delta\ell\ell_{\mathcal{M}_1\text{-}\mathcal{M}_2} = \ln \hat{L}_{\mathcal{M}_1} - \ln \hat{L}_{\mathcal{M}_2}, \tag{11}$$

where $\mathcal{M}_i \in \{\widetilde{\mathrm{M}}(k), \mathrm{M}(k), \mathrm{M}\}$ for $i = 1, 2$ and $k \geq 1$.

For model M, we have

$$\hat{L}_{\mathrm{M}} = \prod_{(C_i, C_j) \in \mathcal{C}^2} \hat{p}_{ij}^{\,n(C_i, C_j)} = \hat{L}_{\mathcal{C}_{i_0}} \cdot \frac{{n_{AA}}^{n_{AA}} {n_{AB}}^{n_{AB}}}{{n_A}^{n_A}},$$

where $\hat{L}_{\mathcal{C}_{i_0}}$ is given by (4) and $n_{AA} = n(A, A)$, $n_{AB} = n(A, B)$ and $n_A = n_{AA} + n_{AB}$. With the use of the auxiliary function ϕ defined as

$$\phi(t_1, t_2) = \varphi(t_1) + \varphi(t_2) - \varphi(t_1 + t_2), \tag{12}$$

where

$$\varphi(t) = \begin{cases} t \ln t & \text{if } t > 0 \\ 0 & \text{if } t = 0, \end{cases} \tag{13}$$

the log-likelihood of model M can be rewritten as

$$\ln \hat{L}_{\mathrm{M}} = \ln \hat{L}_{\mathcal{C}_{i_0}} + \phi(n_{AA}, n_{AB}). \tag{14}$$

The log-likelihoods of models $\mathrm{M}(k)$ and $\widetilde{\mathrm{M}}(k)$ can also be expressed using the function ϕ. We hereby need some supporting lemmas which are proven in Appendix Appendix A. By lemma 1

$$\begin{aligned} \ln \hat{L}_{\mathrm{M}(k)} &= \ln \hat{L}_{\mathcal{C}_{i_0}} + \ln \hat{L}_{\mathcal{S}} \\ &= \ln \hat{L}_{\mathcal{C}_{i_0}} + \phi(n_{11}, n_{12}) + \phi(n_{11} + n_{12}, n_{13}) + \phi(n_{22}, n_{23}) \end{aligned} \tag{15}$$

and by lemma 2

$$\begin{aligned} \ln \hat{L}_{\widetilde{\mathrm{M}}(k)} &= \ln \hat{L}_{\mathcal{C}_{i_0}} + \ln \hat{L}_{\tilde{\mathcal{S}}} \\ &= \ln \hat{L}_{\mathcal{C}_{i_0}} + \phi(n_{11} + n_{12}, n_{13}) + \phi(n_{22}, n_{23}). \end{aligned} \tag{16}$$

Using (14), (15) and (16), we can now express the differences in maximized log-likelihoods in terms of the one-step transition frequencies $n_{ij} = n(S_i, S_j)$ $(1 \leq i \leq j \leq 3)$, n_{AA} and n_{AB}:

$$\Delta\ell\ell_{\mathrm{M}(k)\text{-}\mathrm{M}} = \phi(n_{11}, n_{12}) + \phi(n_{11} + n_{12}, n_{13}) + \phi(n_{22}, n_{23}) - \phi(n_{AA}, n_{AB}) \tag{17}$$

$$\Delta\ell\ell_{\widetilde{\mathrm{M}}(k)\text{-}\mathrm{M}} = \phi(n_{11} + n_{12}, n_{13}) + \phi(n_{22}, n_{23}) - \phi(n_{AA}, n_{AB}) \tag{18}$$

$$\Delta\ell\ell_{\mathrm{M}(k)\text{-}\widetilde{\mathrm{M}}(k)} = \phi(n_{11}, n_{12}) \tag{19}$$

We are now ready to formulate our main results on the signs of the log-likelihood differences.

Theorem 1 *For an AB-complete data set satisfying $n_{AA} > 2n_{AB}$ and for $k \geq 1$, it holds that $\Delta\ell\ell_{\mathrm{M}(k)\text{-}\mathrm{M}} < 0$ if*

$$\hat{\tau}_{23} > \hat{p}_{AB} \quad \textit{and} \quad \left(n_2 \geq \tfrac{3}{2} n_{AB} \;\; \textit{or} \;\; n_{12} \leq \tfrac{3}{4} n_{AB}\right). \tag{20}$$

Proof See Appendix A.

Theorem 2 *For an AB-complete data set, it holds that $\Delta\ell\ell_{\widetilde{\mathrm{M}}(k)\text{-}\mathrm{M}} \geq 0$ with equality if and only if $\hat{\tau}_{13} = \hat{\tau}_{23} = \hat{p}_{AB}$.*

Proof See Appendix A.

Theorem 3 *For an AB-complete data set, it holds that $\Delta\ell\ell_{\mathrm{M}(k)\text{-}\widetilde{\mathrm{M}}(k)} \leq 0$ with equality if and only if $n_{11} = 0$ or $n_{12} = 0$.*

Proof See Appendix A.

The above theorems allow to draw some conclusions about model selection, when the models $\mathrm{M}(k)$, $\widetilde{\mathrm{M}}(k)$ and M are compared using common information-theoretic criteria such as the Akaike Information Criterion (AIC; Akaike, 1973) and the Bayesian Information Criterion (BIC; Schwarz, 1978). Both criteria assess model fit penalized for the number of estimated parameters. AIC and BIC are defined as:

$$\mathrm{AIC} = -2 \ln \hat{L} + 2K, \quad \mathrm{BIC} = -2 \ln \hat{L} + K \cdot \ln n,$$

where $\hat{L}$ is the maximized likelihood of the data given the model, K is the number of estimable (free) parameters of the model and n is the number of observations in the dataset at hand. In an information-theoretic approach, models with smaller AIC/BIC-values are 'preferred' over models with larger AIC/BIC-values. Among the candidate models, the model with the smallest AIC-value need not be the model with the smallest BIC-value and vice versa. For more details on the interpretation of information-theoretic criteria, we refer to Anderson and Burnham (2004). In the context of Markov chains, AIC and BIC have been used and studied to estimate the order of the chain (Katz, 1981). Cartella et al. (2015) determine for a hidden semi-Markov chain

the optimal number of hidden states by AIC. A discussion on selecting the number of latent classes in a latent Markov model using AIC and BIC is given by Bacci et al. (2014).

The AIC- and BIC-differences between two candidate models $\mathcal{M}_1$ and $\mathcal{M}_2$ are expressible in terms of the corresponding log-likelihood difference and the difference in number of estimable parameters:

$$\Delta\mathrm{AIC}_{\mathcal{M}_1\text{-}\mathcal{M}_2} = -2\Delta\ell\ell_{\mathcal{M}_1\text{-}\mathcal{M}_2} + 2\Delta K_{\mathcal{M}_1\text{-}\mathcal{M}_2}, \tag{21}$$

$$\Delta\mathrm{BIC}_{\mathcal{M}_1\text{-}\mathcal{M}_2} = -2\Delta\ell\ell_{\mathcal{M}_1\text{-}\mathcal{M}_2} + \Delta K_{\mathcal{M}_1\text{-}\mathcal{M}_2} \cdot \ln n. \tag{22}$$

Using the terminology of Bartolucci et al. (2012, Chapter 4), the Markov chain models $\mathrm{M}(k)$ and $\widetilde{\mathrm{M}}(k)$ are 'constrained' in the sense that the transition matrix is parameterized in a way specific to the system studied. By the constraints (3) and (7), we have $\Delta K_{\mathrm{M}(k)\text{-}\mathrm{M}} = 2$, $\Delta K_{\widetilde{\mathrm{M}}(k)\text{-}\mathrm{M}} = 1$ and $\Delta K_{\mathrm{M}(k)\text{-}\widetilde{\mathrm{M}}(k)} = 1$. Under the conditions of Theorem 1, we have $\Delta\ell\ell_{\mathrm{M}(k)\text{-}\mathrm{M}} < 0$, hence, by (21) and (22), both $\Delta\mathrm{AIC}_{\mathrm{M}(k)\text{-}\mathrm{M}}$ and $\Delta\mathrm{BIC}_{\mathrm{M}(k)\text{-}\mathrm{M}}$ will be positive. In this case, both AIC and BIC lead to the non-selection of model $\mathrm{M}(k)$ over the baseline Markov model M. More generally, this conclusion holds when $\Delta\ell\ell_{\mathrm{M}(k)\text{-}\mathrm{M}} < \min\{2, \ln n\}$.

Furthermore, the signs of $\Delta\mathrm{AIC}_{\widetilde{\mathrm{M}}(k)\text{-}\mathrm{M}}$ and $\Delta\mathrm{BIC}_{\widetilde{\mathrm{M}}(k)\text{-}\mathrm{M}}$ depend on the value of the log-likelihood difference $\Delta\ell\ell_{\widetilde{\mathrm{M}}(k)\text{-}\mathrm{M}}$. If $\hat{\tau}_{13} = \hat{\tau}_{23} = \hat{p}_{AB}$ (no duration effect), we have $\Delta\ell\ell_{\widetilde{\mathrm{M}}(k)\text{-}\mathrm{M}} = 0$ by Theorem 2 and thus $\Delta\mathrm{AIC}_{\widetilde{\mathrm{M}}(k)\text{-}\mathrm{M}}$ and $\Delta\mathrm{BIC}_{\widetilde{\mathrm{M}}(k)\text{-}\mathrm{M}}$ will be positive, so that AIC and BIC will not select $\widetilde{\mathrm{M}}(k)$ over the baseline Markov model M in that case. The same conclusion holds when $0 < \Delta\ell\ell_{\widetilde{\mathrm{M}}(k)\text{-}\mathrm{M}} \leq 1$ for AIC; if BIC is used, it holds only when $n > \mathrm{e}^2$. In practice, the number of observations n in the panel data set is sufficiently high, hence we assume $n > \mathrm{e}^2$ from this point on. If $1 < \Delta\ell\ell_{\widetilde{\mathrm{M}}(k)\text{-}\mathrm{M}} \leq \ln\sqrt{n}$, $\widetilde{\mathrm{M}}(k)$ is chosen over M by AIC but not by BIC. If $\Delta\ell\ell_{\widetilde{\mathrm{M}}(k)\text{-}\mathrm{M}} > \ln\sqrt{n} > 1$, both AIC and BIC select model $\widetilde{\mathrm{M}}(k)$ over M. Finally, model $\widetilde{\mathrm{M}}(k)$ is selected over model $\mathrm{M}(k)$ by both AIC and BIC, because of Theorem 3. These results on pairwise comparison of the models $\mathrm{M}(k)$, $\widetilde{\mathrm{M}}(k)$ and M are summarized in Table 1.

Table 1: Model selection via AIC and BIC in case $n > \mathrm{e}^2$ for various log-likelihood difference values.

$\mathcal{M}_1$	$\mathcal{M}_2$	$\Delta K_{\mathcal{M}_1\text{-}\mathcal{M}_2}$	$\Delta\ell\ell_{\mathcal{M}_1\text{-}\mathcal{M}_2}$	AIC	BIC
$\mathrm{M}(k)$	M	2	$\cdot < 2$	M	M
$\widetilde{\mathrm{M}}(k)$	M	1	$0 \leq \cdot < 1$	M	M
$\widetilde{\mathrm{M}}(k)$	M	1	$1 < \cdot < \ln\sqrt{n}$	$\widetilde{\mathrm{M}}(k)$	M
$\widetilde{\mathrm{M}}(k)$	M	1	$\cdot > \ln\sqrt{n}$	$\widetilde{\mathrm{M}}(k)$	$\widetilde{\mathrm{M}}(k)$
$\mathrm{M}(k)$	$\widetilde{\mathrm{M}}(k)$	1	$-\infty < \cdot < \infty$	$\widetilde{\mathrm{M}}(k)$	$\widetilde{\mathrm{M}}(k)$

Noteworthy is the agreement of AIC and BIC on the selection of model $\widetilde{\mathrm{M}}(k)$ over M whenever the number of observations n does not exceed the value $\mathrm{e}^{2\Delta\ell\ell_{\widetilde{\mathrm{M}}(k)\text{-}\mathrm{M}}}$.

4 Illustrative example

We use a real-world panel data set of academic staff of a Belgian University from 1999 to 2013. To study the career advancement from grade A to grade B, we consider all the faculty members who entered grade A and eventually promoted to grade B during the aforementioned time horizon. For A = "lecturer" and B = "senior lecturer", we thus observe $n_{AB} = 68$ faculty members for which $n_{AA} = 172$. Their career progression from lecturer to senior lecturer takes $\frac{n_{AA}}{n_{AB}} + 1 = 3.52$ years on average with a maximum of $T = 10$ years. To illustrate Theorems 1 and 2, we compare the maximized likelihood of the data given models $\mathrm{M}(k)$ and $\widetilde{\mathrm{M}}(k)$ to the baseline Markov chain model M for values of k ranging from 1 to $T-2$. The results are shown in Table 2. Therein, $\Delta\ell\ell_{\mathrm{M}(k)\text{-M}} = \psi(n_{11}, n_{12})$ where ψ is defined by (24), and $\Delta\ell\ell_{\widetilde{\mathrm{M}}(k)\text{-M}} = \Delta\ell\ell_{\mathrm{M}(k)\text{-M}} - \phi(n_{11}, n_{12})$ using (17) and (18).

Table 2: Model comparison results for the career progression from lecturer (A) to senior lecturer (B). Figures in bold correspond to the value of k where $\Delta\ell\ell_{\widetilde{\mathrm{M}}(k)\text{-M}}$ is maximal.

k	n_{11}	n_{12}	n_2	$\hat{\tau}_{23}$	$\Delta\ell\ell_{\mathrm{M}(k)\text{-M}}$	$\Delta\ell\ell_{\widetilde{\mathrm{M}}(k)\text{-M}}$
1	55	43	117	0.37	−63.18	4.02
2	98	34	74	0.46	−67.46	**7.84**
3	132	18	40	0.45	−51.97	3.07
4	150	11	22	0.50	−37.59	2.55
5	161	5	11	0.45	−21.67	0.76
6	166	4	6	0.67	−17.00	1.95
7	170	1	2	0.50	−5.93	0.21
8	171	1	1	1.00	−4.88	1.27

$n_{AB} = 68$, $n_{AA} = 172$, $\hat{p}_{AB} = 0.28$

For all tabulated values of k, $\Delta\ell\ell_{\mathrm{M}(k)\text{-M}}$ is negative. This agrees with Theorem 1, since, for all k, $\hat{\tau}_{23} > \hat{p}_{AB}$ and $n_{12} < \frac{3}{4}n_{AB} = 51$. Remark that all entries in the last column of Table 2 are positive, as they should be by Theorem 2.

The model comparison results for the career progression from senior lecturer (A) to professor (B) are displayed in Table 3. Again, $\Delta\ell\ell_{\mathrm{M}(k)\text{-M}} < 0$ for all values of k. This is also in agreement with Theorem 1 because $\hat{\tau}_{23} > \hat{p}_{AB}$ and $n_2 > \frac{3}{2}n_{AB} = 99$ (if $k \leq 2$) or $n_{12} < \frac{3}{4}n_{AB} = 49.5$ (if $k \geq 3$).

In Tables 2 and 3, $\Delta\ell\ell_{\widetilde{\mathrm{M}}(k)\text{-M}}$ attains its maximum value at $k = 2$. Since $\Delta\ell\ell_{\widetilde{\mathrm{M}}(2)\text{-M}} > 1$ in both tables, model $\widetilde{\mathrm{M}}(2)$ is selected over the baseline model M by AIC (see Table 1). According to Table 1, BIC selects model $\widetilde{\mathrm{M}}(2)$ over model M in both career progression examples, as the number n of observations in the panel dataset, calculated as the academic staff size multiplied by the number of years (14), is well below $\mathrm{e}^{2(7.84)} \approx 6452640$ for a single university institution in Belgium.

Table 3: Model comparison results for the career progression from senior lecturer (A) to professor (B). Figures in bold correspond to the value of k where $\Delta\ell\ell_{\widetilde{\mathrm{M}(k)\text{-M}}}$ is maximal.

k	n_{11}	n_{12}	n_2	$\hat{\tau}_{23}$	$\Delta\ell\ell_{\mathrm{M}(k)\text{-M}}$	$\Delta\ell\ell_{\widetilde{\mathrm{M}(k)\text{-M}}}$
1	62	55	178	0.31	−68.91	11.98
2	117	50	123	0.41	−79.75	**22.18**
3	167	28	73	0.38	−72.91	7.32
4	195	17	45	0.38	−55.53	3.67
5	212	11	28	0.39	−41.32	2.51
6	223	8	17	0.47	−31.86	2.90
7	231	6	9	0.67	−23.61	4.37
8	237	2	3	0.67	−10.14	1.42
9	239	1	1	1.00	−4.94	1.54

$n_{AB} = 66$, $n_{AA} = 240$, $\hat{p}_{AB} = 0.22$

Considering the above observations, the promotion probability of individuals who remain at most two years in their grade appears to contrast with the promotion probability of those having a longer length of service in that grade. In Figures 1 and 2, the promotion rates for different grade seniority values are shown together with error bars based on one standard error of proportion estimation. The horizontal dotted line represents $\hat{p}_{AB}$. It is clear from these figures that $k = 2$ is the only value that splits the graph in two homogeneous parts with similar promotion rates.

5 Discussion

Bartholomew et al. (1991) have suggested that transitions depending on the length of stay in a state could be incorporated into the Markov chain framework by a suitable definition of the states using length-of-stay intervals. Formalising this idea, the model M(k) tries to capture the situation in which transition probabilities over the length-of-stay intervals $[0, k]$ and (k, ∞) are different. Although the state space definition of model M(k) seems a natural choice, there are some issues related to it. First, the state-homogeneity regarding the transition probabilities, which is required in a Markov model, cannot be fulfilled when the states are based on duration intervals. In addition, we prove Theorem 1 which entails that, for non-censored panel data satisfying certain conditions, model M(k) cannot be preferred over the baseline Markov chain model M which ignores the duration effects, when AIC and BIC are used as model selection tool criteria. In the light of the presumed duration effect, the result of Theorem 1 surprises. So far, it is unclear whether the result still holds for a censored data set.

In a manpower planning context, the conditions of Theorem 1 can be interpreted as follows. The inequality $n_{AA} > 2n_{AB}$ reflects that the n_{AB} promotees in the data-set stay on average more than three periods of time in grade A before promoting to B. The employee who promotes to B at the moment he attains a grade seniority of at most k in A, can be called a fast-tracker.

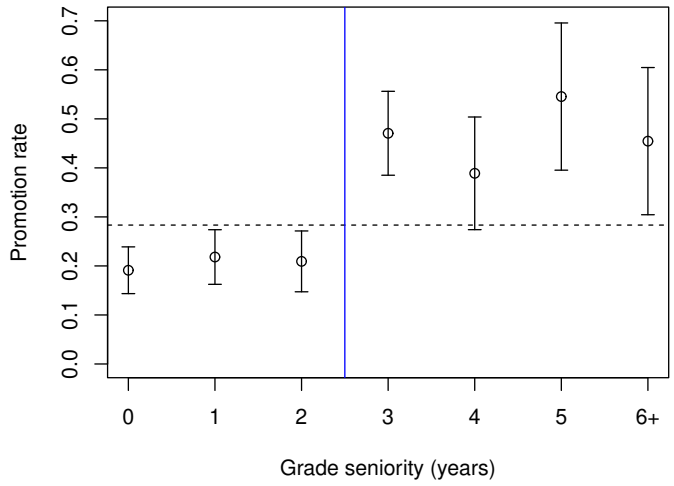


Fig. 1: Promotion rates from lecturer to senior lecturer by grade seniority, with error bars based on one standard error of proportion estimation.

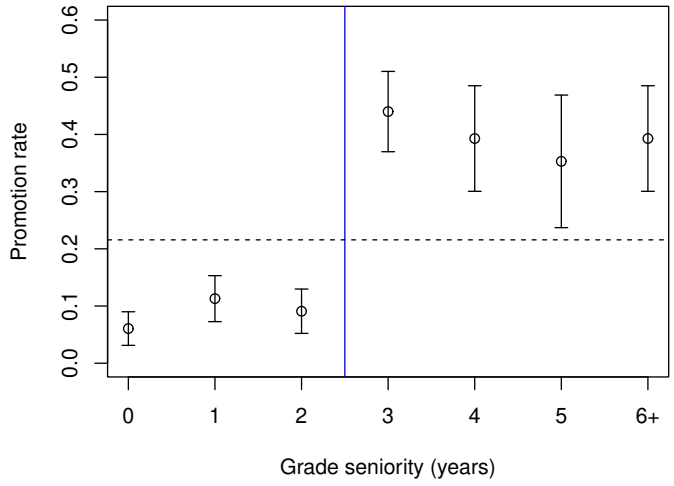


Fig. 2: Promotion rates from senior lecturer to professor by grade seniority, with error bars based on one standard error of proportion estimation.

Employees who remain longer in A before promoting to B are then named slow-trackers. The inequality $\hat{\tau}_{23} > \hat{p}_{AB}$ reflects that fast-track promotions are not as common as slow-track promotions in the data-set. The inequality $n_2 \geq \frac{3}{2}\, n_{AB}$ signifies the employees spend on average at least 3/2 periods of time in a slow-track career progression state in A before switching to B. Finally, the inequality $n_{12} \leq \frac{3}{4}\, n_{AB}$ expresses that at least 25% of the promotees in the data-set are fast-trackers. These conditions seem realistic in practice, as our illustrative example in Section 4 demonstrates.

To avoid the issues of the duration-interval model $\mathrm{M}(k)$ mentioned above, we have adapted its state space by disaggregating the state $A_{[0,k]}$ into its duration-value component states $A_0, \ldots, A_k$ having equal one-step transition probabilities to category B. The Markov chain so obtained is denoted $\widetilde{\mathrm{M}}(k)$. The states $A_0, \ldots, A_k$ have the property that the only one-step transition from A_i $(0 \leq i \leq k)$ is either to B, or to A_{i+1} (if $i < k$) or $A_{(k,+\infty)}$ (if $i = k$). These duration-value component states are examples of what Sonnenberg and Beck (1993) call *tunnel states*, i.e. states that can only be visited in a fixed order and without transitions to themselves.

Let k^* be a duration-of-stay cut-off value for which the likelihood difference $\Delta\ell\ell_{\widetilde{\mathrm{M}}(k^*)\text{-M}}$ between models $\widetilde{\mathrm{M}}(k^*)$ and M is maximized. Provided that complete information on the history from category A to category B is available and used, we have by Theorem 2 that $\Delta\ell\ell_{\widetilde{\mathrm{M}}(k^*)\text{-M}}$ is non-negative. If in addition $\Delta\ell\ell_{\widetilde{\mathrm{M}}(k^*)\text{-M}} > 1$, the AIC will select model $\widetilde{\mathrm{M}}(k^*)$ over M, see Table 1. In both our examples, we have found $k^* = 2$. Figures 1 and 2 suggest that this value of k^* corresponds with the clearest cut between the fast-track and slow-track promotion rates.

To introduce and study properties of the duration-interval model and duration-value model, we have focused on the transitions between two arbitrarily chosen categories A and B, for which a duration effect is present. Nevertheless our finding can be generalized to the case of multiple categories with duration effects. By a recursive argument, the state selection by duration intervals for all involved categories can never result in a Markov model with smaller AIC-value.

We conclude with the following thoughts on the duration-interval model. The hidden Markov model has a two-layered structure consisting of observable variables and hidden states. Similarly, the duration-interval model has also two layers, being the categories and the corresponding states based on duration intervals. In the hidden Markov model, AIC and BIC enable to determine the optimal number of hidden states (Bacci et al., 2014; Cartella et al., 2015). In this way, the AIC and BIC criterion are useful in determining the best level of disaggregation of the categories into duration intervals. Suppose this approach results in a situation where the duration effect is described by κ cut-points, say $k_1 < \cdots < k_\kappa$ where $\kappa \geq 2$, instead of one cut-point k as described in the present paper. Then it would be of interest to see if the unexpected result of Theorem 1 continues to hold in this case, i.e. does the corresponding duration-interval model $\mathrm{M}(k_1, \ldots, k_\kappa)$ still have a poorer likelihood fit to a set

of panel data than the baseline Markov model? To our knowledge, this is an open question.

References

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: 2nd Intr. Symp. on Information Theory, Budapest, 1973, Akademiai Kiado

Anderson D, Burnham K (2004) Model selection and multi-model inference. Second NY: Springer-Verlag 63(2020):10

Bacci S, Pandolfi S, Pennoni F (2014) A comparison of some criteria for states selection in the latent Markov model for longitudinal data. Advances in Data Analysis and Classification 8(2):125–145

Bányai T, Landschützer C, Bányai Á (2018) Markov-chain simulation-based analysis of human resource structure: How staff deployment and staffing affect sustainable human resource strategy. Sustainability 10(10):3692

Barbu V, Limnios N (2008) Semi-Markov chains and hidden semi-Markov models toward applications: their use in reliability and DNA analysis. Springer Science & Business Media, New York

Bartholomew D, Forbes A, McClean S (1991) Statistical Techniques for Manpower Planning. Wiley, Chichester

Bartolucci F, Farcomeni A, Pennoni F (2012) Latent Markov Models for Longitudinal Data. Chapman and Hall/CRC, Boca Raton, London, New York

Cartella F, Lemeire J, Dimiccoli L, Sahli H (2015) Hidden semi-Markov models for predictive maintenance. Mathematical Problems in Engineering 2015

D'Amico G, Janssen J, Manca R (2006) Homogeneous semi-Markov reliability models for credit risk management. Decisions in Economics and Finance 28(2):79–93

De Feyter T (2006) Modelling heterogeneity in manpower planning: dividing the personnel system into more homogeneous subgroups. Applied Stochastic Models in Business and Industry 22(4):321–334

Dewar M, Wiggins C, Wood F (2012) Inference in hidden Markov models with explicit state duration distributions. IEEE Signal Processing Letters 19(4):235–238

Duffy S, Day N, Tabár L, Chen H, Smith T (1997) Markov models of breast tumor progression: some age-specific results. Journal of the National Cancer Institute Monographs 1997(22):93–97

Durland JM, McCurdy TH (1994) Duration-dependent transitions in a Markov model of us gnp growth. Journal of Business & Economic Statistics 12(3):279–288

Forbes J (1987) Early intraorganizational mobility: Patterns and influences. Academy of Management Journal 30(1):110–125

Frühwirth-Schnatter S, Pittner S, Weber A, Winter-Ebmer R (2018) Analysing plant closure effects using time-varying mixture-of-experts Markov chain clustering. Annals of Applied Statistics 12(3):1796–1830

Jiang Y, Sinha K (1989) Bridge service life prediction model using the Markov chain. Transportation research record 1223:24–30

Katz RW (1981) On some criteria for estimating the order of a Markov chain. Technometrics 23(3):243–249

Langeheine R, Van de Pol F (1994) Discrete-time mixed Markov latent class models. Analyzing social and political change: A casebook of methods pp 170–197

Longini I, Clark W, Gardner L, Brundage J (1991) The dynamics of CD4+ T-lymphocyte decline in HIV-infected individuals: a Markov modeling approach. Journal of Acquired Immune Deficiency Syndromes 4(11):1141–1147

Lyness K, Thompson D (2000) Climbing the corporate ladder: do female and male executives follow the same route? Journal of applied psychology 85(1):86

McFarland D (1970) Intragenerational social mobility as a Markov process: Including a time-stationary Markovian model that explains observed declines in mobility rates over time. American Sociological Review 35(3):463–476

McGinnis R (1968) A stochastic model of social mobility. American Sociological Review 33(5):712–722

Parker B, Caine D (1996) Holonic modelling: human resource planning and the two faces of janus. International Journal of Manpower 17(8):30–45

Patten S (2005) Markov models of major depression for linking psychiatric epidemiology to clinical practice. Clinical Practice and Epidemiology in Mental Health 1(2):11

Rombaut E, Guerry M (2015) Decision trees as a classification technique in manpower planning. The 16th Conference of the Applied Stochastic Models and Data Analysis International Society pp 863–877

Schwarz G (1978) Estimating the dimension of a model. The Annals of Statistics 6(2):461–464

Sonnenberg FA, Beck JR (1993) Markov models in medical decision making: a practical guide. Medical decision making 13(4):322–338

Uche P (1990) Non-homogeneity and transition probabilities of a Markov chain. International Journal of Mathematical Education in Science and Technology 21(2):295–301

Ugwuowo F, McClean S (2000) Modelling heterogeneity in a manpower system: a review. Applied Stochastic Models in Business and Industry 16(2):99–110

Appendix A Proofs of theorems and lemmas

Let us recall the definition of the following functions:

$$\varphi(t) = \begin{cases} t \ln t & \text{if } t > 0 \\ 0 & \text{if } t = 0, \end{cases} \tag{13}$$

$$\phi(t_1, t_2) = \varphi(t_1) + \varphi(t_2) - \varphi(t_1 + t_2). \tag{12}$$

and

$$\psi(x, y) = \phi(x, y) + \phi(x + y, n_{AB} - y) + \phi(n_{AA} - x - y, y) - \phi(n_{AA}, n_{AB}). \tag{24}$$

Lemma 1 $\ln \hat{L}_{\mathcal{S}} = \phi(n_{11}, n_{12}) + \phi(n_{11} + n_{12}, n_{13}) + \phi(n_{22}, n_{23})$

Proof Using (4), (5) and the fact that $n_2 - n_{23} = n_{22}$, we obtain

$$\begin{aligned}\ln \hat{L}_{\mathcal{S}} &= \varphi(n_{11}) + \varphi(n_{12}) + \varphi(n_{13}) + \varphi(n_{22}) + \varphi(n_{23}) - \varphi(n_1) - \varphi(n_2) \\ &= \varphi(n_{11}) + \varphi(n_{12}) - \varphi(n_{11} + n_{12}) \\ &\quad + \varphi(n_{11} + n_{12}) + \varphi(n_{13}) - \varphi(n_1) \\ &\quad + \varphi(n_{22}) + \varphi(n_{23}) - \varphi(n_2) \\ &= \phi(n_{11}, n_{12}) + \phi(n_{11} + n_{12}, n_{13}) + \phi(n_{22}, n_{23}).\end{aligned}$$

Lemma 2 $\ln \hat{L}_{\tilde{\mathcal{S}}} = \phi(n_{11} + n_{12}, n_{13}) + \phi(n_{22}, n_{23})$

Proof It follows from (10) that

$$\hat{L}_{\tilde{\mathcal{S}}} = \frac{(n_{11} + n_{12})^{n_{11}+n_{12}}\, {n_{13}}^{n_{13}}\, {n_{22}}^{n_{22}}\, {n_{23}}^{n_{23}}}{{n_1}^{n_1}\, {n_2}^{n_2}},$$

using $n_1 = n_{11} + n_{12} + n_{13}$ and $n_2 = n_{22} + n_{23}$. Hence,

$$\begin{aligned}\ln \hat{L}_{\tilde{\mathcal{S}}} &= \varphi(n_{11} + n_{12}) + \varphi(n_{13}) + \varphi(n_{22}) + \varphi(n_{23}) - \varphi(n_1) - \varphi(n_2) \\ &= \big(\varphi(n_{11} + n_{12}) + \varphi(n_{13}) - \varphi(n_1)\big) + \big(\varphi(n_{22}) + \varphi(n_{23}) - \varphi(n_2)\big) \\ &= \phi(n_{11} + n_{12}, n_{13}) + \phi(n_{22}, n_{23}).\end{aligned}$$

Lemma 3 *The function ϕ, defined by* (12)*, is homogeneous of the first degree, i.e.*

$$\phi(tx, ty) = t\,\phi(x, y) \quad \textit{for all } t, x, y \geq 0$$

Proof By (13), we have that $\varphi(tu) = t\,\varphi(u) + u\,\varphi(t)$ for all $t, u \geq 0$. The result then follows from (12).

Lemma 4 *The function ϕ, defined by* (12)*, is convex.*

Proof Let $u, v > 0$. We prove that the Hessian matrix of ϕ at (u, v), denoted H, is positive semi-definite. Using standard calculus,

$$H = \begin{bmatrix} \frac{v}{u(u+v)} & -\frac{1}{u+v} \\ -\frac{1}{u+v} & \frac{u}{v(u+v)} \end{bmatrix} = \frac{1}{uv(u+v)} \begin{bmatrix} v^2 & -uv \\ -uv & u^2 \end{bmatrix}. \tag{23}$$

The eigenvalues of H are 0 and $\frac{u^2+v^2}{uv(u+v)}$. They are both non-negative, hence H is positive semi-definite.

Lemma 5 *For the function ϕ, defined by* (12)*, holds*

$$\phi(a, b) + \phi(c, d) \geq \phi(a + c, b + d) \quad \textit{for all } a, b, c, d \geq 0$$

with equality if and only if $ad = bc$.

Proof The inequality is an immediate consequence of the convexity and homogeneity properties of the function ϕ (lemmas 3 and 4):

$$\phi(a, b) + \phi(c, d) \geq 2\,\phi(\tfrac{a+c}{2}, \tfrac{b+d}{2}) = \phi(a + c, b + d).$$

Suppose $ad = bc$. If $ad = 0 = bc$, then equality holds surely because $\phi(u, v) = 0$ if $u = 0$ or $v = 0$. If $ad = bc \neq 0$, then $c = ta$ and $d = tb$ for some number $t \neq 0$. Hence, by the homogeneity of ϕ,

$$\begin{aligned}\phi(a, b) + \phi(c, d) &= \phi(a, b) + \phi(ta, tb) \\ &= (1 + t)\phi(a, b) \\ &= \phi((1 + t)a, (1 + t)b) = \phi(a + c, b + d).\end{aligned}$$

Now suppose equality holds. Then, by homogeneity of ϕ,

$$\frac{\phi(a,b)+\phi(c,d)}{2} = \phi(\tfrac{a+c}{2}, \tfrac{b+d}{2}).$$

Consequently, the function $f(t) = \phi(u,v)$ with $u = a + t(c-a)$ and $v = b + t(d-b)$ is linear on $[0,1]$, because ϕ is convex. So, $f''(t) = 0$ for all $t \in (0,1)$. Using the Hessian matrix of ϕ in (23), we obtain

$$\begin{aligned} f''(t) &= (c-a)^2 H_{11} + 2(c-a)(d-b)H_{12} + (d-b)^2 H_{22} \\ &= \frac{[(b-d)u - (c-a)v]^2}{uv(u+v)} = \frac{(ad-bc)^2}{uv(u+v)}, \end{aligned}$$

and the result $ad = bc$ thus follows from $f''(t) = 0$.

Lemma 6 *For the function ϕ, as defined in (12), holds that $\phi_y : t \mapsto \phi(t,y)$ is strictly decreasing and strictly convex for all $y > 0$.*

Proof Using standard calculus, ${\phi_y}'(t) = \ln t - \ln(t+y) < 0$ and ${\phi_y}''(t) = \frac{y}{t(t+y)} > 0$, if $t > 0$.

Lemma 7 *If $n_{AA} > 2n_{AB}$, it holds that $\psi(n_{AA} - \frac{3}{2}n_{AB}, n_{AB}) < 0$.*

Proof Let $\alpha = n_{AA}/n_{AB}$. By (24) and Lemma 3,

$$\psi(n_{AA} - \tfrac{3}{2}n_{AB}, n_{AB}) = h(\alpha) n_{AB}$$

where

$$h(t) = \phi_1(t - \tfrac{3}{2}) + \phi_1(\tfrac{1}{2}) - \phi_1(t)$$

and ϕ_1 is the function $u \mapsto \phi(u,1)$. Since ϕ_1 is strictly convex (Lemma 6), its first derivative is strictly increasing and therefore h is strictly decreasing. Furthermore,

$$h(2) = \phi(\tfrac{1}{2},1) + \phi(\tfrac{1}{2},1) - \phi(2,1) = 2\phi(\tfrac{1}{2},1) - \phi(2,1) = 0$$

by Lemma 3. Consequently, $h(\alpha) < 0$ since $\alpha > 2$.

Lemma 8 *If $n_{AA} > 2n_{AB}$, it holds that $\psi(n_{AA} - \frac{3}{4}n_{AB}, \frac{3}{4}n_{AB}) < 0$.*

Proof Let $\alpha = n_{AA}/n_{AB}$ and $\beta = 3/4$. By (24) and Lemma 3,

$$\psi(n_{AA} - \beta n_{AB}, \beta n_{AB}) = h(\alpha) n_{AB}$$

where

$$h(t) = \phi(t - \beta, \beta) + \phi(t, 1-\beta) - \phi(t,1).$$

With the use of the function $\phi_1 : u \mapsto \phi(u,1)$ and equations (13) and (12) we can rewrite $h(t)$ as

$$h(t) = \phi_1(t-\beta) - \phi_1(t) + \phi(\beta, 1-\beta).$$

Since ϕ_1 is strictly convex (Lemma 6), its first derivative is strictly increasing and therefore h is strictly decreasing. The result now follows from the fact that $h(2) < 0$.

Lemma 9 *If $n_{AA} \geq 3n_{AB}$, it holds that $\psi(n_{AB}, n_{AB}) < 0$.*

Proof Let $\alpha = n_{AA}/n_{AB}$. By (24) and Lemma 3,

$$\psi(n_{AB}, n_{AB}) = h(\alpha) n_{AB}$$

where

$$h(t) = \phi(1,1) + \phi_1(t-2) - \phi_1(t)$$

and ϕ_1 is the function $u \mapsto \phi(u,1)$. Since ϕ_1 is strictly convex (Lemma 6), its first derivative is strictly increasing and therefore h is strictly decreasing. Furthermore, $h(3) = \ln \frac{16}{27} < 0$. Hence, since $\alpha \geq 3$ the monotonicity of h yields $h(\alpha) < 0$ and the result follows.

Lemma 10 *If $n_{AA} > 2n_{AB}$, it holds that $\psi(\frac{n_{AA}-n_{AB}}{2}, n_{AB}) < 0$.*

Proof Let $\alpha = n_{AA}/n_{AB}$. By (24) and lemmas 3 and 6,

$$\psi(\tfrac{n_{AA}-n_{AB}}{2}, n_{AB}) = h(\alpha) n_{AB},$$

where

$$h(t) = \phi(t-1, 2) - \phi(t, 1).$$

Using standard calculus,

$$h'(t) = \ln(t-1) - \ln t < 0,$$

so that the function h is strictly decreasing. Furthermore, $h(2) = 0$ which can be verified by straightforward computation. Hence, since $\alpha > 2$ the monotonicity of h yields $h(\alpha) < 0$ and the result follows.

Lemma 11 $\psi(0, \frac{n_{AA}n_{AB}}{n_{AA}+n_{AB}}) = 0$.

Proof Denote $\rho = \frac{n_{AA}n_{AB}}{n_{AA}+n_{AB}}$. Then, $n_{AB} - \rho = \frac{n_{AB}}{n_{AA}}\rho$ and $n_{AA} - \rho = \frac{n_{AA}}{n_{AB}}\rho$, so that, using (24) and Lemma 3,

$$\begin{aligned}
\psi(0, \rho) &= \phi(\theta, n_{AB} - \rho) + \phi(n_{AA} - \rho, \rho) - \phi(n_{AA}, n_{AB}) \\
&= \phi(\rho, \tfrac{n_{AB}}{n_{AA}}\rho) + \phi(\tfrac{n_{AA}}{n_{AB}}\rho, \rho) - \phi(n_{AA}, n_{AB}) \\
&= \tfrac{\rho}{n_{AA}}\phi(n_{AA}, n_{AB}) + \tfrac{\rho}{n_{AB}}\phi(n_{AA}, n_{AB}) - \phi(n_{AA}, n_{AB}) \\
&= \left(\tfrac{\rho}{n_{AA}} + \tfrac{\rho}{n_{AB}} - 1\right) \phi(n_{AA}, n_{AB}) = 0.
\end{aligned}$$

Lemma 12 $\psi(0, n_{AB}) = \psi(n_{AA} - n_{AB}, n_{AB}) > 0$.

Proof By (24), $\psi(0, n_{AB})$ and $\psi(n_{AA} - n_{AB}, n_{AB})$ are both equal to $\phi_{n_{AB}}(n_{AA} - n_{AB}) - \phi_{n_{AB}}(n_{AA})$, where $\phi_{n_{AB}}$ is the function $u \mapsto \phi(u, n_{AB})$. According to Lemma 6, $\phi_{n_{AB}}$ is strictly decreasing and the result follows.

Lemma 13 *For an AB-complete data set, we have that $\hat{\tau}_{23} > \hat{p}_{AB}$ is equivalent to $g(n_{11}, n_{12}) > 0$, where*

$$g(x, y) = n_{AB}x + (n_{AA} + n_{AB})y - n_{AA}n_{AB}.$$

Hence, if $\hat{\tau}_{23} > \hat{p}_{AB}$, the point (n_{11}, n_{12}) in xy-plane lies above the line through the points $(0, \frac{n_{AA}n_{AB}}{n_{AA}+n_{AB}})$ and $(n_{AA}, 0)$.

Proof Because of AB-completeness, it holds that $n_{23} = n_{12}$. Furthermore, $n_2 = n_{22} + n_{23} = n_{22} + n_{12} = n_{AA} - n_{11}$. Consequently, since $\hat{\tau}_{23} = n_{23}/n_2$ and $\hat{p}_{AB} = n_{AB}/(n_{AA} + n_{AB})$, we have

$$\begin{aligned}
\hat{\tau}_{23} > \hat{p}_{AB} &\Leftrightarrow \frac{n_{12}}{n_{AA} - n_{11}} > \frac{n_{AB}}{n_{AA} + n_{AB}} \\
&\Leftrightarrow n_{12}(n_{AA} + n_{AB}) > (n_{AA} - n_{11})n_{AB} \\
&\Leftrightarrow g(n_{11}, n_{12}) > 0.
\end{aligned}$$

Proof of theorem 1

Proof In expression (17), we can eliminate the variables n_{13}, n_{22} and n_{23}, since $n_{AA} = n_{11} + n_{12} + n_{22}$, and, by the AB-completeness assumption, $n_{AB} = n_{13} + n_{23}$ and $n_{12} = n_{23}$. Hence, $\Delta\ell\ell_{\mathrm{M}(k)\text{-M}} = \psi(n_{11}, n_{12})$, where

$$\psi(x, y) = \phi(x, y) + \phi(x + y, n_{AB} - y) + \phi(n_{AA} - x - y, y) - \phi(n_{AA}, n_{AB}). \tag{24}$$

The function ψ is convex, since each of the first three terms in (24) is a convex function of (x, y), as a composition of the convex function ϕ (Lemma 4) and an affine function from $\mathbb{R}^2$ to $\mathbb{R}^2$.

Let $\alpha = n_{AA} - n_{AB}$. By assumption, $\alpha > 0$. Let $\beta = \min\{n_{AB}, \frac{\alpha}{2}\}$ and denote the following points in xy-plane: $\mathbf{a}(0, \frac{n_{AA}n_{AB}}{n_{AA}+n_{AB}})$, $\mathbf{b}(\beta, n_{AB})$, $\mathbf{c}(\alpha, n_{AB})$, $\mathbf{d}(0, n_{AA})$, $\mathbf{e}(n_{AA} - \frac{3}{2}n_{AB}, n_{AB})$ and $\mathbf{f}(n_{AA} - \frac{3}{4}n_{AB}, \frac{3}{4}n_{AB})$. Because $n_{AA} > 2n_{AB}$, the 3-simplex **ecf** is a subset of the 4-simplex **abcd**, see figure 3.

Take $k \geq 1$ sufficiently large, so that $\hat{\tau}_{23} > \hat{p}_{AB}$. Then, using Lemma 13 and the fact that $n_{11} \geq n_{12}$ whenever $k \geq 1$, the point $\mathbf{p}(n_{11}, n_{12})$ in xy-plane must be contained in the 4-simplex **abcd**.

Suppose $\Delta\ell\ell_{\mathrm{M}(k)\text{-M}} \geq 0$. We prove that $\mathbf{p}$ must then belong to the 3-simplex **ecf**. First, we observe that ψ is non-positive on the 5-simplex **abefd**, because ψ is convex and ψ is non-positive in all vertices of **abefd** (lemmas 11, 9, 10, 7, 8). Hence, because $\mathbf{p} \in$ **abcd** and $\psi(\mathbf{p}) = \Delta\ell\ell_{\mathrm{M}(k)\text{-M}} \geq 0$, we have $\mathbf{p} \in$ **ecf**. Consequently, $n_{11} > n_{AA} - \frac{3}{2}n_{AB}$ and $n_{12} > \frac{3}{4}n_{AB}$. But then, (20) cannot be satisfied, since $n_{AA} - n_{11} = n_{22} + n_{12} = n_{22} + n_{23} = n_2$.

Proof of theorem 2

Proof Denote $n_{11} + n_{12} = a$, $n_{13} = b$, $n_{22} = c$ and $n_{23} = d$. Then, $n_{AA} = a + c$ and $n_{AB} = b + d$. Hence, using (18), we have

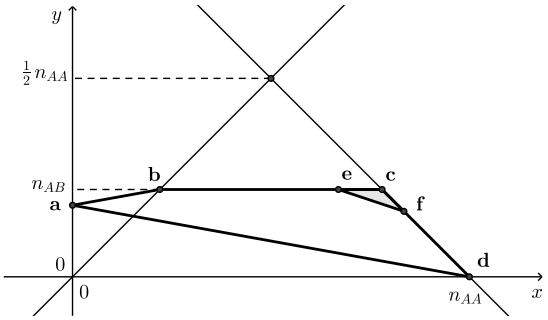
$$\Delta\ell\ell_{\widetilde{\mathrm{M}(k)\text{-M}}} = \phi(a, b) + \phi(c, d) - \phi(a + c, b + d),$$

which is non-negative by Lemma 5.

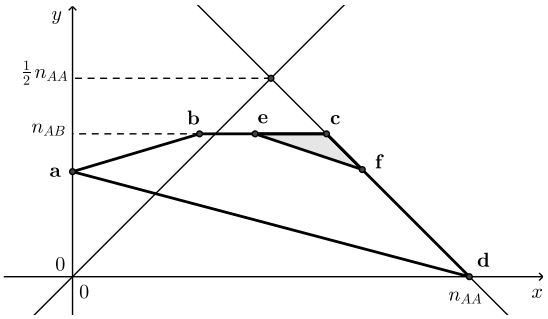
Furthermore, $\Delta\ell\ell_{\widetilde{\mathrm{M}(k)\text{-M}}} = 0$ is equivalent to $\phi(a, b) + \phi(c, d) = \phi(a + c, b + d)$, which in turn is equivalent to $ad = bc$ by Lemma 5. Finally, $\hat{\tau}_{13} = \hat{\tau}_{23} = \hat{p}_{AB}$ if and only if $ad = bc$, since $\hat{\tau}_{13} = \frac{b}{a+b}$, $\hat{\tau}_{23} = \frac{d}{c+d}$ and $\hat{p}_{AB} = \frac{b+d}{a+c+b+d}$.

Proof of theorem 3

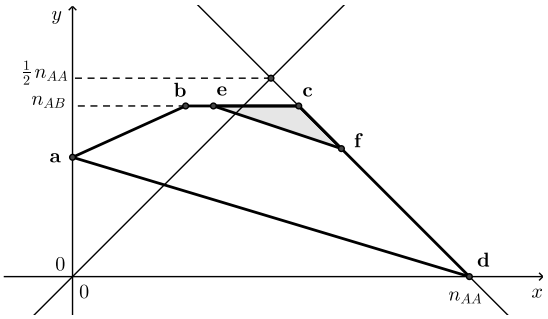
Proof If $n_{11} > 0$ and $n_{12} > 0$, we have by the binomial theorem that $n_{11}^{n_{11}} n_{12}^{n_{12}} < (n_{11} + n_{12})^{n_{11}+n_{12}}$, hence $\phi(n_{11}, n_{12}) < 0$ using (12). The theorem now follows from (19) and the fact that $\phi(n_{11}, n_{12}) = 0$ if $n_{11} = 0$ or $n_{12} = 0$.



(a) case: $n_{AB} \leq \frac{1}{3}\,n_{AA}$



(b) case: $\frac{1}{3}\,n_{AA} < n_{AB} \leq \frac{2}{5}\,n_{AA}$



(c) case: $\frac{2}{5}\,n_{AA} < n_{AB} < \frac{1}{2}\,n_{AA}$

Fig. 3: Location of the points **a** through **f** in the xy-plane