J. R. Statist. Soc. B (2005)
67, *Part* 3, *pp.* 395–410

Statistical inference for discretely observed Markov jump processes

Mogens Bladt

Universidad Nacional Autónoma de México, México

and Michael Sørensen

University of Copenhagen, Denmark

[Received September 2003. Final revision January 2005]

Summary. Likelihood inference for discretely observed Markov jump processes with finite state space is investigated. The existence and uniqueness of the maximum likelihood estimator of the intensity matrix are investigated. This topic is closely related to the imbedding problem for Markov chains. It is demonstrated that the maximum likelihood estimator can be found either by the EM algorithm or by a Markov chain Monte Carlo procedure. When the maximum likelihood estimator does not exist, an estimator can be obtained by using a penalized likelihood function or by the Markov chain Monte Carlo procedure with a suitable prior. The methodology and its implementation are illustrated by examples and simulation studies.

Keywords: EM algorithm; Imbedding problem; Likelihood inference; Markov chain Monte Carlo methods

1. Introduction

Markov jump processes with finite state space have many applications and, if a continuous record of such a process has been observed, likelihood inference concerning the transition intensities is simple and well known; see for example Billingsley (1961), Jacobsen (1982) and Küchler and Sørensen (1997). If a Markov jump process is observed only at discrete time points, the situation is more complex. Discretely observed diffusion processes have been studied intensively in the last decade. A few recent references are Kessler and Sørensen (1999), Hoffmann (1999), Roberts and Stramer (2001), Elerian *et al.* (2001), Aït-Sahalia (2002) and Bibby *et al.* (2004). For Markov jump processes not much research has been done on the discretely sampled case. Discretely sampled birth processes and birth-and-death processes were investigated in Keiding (1974, 1975). An important application of Markov jump processes in mathematical finance is in credit risk modelling, where the transitions between different credit ratings are modelled by a Markov jump process; see Jarrow *et al.* (1997). This led Israel *et al.* (1997) to propose a method of estimating the jump intensities from discrete time observations. Their method is not efficient, however, and *ad hoc* modification of the estimator is required to obtain an intensity matrix.

In this paper we discuss the problems that are related to maximum likelihood estimation of the intensity matrix based on a discretely sampled Markov jump process and demonstrate that the maximum likelihood estimator can be found either by the EM algorithm or by a Markov chain Monte Carlo (MCMC) procedure. It is possible that the maximum likelihood estimator

Address for correspondence: Michael Sørensen, Department of Applied Mathematics and Statistics, University of Copenhagen, Universitetsparken 5, DK-2100 Copenhagen Ø, Denmark.
E-mail: michael@math.ku.dk

1369–7412/05/67395

does not exist, but this problem can be overcome by using a penalized likelihood function or the MCMC estimator with a suitable prior.

The problems of identifiability and of existence and uniqueness of the maximum likelihood estimator are closely related to a classical problem in probability theory: the imbedding problem for Markov chains. This is the question about whether a given discrete time Markov chain can be obtained by discrete time sampling of a continuous time Markov jump process. In Section 2 we review results on the imbedding problem that we need for our discussion of maximum likelihood estimation. We also present the various likelihood functions that are used in later sections, give a result on existence and uniqueness of the maximum likelihood estimator and study in detail the instructive case of a two-state process where the problem of possible non-existence of the maximum likelihood estimator can be discussed explicitly. In Section 3 we demonstrate how the EM algorithm can be implemented and give a result on the convergence of the algorithm. The problems of non-existence of the maximum likelihood estimator can be avoided by using the MCMC procedure that is presented in Section 4. In fact, a Gibbs sampler with a conjugate prior turns out to be sufficient to solve the problem. In Section 5 implementation problems are discussed and illustrated by examples and simulation studies. In particular, it is demonstrated that the proposed MCMC methodology is applicable even when the number of states is as large as 50.

2. The likelihood function

Let X be a Markov jump process with finite state space $E=\{1,\ldots,m\}$ and intensity matrix (infinitesimal generator) $\mathbf{Q}=\{q_{ij}\}$. If X has been observed continuously in the time interval $[0,\tau]$, i.e. if the data are $\{X(t)|0\leqslant t\leqslant\tau\}$, maximum likelihood estimation of $\mathbf{Q}$ is an easy task that has been considered by several researchers (e.g. Billingsley (1961), Jacobsen (1982) and Küchler and Sørensen (1997)). The likelihood function is given by

$$L_\tau^{(\mathrm{c})}(\mathbf{Q})=\prod_{i=1}^{m}\prod_{j\neq i} q_{ij}^{N_{ij}(\tau)}\exp\{-q_{ij}R_i(\tau)\}. \tag{2.1}$$

The superscript (c) indicates continuous time observation. The process $N_{ij}(t)$ is the number of transitions from state i to state j in the time interval $[0,t]$, whereas

$$R_i(t)=\int_0^t I\{X(s)=i\}\,\mathrm{d}s \tag{2.2}$$

is the time that is spent in state i before time t. For details see for example Jacobsen (1982). It is not difficult to see that the maximum likelihood estimator of $\mathbf{Q}$ is

$$\hat{q}_{ij}^{(\mathrm{c})}(\tau)=N_{ij}(\tau)/R_i(\tau), \tag{2.3}$$

provided, of course, that $R_i(\tau)>0$. If the process has not been in state i, there is no information about q_{ij} in the data, and the maximum likelihood estimator of q_{ij} does not exist.

The continuous observation likelihood function will play a role in later sections, but in the present paper we are mainly interested in inference about the intensity matrix $\mathbf{Q}$ based on a sample of observations of X at discrete time points, i.e. $\{X(t_1),\ldots,X(t_n)\}$. Also for discrete time observations the likelihood function is in theory simple. The process $Y_i=X(t_i)$ is a discrete time Markov chain, in general time inhomogeneous, for which the transition matrix at time i is $P^{\Delta_i}(\mathbf{Q})$, where $\Delta_i=t_{i+1}-t_i$ and

$$P^t(\mathbf{Q})=\exp(t\mathbf{Q}),\qquad t>0, \tag{2.4}$$

with exp(·) denoting the matrix exponential function. Hence the likelihood function for the discrete time data is given by

$$L_n(\mathbf{Q}) = \prod_{i=1}^{n-1} P^{\Delta_i}(\mathbf{Q})_{x_i x_{i+1}}, \qquad \mathbf{Q} \in \mathcal{Q}, \tag{2.5}$$

where $x_1, \ldots, x_n$ denote the observed values of X. For a matrix A we denote the ijth entry by A_{ij}. The set of all intensity matrices is denoted by $\mathcal{Q}$. This is the set of matrices for which the off-diagonal entries are non-negative and the sum of the entries in each row equals 0. In the case of equidistant observation times, i.e. when $\Delta_i = \Delta$ for some $\Delta > 0$, the Markov chain Y is time homogeneous with transition matrix $P^{\Delta}(\mathbf{Q})$, so the likelihood function simplifies somewhat to

$$L_n(\mathbf{Q}) = \prod_{i=1}^{m} \prod_{j=1}^{m} P^{\Delta}(\mathbf{Q})_{ij}^{K_{ij}(n)}, \qquad \mathbf{Q} \in \mathcal{Q}, \tag{2.6}$$

where $K_{ij}(n)$ is the number of transitions from state i to state j in the discrete time Markov chain $\{X(t_1), \ldots, X(t_n)\}$. We shall mainly consider the case of equidistant observation times.

For the full class of time homogeneous Markov chains with state space $\{1, \ldots, m\}$, the likelihood function based on observations of the state of the chain at the first n time points is

$$L_n(\mathbf{P}) = \prod_{i=1}^{m} \prod_{j=1}^{m} \mathbf{P}_{ij}^{K_{ij}(n)}, \qquad \mathbf{P} \in \mathcal{P}, \tag{2.7}$$

where $K_{ij}(n)$ is again the number of transitions from i to j before time n, and where $\mathcal{P}$ denotes the set of $m \times m$ transition matrices (stochastic matrices), i.e. $(m \times m)$-matrices with non-negative entries for which the sum of the entries in each row is equal to 1. This likelihood function is identical to the likelihood function for m independent multinomial distributions, so the maximum likelihood estimator of the parameter $\mathbf{P}$ is

$$\hat{\mathbf{P}}_{ij} = K_{ij}(n)/K_{i.}(n) \tag{2.8}$$

where

$$K_{i.}(n) = \sum_{j=1}^{m} K_{ij}(n).$$

Define

$$\mathcal{P}_0 = \{\exp(\mathbf{Q}) | \mathbf{Q} \in \mathcal{Q}\}, \tag{2.9}$$

the set of transition matrices that correspond to discrete time observation of a continuous time Markov jump process. Now suppose that we calculate $\hat{\mathbf{P}}$ by equation (2.8) based on our discrete time observations of a continuous time Markov jump process. If $\hat{\mathbf{P}} \in \mathcal{P}_0$, there is a $\hat{\mathbf{Q}} \in \mathcal{Q}$ such that $P^{\Delta}(\hat{\mathbf{Q}}) = \hat{\mathbf{P}}$, and the likelihood function (2.6) attains its maximal value at $\hat{\mathbf{Q}}$, which is thus the maximum likelihood estimator. There are, however, two problems here. One is that the set $\mathcal{P}_0$ is very complicated (except when $m = 2$); the other is that the matrix exponential function is not an injection in all parts of its domain, so $\hat{\mathbf{Q}}$ need not be unique. When $\hat{\mathbf{P}} \notin \mathcal{P}_0$, the situation is not clear owing to the complicated structure of $\mathcal{P}_0$, but it seems not to be uncommon that the maximum likelihood estimator does not exist, in particular when the time between observations Δ is large. General results on the existence and uniqueness of the maximum likelihood estimator are summarized in theorem 1 below, in particular, the probability that $\hat{\mathbf{P}} \in \mathcal{P}_0$ goes to 1 as $n \to \infty$. We shall give a complete discussion of the case $m = 2$, where the maximum likelihood estimator does not exist when $\hat{\mathbf{P}} \notin \mathcal{P}_0$.

The problem of identifying the set $\mathcal{P}_0$ has a long history and was first posed by Elfving (1937). It is usually referred to as the *imbedding problem* for finite Markov chains. Kingman (1962) showed that $\mathcal{P}_0 = \mathcal{P}_+$ when $m = 2$, where

$$\mathcal{P}_+ = \{\mathbf{P} \in \mathcal{P} | \det(\mathbf{P}) > 0\},$$

and derived the following general results about $\mathcal{P}_0$. For $m \geqslant 3$, $\mathcal{P}_0$ is a (relatively) closed subset of $\mathcal{P}_+$ with a complex geometric shape. In particular, it is not convex. Its relative interior as a subset of $\mathcal{P}$ is non-empty, so its dimension is $m(m-1)$. Let $\delta\mathcal{P}_0$ denote the boundary of $\mathcal{P}_0$ relative to $\mathcal{P}_+$. Then

$$\delta\mathcal{P}_0 = (\cup_{i \neq j} E_{ij}) \cup \mathcal{E}, \tag{2.10}$$

where E_{ij} is a non-empty subset of the set of exponentials of intensity matrices with $q_{ij} = 0$, and $\mathcal{E}$ is a non-empty subset of the $m \times m$ transition matrices with fewer than m distinct eigenvalues. For details see Kingman (1962). Johansen (1974) gave an explicit description of $\mathcal{P}_0$ for $m = 3$, which already at this low dimension is somewhat complicated.

The second problem is whether there are two or more intensity matrices $\mathbf{Q}$ for which the corresponding transition matrix $\exp(\Delta\mathbf{Q})$ is the same, i.e. do two or more continuous time Markov jump processes exist for which the discrete time sample $(X(\Delta), \ldots, X(n\Delta))$ has the same distribution? In statistical terms this is the question of whether or not the parameterization of the distribution of the data $X(\Delta), \ldots, X(n\Delta)$ by $\mathbf{Q}$ is identifiable. Let $\mathcal{P}_{00}$ denote the subset of $\mathcal{P}_0$ of transition matrices $\mathbf{P} \in \mathcal{P}_0$, for which $\mathbf{Q}$ is uniquely determined by $\mathbf{P} = \exp(\mathbf{Q})$. For $m = 2$, $\mathcal{P}_{00} = \mathcal{P}_0 = \mathcal{P}_+$. The characterization of the set $\mathcal{P}_{00}$ is the classical problem of when the real logarithm of a matrix is unique, which was solved for general matrices by Culver (1966). His general result is that $\mathcal{P}_{00}$ consists of the transition matrices $\mathbf{P} \in \mathcal{P}_0$, for which all eigenvalues of $\mathbf{P}$ are positive and no elementary divisor (Jordan block) of $\mathbf{P}$ belonging to any eigenvalue appears more than once. Thus, once $\hat{\mathbf{P}}$ has been calculated from equation (2.8), it is in principle easy to check whether it determines an estimator of the intensity matrix uniquely (provided that $\hat{\mathbf{P}} \in \mathcal{P}_0$). If $\mathbf{P} \notin \mathcal{P}_{00}$, there are infinitely many solutions $\mathbf{X}$ to the equation $\mathbf{P} = \exp(\mathbf{X})$, not all of which are intensity matrices. The set of solutions is countable if all real eigenvalues of $\mathbf{P}$ are positive with their Jordan blocks appearing only once and any complex eigenvalue belongs to only one Jordan block. Otherwise there are uncountably many solutions. Cuthbert (1973) showed that in the countable case only a finite subset of the solutions are intensity matrices.

Simple necessary conditions for a transition matrix $\mathbf{P}$ to belong to $\mathcal{P}_{00}$ were given by Cuthbert (1972, 1973). A simple, but crude, condition for $\mathbf{P} \in \mathcal{P}_0$ to belong to $\mathcal{P}_{00}$ is that

$$\inf_i (\mathbf{P}_{ii}) \geqslant \tfrac{1}{2}. \tag{2.11}$$

A less crude criterion for $\mathbf{P} \in \mathcal{P}_0$ to belong to $\mathcal{P}_{00}$ is that

$$\inf_i (\mathbf{P}_{ii}) \cdot \det(\mathbf{P}) > \exp(-\pi) \prod_i \mathbf{P}_{ii}; \tag{2.12}$$

see Cuthbert (1973) ($\exp(-\pi) \simeq 0.0432$).

We can now summarize the results on existence and uniqueness of the maximum likelihood estimator.

Theorem 1. If $\hat{\mathbf{P}}$ given by equation (2.8) belongs to $\mathcal{P}_0$, then the maximum likelihood estimator of the intensity matrix $\hat{\mathbf{Q}}$ exists and is the solution to $\hat{\mathbf{P}} = \exp(\Delta\hat{\mathbf{Q}})$. If $\hat{\mathbf{P}} \notin \mathcal{P}_0$, then either the maximum likelihood estimator $\hat{\mathbf{Q}}$ exists and satisfies the condition that $\exp(\Delta\hat{\mathbf{Q}}) \in \delta\mathcal{P}_0$ (given by equation (2.10)), or the likelihood function given by equation (2.6) has no maximum in $\mathcal{Q}$.

If the true transition matrix $\mathbf{Q}_0$ satisfies the condition that $\exp(\Delta\mathbf{Q}_0)\in \mathrm{int}(\mathcal{P}_0)$, and if the Markov process is ergodic, then the probability that the maximum likelihood estimator exists goes to 1 as $n\to\infty$, and $\exp(\Delta\hat{\mathbf{Q}})\to\exp(\Delta\mathbf{Q}_0)$ almost surely. Moreover, if $\mathbf{Q}_0$ satisfies the condition that $\exp(\Delta\mathbf{Q}_0)\in \mathrm{int}(\mathcal{P}_{00})$, then the probability that the maximum likelihood estimator is unique goes to 1 and $\hat{\mathbf{Q}}\to\mathbf{Q}_0$ almost surely as $n\to\infty$. The condition $\exp(\Delta\mathbf{Q}_0)\in\mathrm{int}(\mathcal{P}_{00})$ is satisfied when Δ is sufficiently small.

Proof. The situation where $\hat{\mathbf{P}}\in\mathcal{P}_0$ is trivial and was discussed above. Next assume that $\hat{\mathbf{P}}\notin\mathcal{P}_0$ and define the set

$$\mathcal{P}_c=\{\mathbf{P}\in\mathcal{P}|\log\{L(\mathbf{P})\}\geqslant -c\},$$

where $L(\mathbf{P})$ is the likelihood function for the full class of Markov chains given by equation (2.7) and $c>0$. Consider the compact set $\mathcal{P}_c\cap\bar{\mathcal{P}}_0$ for a $c>0$ that is sufficiently large that $\mathcal{P}_c\cap\bar{\mathcal{P}}_0$ is not empty. Here $\bar{\mathcal{P}}_0$ denotes the set $\bar{\mathcal{P}}_0=\mathcal{P}_0\cup\{\mathbf{P}\in\mathcal{P}|\det(\mathbf{P})=0\}$. The continuous function $L(\mathbf{P})$ has a maximum $\tilde{\mathbf{P}}$ in $\mathcal{P}_c\cap\bar{\mathcal{P}}_0$ and, since $L(\mathbf{P})$ increases whenever $\mathbf{P}$ is moved in the direction of $\hat{\mathbf{P}}$, $\tilde{\mathbf{P}}$ is on the boundary of $\mathcal{P}_c\cap\bar{\mathcal{P}}_0$. Thus either $\tilde{\mathbf{P}}\in\delta\mathcal{P}_0$, in which case there is a $\mathbf{Q}$ such that $\exp(\Delta\hat{\mathbf{Q}})=\tilde{\mathbf{P}}$ (remember that $\mathcal{P}_0$ is closed relative to $\mathcal{P}_+$), or $\det(\tilde{\mathbf{P}})=0$, in which case the likelihood function does not have a maximum in $\mathcal{Q}$.

Now assume that $\exp(\Delta\mathbf{Q}_0)\in\mathrm{int}(\mathcal{P}_0)$. From a well-known result for Markov processes (see for example theorem 1.1 in Billingsley (1961)), we know that $\hat{\mathbf{P}}\to\exp(\Delta\mathbf{Q}_0)\in\mathrm{int}(\mathcal{P}_0)$ almost surely as $n\to\infty$. Therefore the probability that $\hat{\mathbf{P}}\in\mathrm{int}(\mathcal{P}_0)$ goes to 1 as $n\to\infty$. The claim about uniqueness and consistency of the maximum likelihood estimator is shown in the same way. That $\exp(\Delta\hat{\mathbf{Q}})\in\mathrm{int}(\mathcal{P}_{00})$ when Δ is sufficiently small follows from inequality (2.11). □

The situation that $\det(\tilde{\mathbf{P}})=0$, where the maximum likelihood estimator does not exist, is more likely to happen when the determinant of $\exp(\Delta\mathbf{Q}_0)$ is close to 0. When the Markov process is ergodic, $\exp(\Delta\mathbf{Q}_0)$ converges as $\Delta\to\infty$ to the singular matrix, where all rows are equal to the row vector $\boldsymbol{\pi}$ given by $\boldsymbol{\pi}\mathbf{Q}_0=0$ (the stationary distribution). Hence the propensity of the maximum likelihood estimator not to exist increases with Δ (at least when Δ is sufficiently large).

For a finite sample size all that we can say for sure about uniqueness is that the maximum likelihood estimator is unique when $\hat{\mathbf{P}}\in\mathcal{P}_{00}$ and that the maximum likelihood estimator is not unique when $\hat{\mathbf{P}}\in\mathcal{P}_0\backslash\mathcal{P}_{00}$. If $\hat{\mathbf{P}}\notin\mathcal{P}_0$, we cannot be sure that the maximum likelihood estimator is unique, even when $\tilde{\mathbf{P}}\in\mathcal{P}_0$, because of the complicated geometric structure of the set $\mathcal{P}_0$.

2.1. *Example 1*

Let us consider the case of a Markov process with two states in more detail. This case is simpler than when $m>2$ because here $\mathcal{P}_0=\mathcal{P}_+$, but the statistical problems occur at the boundary where $\det(\mathbf{P})=0$, so the two-state example is instructive.

For an intensity matrix

$$\begin{pmatrix}-\alpha & \alpha\\ \beta & -\beta\end{pmatrix},$$

where $\alpha,\beta\geqslant 0$, the eigenvalues are 0 and $-(\alpha+\beta)$. The corresponding transition matrix is

$$P^{\Delta}(\mathbf{Q})=\frac{1}{\alpha+\beta}\begin{pmatrix}\beta+\alpha\exp\{-\Delta(\alpha+\beta)\} & \alpha[1-\exp\{-\Delta(\alpha+\beta)\}]\\ \beta[1-\exp\{-\Delta(\alpha+\beta)\}] & \alpha+\beta\exp\{-\Delta(\alpha+\beta)\}\end{pmatrix}$$

with eigenvalues 1 and $\rho=\exp\{-\Delta(\alpha+\beta)\}$. It is convenient to introduce a new parameterization of the model:

$$\pi_{11} = P^{\Delta}(\mathbf{Q})_{11} = 1 - (1-\rho)\alpha/(\alpha+\beta),$$
$$\pi_{21} = P^{\Delta}(\mathbf{Q})_{21} = (1-\rho)\beta/(\alpha+\beta).$$

We ignore the trivial case where $\alpha = \beta = 0$. The set of parameter values is

$$\Pi_0 = \{(\pi_{11}, \pi_{21}) | 0 \leqslant \pi_{21} < \pi_{11} \leqslant 1\}.$$

Note that Π_0 is a parameterization of $\mathcal{P}_0$, and $\mathcal{P} = [0,1]^2$. The determinant of $P^{\Delta}(\mathbf{Q})$ equals $\pi_{11} - \pi_{21}$, so the diagonal $\pi_{11} = \pi_{21}$ corresponds to the problematic boundary of $\mathcal{P}_0$, where $\det(\mathbf{P}) = 0$. The likelihood function is

$$L(\pi_{11}, \pi_{21}) = \pi_{11}^{K_{11}(n)} (1-\pi_{11})^{K_{12}(n)} \pi_{21}^{K_{21}(n)} (1-\pi_{21})^{K_{22}(n)},$$

so the maximum likelihood estimator of π_{11} is $\hat{\pi}_{11} = K_{11}(n)/K_{1\cdot}(n)$. If

$$K_{21}(n)/K_{2\cdot}(n) < K_{11}(n)/K_{1\cdot}(n),$$

i.e. if $\hat{\mathbf{P}} \in \mathcal{P}_0$, then $\hat{\pi}_{21} = K_{21}(n)/K_{2\cdot}(n)$. Otherwise, the profile likelihood $\tilde{L}(\pi_{21}) = L(\hat{\pi}_{11}, \pi_{21})$, where $0 \leqslant \pi_{21} < \hat{\pi}_{11}$, keeps growing as π_{21} approaches the boundary point $\hat{\pi}_{11}$. Thus in this case the likelihood function does not have a maximum in Π_0, and the maximum likelihood estimator does not exist. This situation is more likely to happen when the true values of π_{21} and π_{11} are close, which happens when $\Delta(\alpha+\beta)$ is large because then both probabilities are close to the probability of state 1 in the stationary distribution, $\beta/(\alpha+\beta)$.

Since $\alpha + \beta = -\log(\pi_{11} - \pi_{21})/\Delta$, we see that the likelihood function grows (slightly) as $\alpha + \beta \to \infty$. If we have reason to believe that $\alpha + \beta$ is not large, we can avoid the problem by penalizing the likelihood with a prior, for instance

$$\phi(\alpha, \beta) \propto \alpha^a \exp(-b\alpha) \beta^c \exp(-d\beta),$$

which is the conjugate prior for the continuous time model with likelihood function (2.1). The exponential functions ensure that the posterior distribution goes to zero at the critical boundary where $\pi_{11} = \pi_{21}$ so that an estimator that maximizes the posterior exists also when

$$K_{21}(n)/K_{2\cdot}(n) \geqslant K_{11}(n)/K_{1\cdot}(n),$$

i.e. when $\hat{\mathbf{P}} \notin \mathcal{P}_0$. This estimator is not explicit but must be found numerically.

The eigenvalues of $\exp(\Delta\mathbf{Q})$ are $\exp(\Delta\lambda_i)$, $i = 1, \ldots, m$, where $\{\lambda_i\}$ are the eigenvalues of $\mathbf{Q}$. Therefore, as $\exp(\Delta\mathbf{Q})$ goes to the critical boundary, where $\det\{\exp(\Delta\mathbf{Q})\} \to 0$, one or more of the eigenvalues of $\mathbf{Q}$ must go to $-\infty$ (Δ is fixed). Therefore the idea that is presented in example 1 of penalizing the likelihood function (2.6), which is bounded, by the conjugate prior for the continuous time likelihood function (2.1) will in general ensure that there are no problems with existence of an estimator that maximizes the posterior. A general MCMC method along these lines is presented in Section 4.

Asymptotic normality of the maximum likelihood estimator can be established by standard arguments, or follows from results in Billingsley (1961), provided that $\exp(\Delta\mathbf{Q}_0) \in \text{int}(\mathcal{P}_{00})$, that $(\mathbf{Q}_0)_{ij} > 0$ for $i \neq j$ and that the process is ergodic. As earlier $\mathbf{Q}_0$ denotes the true intensity matrix. The expression for the asymptotic variance of the maximum likelihood estimator is very complicated and involves infinite sums. If the maximum likelihood estimator is found by the EM algorithm that is discussed in the following section, the Fisher information matrix can be calculated by means of a formula that was given by Oakes (1999). If $(\mathbf{Q}_0)_{ij} = 0$ for one or more pairs $i \neq j$, a result about asymptotic normality of the maximum likelihood estimator can be obtained if the parameter space is reduced by fixing these intensities at zero, provided that the process is still irreducible.

3. The expectation–maximization algorithm

For a discretely sampled Markov jump process it is natural to use the EM algorithm for optimizing the likelihood function: there is a simple expression for the maximum likelihood estimator when complete continuous time data $X=\{X(t)|0\leqslant t\leqslant\tau\}$ are observed, but only the partial data $Y_i=X(t_i)$, $i=1,\ldots,n$, are available. Here $t_1=0$ and $t_n=\tau$. The difficult step in the EM algorithm is the E-step, i.e. the calculation of $\mathbb{E}_{\mathbf{Q}_0}[\log\{L_\tau^{(c)}(\mathbf{Q})\}|Y=y]$, where $Y=\{Y_i|i=1,\ldots,n\}$ and where $\mathbf{Q}_0$ is a given intensity matrix. From equation (2.1) we see that

$$\mathbb{E}_{\mathbf{Q}_0}[\log\{L_\tau^{(c)}(\mathbf{Q})\}|Y=y]=\sum_{i=1}^{m}\sum_{j\neq i}\log(q_{ij})\,\mathbb{E}_{\mathbf{Q}_0}\{N_{ij}(\tau)|Y=y\}-\sum_{i=1}^{m}\sum_{j\neq i}q_{ij}\,\mathbb{E}_{\mathbf{Q}_0}\{R_i(\tau)|Y=y\}.$$

This is the continuous time log-likelihood for data with observed statistics $\mathbb{E}_{\mathbf{Q}_0}\{N_{ij}(\tau)|Y=y\}$ and $\mathbb{E}_{\mathbf{Q}_0}\{R_i(\tau)|Y=y\}$, which is maximized (as a function of $\mathbf{Q}$) by equation (2.3) (the M-step). The only non-trivial task left is hence to evaluate $\mathbb{E}_{\mathbf{Q}_0}\{N_{ij}(\tau)|Y=y\}$ and $\mathbb{E}_{\mathbf{Q}_0}\{R_i(\tau)|Y=y\}$. By the Markov property and the homogeneity of the process, it is sufficient to find

$$\tilde{M}_{ij}^{k}(t)=\mathbb{E}_{\mathbf{Q}_0}\{R_k(t)|X(t)=j,X(0)=i\}\tag{3.1}$$

and

$$\tilde{f}_{ij}^{kl}(t)=\mathbb{E}_{\mathbf{Q}_0}\{N_{kl}(t)|X(t)=j,X(0)=i\}\tag{3.2}$$

because

$$\mathbb{E}_{\mathbf{Q}_0}\{N_{ij}(\tau)|Y=y\}=\sum_{k=1}^{n-1}\tilde{f}_{y_k,y_{k+1}}^{ij}(t_{k+1}-t_k),\tag{3.3}$$

$$\mathbb{E}_{\mathbf{Q}_0}\{R_l(\tau)|Y=y\}=\sum_{k=1}^{n-1}\tilde{M}_{y_k,y_{k+1}}^{l}(t_{k+1}-t_k).\tag{3.4}$$

To calculate equation (3.1), it turns out to be convenient to study the related quantity (here and later we drop the index $\mathbf{Q}_0$ for simplicity)

$$M_{ij}^{k}(t)=\mathbb{E}[R_k(t)\,I\{X(t)=j\}|X(0)=i].$$

The following result can be found in Bladt *et al.* (2002). In formula (3.5) as well as in equation (3.8), δ_{ij} equals 1 if and only if $i=j$ and is 0 otherwise.

Theorem 2. The function M_{ij}^{k} solves the differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}M_{ij}^{k}(t)=\sum_{l}M_{il}^{k}(t)q_{lj}+\exp(t\mathbf{Q})_{ij}\delta_{jk}\tag{3.5}$$

with initial condition $M_{ij}^{k}(0)=0$.

Define $\mathbf{M}_{i\cdot}^{k}(t)=(M_{i1}^{k}(t),\ldots,M_{im}^{k}(t))$ (row vector). Then formula (3.5) may be written as

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{M}_{i\cdot}^{k}(t)=\mathbf{M}_{i\cdot}^{k}(t)\mathbf{Q}+\mathbf{A}_{i}^{k}(t),$$

where $\mathbf{A}_i^k(t)=\mathbf{e}_i'\exp(\mathbf{Q}t)\mathbf{e}_k\mathbf{e}_k'$ with $\mathbf{e}_i$ denoting the unit vector with the ith co-ordinate equal to 1 and with $\mathbf{e}_i'$ denoting its transpose. This is a system of inhomogeneous linear differential equations with initial condition $\mathbf{M}^k_{i\cdot}(0)=0$ which may efficiently be solved numerically by, for example, a fourth-order Runge–Kutta method. Alternatively, we note that the solution to the system of differential equations is given by

$$\begin{aligned}\mathbf{M}^k_{i\cdot}(t)&=\int_0^t \mathbf{A}^k_i(s)\exp\{(t-s)\mathbf{Q}\}\,\mathrm{d}s\\&=\mathbf{e}_i'\int_0^t \exp(s\mathbf{Q})(\mathbf{e}_k\mathbf{e}_k')\exp\{(t-s)\mathbf{Q}\}\,\mathrm{d}s,\end{aligned}$$

which may be evaluated numerically by making a suitable expansion of the matrix exponentials by using, for example, the uniformization method (see Neuts (1995), page 232). Specifically, choose $\lambda\geqslant\max_{i=1,\ldots,m}(-Q_{ii})$, and define

$$\mathbf{B}=\mathbf{I}+\frac{1}{\lambda}\mathbf{Q}=\frac{1}{\lambda}(\lambda\mathbf{I}+\mathbf{Q}).$$

Then $\mathbf{M}^k=\{M^k_{ij}\}_{ij\in E}$ is given by

$$\mathbf{M}^k(t)=\exp(-\lambda t)\lambda^{-1}\sum_{n=0}^{\infty}\frac{(\lambda t)^{n+1}}{(n+1)!}\sum_{l=0}^{n}\mathbf{B}^l(\mathbf{e}_k\mathbf{e}_k')\mathbf{B}^{n-l}.$$

Both methods are equally efficient in lower dimensions ($m<30$). In higher dimensions (from 30 upwards) the uniformization method is more efficient.

Now we can calculate the quantity (3.1) by

$$\tilde{M}^k_{ij}(t)=\frac{M^k_{ij}(t)}{\mathbf{e}_i'\exp(\mathbf{Q}t)\mathbf{e}_j}. \tag{3.6}$$

To calculate the quantity (3.2), the expected number of transitions from state k to state l in a time interval of length t given that the process initiates in state i and terminates in state j, we first consider

$$f^{kl}_{ij}(t)=\mathbb{E}[N_{kl}(t)\;I\{X(t)=j\}|X(0)=i] \tag{3.7}$$

for fixed k and l.

Theorem 3. The function f^{kl}_{ij} that is given by equation (3.7) solves the differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}f^{kl}_{ij}(t)=\sum_{h=1}^{m}f^{kl}_{ih}(t)q_{hj}+q_{kl}\exp(\mathbf{Q}t)_{ik}\delta_{jl}, \tag{3.8}$$

with boundary condition $f^{kl}_{ij}(0)=0$ for all i and j.

Proof. In Bladt *et al.* (2002) it is shown that

$$\begin{aligned}V^*_{ij}(\mathbf{s},Z;t)&=\mathbb{E}\left[\exp\left\{-\sum_{h=1}^{m}s_h\,R_h(t)\right\}\prod_{a,b}z_{ab}^{N_{ab}}\;I\{X(t)=j\}\middle|X(0)=i\right]\\&=\exp[\{\mathbf{Q}\bullet Z+\Delta(\mathbf{s})\mathbf{I}\}t],\end{aligned}$$

where $\mathbf{s}=(s_1,\ldots,s_m)$ and $\mathbf{Z}=\{z_{ab}\}_{a,b=1,\ldots,m}$ are variables, $\mathbf{I}$ denotes the identity matrix, $\bullet$ denotes the Schur product (which is defined as a product between two matrices $\mathbf{A}=\{a_{ij}\}$ and $\mathbf{B}=\{b_{ij}\}$ by $\mathbf{A}\bullet\mathbf{B}=\{a_{ij}b_{ij}\}$) and $\Delta(\mathbf{s})$ is the diagonal matrix with the numbers $s_1,\ldots,s_m$ as its diagonal. Setting $\mathbf{s}=\mathbf{0}$, $z_{ab}=1,(a,b)\neq(k,l)$ and $z_{kl}=z$, the result is easily obtained by differentiation with respect to z and t, evaluating the former at $z=1$. □

We may solve for $f_{ij}^{kl}(t)$ by either the Runge–Kutta method or by the uniformization method. By uniformization we obtain that

$$\mathbf{f}^{kl}(t) = q_{kl} \exp(-\lambda t) \lambda^{-1} \sum_{n=0}^{\infty} \frac{(\lambda t)^{n+1}}{(n+1)!} \sum_{j=0}^{n} \mathbf{B}^j (\mathbf{e}_k \mathbf{e}_l') \mathbf{B}^{n-j}, \tag{3.9}$$

where λ and $\mathbf{B}$ are as for $M_{ij}^k(t)$.

We can now calculate the quantity $\tilde{f}_{ij}^{kl}(s,t)$ that is defined by equation (3.2):

$$\tilde{f}_{ij}^{kl}(t) = \frac{f_{ij}^{kl}(t)}{\mathbf{e}_i' \exp\{\mathbf{Q}(t)\} \mathbf{e}_j}. \tag{3.10}$$

Summing up, the EM algorithm for maximum likelihood estimation of $\hat{\mathbf{Q}}$ is as follows.

Let $\mathbf{Q}_0$ be any intensity matrix for a Markov jump process with state space E. Initially set $\mathbf{Q} = \mathbf{Q}_0$.

Step 1: calculate $\tilde{M}_{y_i, y_{i+1}}^k (t_{i+1} - t_i)$ and $\tilde{f}_{y_i, y_{i+1}}^{kl} (t_{i+1} - t_i)$ for all k and l under the model with intensity matrix $\mathbf{Q}$ by equations (3.6) and (3.10).
Step 2: calculate $\mathbb{E}_{\mathbf{Q}}\{R_i(\tau) | Y = y\}$ and $\mathbb{E}_{\mathbf{Q}}(N_{ij} | Y = y)$ by equations (3.3) and (3.4).
Step 3: calculate $\hat{\mathbf{Q}}$ by $\hat{\mathbf{Q}}_{ij} = \mathbb{E}_{\mathbf{Q}}(N_{ij} | Y = y) / \mathbb{E}_{\mathbf{Q}}\{R_i(\tau) | Y = y\}$ for all $i \neq j$.
Step 4: $\mathbf{Q} := \hat{\mathbf{Q}}$. Go to step 1.

Let $\mathbf{Q}_0, \mathbf{Q}_1, \mathbf{Q}_2, \ldots$ be a sequence of intensity matrices obtained by the EM algorithm. Then certainly $L_n(\mathbf{Q}_{k+1}) \geqslant L_n(\mathbf{Q}_k)$ for $k = 0, 1, 2, \ldots$, where L_n is the discrete time likelihood function (2.5); see Dempster *et al.* (1977). Regularity conditions for the sequence to converge to a (possibly local) maximum of the likelihood function were given by Wu (1983); see also McLachlan and Krishnan (1997). Unfortunately, one of Wu's conditions, condition (3.19) in McLachlan and Krishnan (1997), is not satisfied by the model that is treated here. In the two-state case that was considered in example 1 it is obvious that there is a problem at the boundary where $\pi_{11} = \pi_{21}$, which does not belong to the parameter space. For general m there is a similar problem at the boundary where $\det\{\exp(\mathbf{Q})\} \rightarrow 0$. One way around this problem is to use the slightly smaller parameter space

$$\mathcal{Q}_\varepsilon = \{\mathbf{Q} \in \mathcal{Q} | \det\{\exp(\mathbf{Q})\} \geqslant \varepsilon\}$$

for some small $\varepsilon > 0$. With this restricted parameter set, it is clear that condition (3.19) in McLachlan and Krishnan (1997) is satisfied, because the discrete time likelihood function L_n is essentially a multinomial likelihood with an unusual parameter space. Let us consider the rest of the conditions (3.18)–(3.21) and (3.23) in McLachlan and Krishnan (1997), which by their theorem 3.2 would imply the convergence of the sequence $\{\mathbf{Q}_k\}$. Condition (3.18) with $d = m(m-1)$ is trivial, and condition (3.20) that the function $\mathbf{Q} \mapsto L_n(\mathbf{Q})$ is continuous and differentiable in the interior of the parameter space follows from the fact that the function $\mathbf{Q} \mapsto \exp(\mathbf{Q})$ is continuous on $\mathcal{Q}$ and differentiable on the interior of $\mathcal{Q}$, i.e. where $q_{ij} > 0$ for all $i \neq j$; see for example Neuts (1995). The continuity of the function

$$(\mathbf{Q}, \mathbf{Q}_0) \rightarrow \mathbb{E}_{\mathbf{Q}_0}[\log\{L_\tau^{(c)}(\mathbf{Q})\} | Y = y],$$

condition (3.23), is obvious from the expressions that were derived previously for $\tilde{M}_{ij}(t)$ and $\tilde{f}_{ij}^{kl}(t)$ as functions of the parameter $\mathbf{Q}_0$. Finally, condition (3.21) that $\mathbf{Q}_{k+1}$ solves

$$\partial \mathbb{E}_{\mathbf{Q}_k}[\log\{L_\tau^{(c)}(\mathbf{Q})\} | Y = y] / \partial \mathbf{Q} = 0$$

is satisfied for the full parameter space $\mathcal{Q}$, provided that the initial matrix $\mathbf{Q}_0$ is chosen in the interior of $\mathcal{Q}$. To see this, note that for any $\mathbf{Q}_0$ in the interior of $\mathcal{Q}$ the expected holding times and the expected numbers of jumps are strictly positive for all possible states. Therefore the maximum likelihood estimator that is obtained by using these expected values as the statistics in L_n has strictly positive off-diagonal elements (see equation (2.3)), and hence $\mathbf{Q}_1$ belongs to the interior of $\mathcal{Q}$. Iteration of this argument shows that $\mathbf{Q}_k$ belongs to the interior of $\mathcal{Q}$ for all k. (Note that some $(\mathbf{Q}_k)_{ij}$ may converge to 0 as $k \to \infty$.) However, for the restricted parameter space $\mathcal{Q}_\varepsilon$, it may happen that the sequence $\mathbf{Q}_k$ converges to the boundary where $\det\{\exp(\mathbf{Q})\} = \varepsilon$ and that $\det\{\exp(\mathbf{Q})_k\} = \varepsilon$ for some k. Then condition (3.21) in McLachlan and Krishnan (1997) will typically not be satisfied. In view of theorem 3.2 in McLachlan and Krishnan (1997) we can summarize the discussion as follows.

Theorem 4. Suppose that the initial matrix $\mathbf{Q}_0$ belongs to the interior of the parameter space $\mathcal{Q}$, i.e. that $(\mathbf{Q}_0)_{ij} > 0$ for all $i \neq j$. Then the sequence $\{\mathbf{Q}_k\}$ will either converge to a stationary point of the likelihood function L_n or $\det\{\exp(\mathbf{Q}_k)\} \to 0$.

If the latter possibility occurs, it is an indication that the maximum likelihood estimator does not exist. Indeed, the problems with the EM algorithm are closely related to the problems with the maximum likelihood estimator that were discussed in the previous section. Obviously, it is a good idea to choose the initial matrix $\mathbf{Q}_0$ in such a way that $\det\{\exp(\mathbf{Q}_k)\}$ is far from 0. If $\mathbf{Q}_0$ is chosen such that some $(\mathbf{Q}_0)_{ij} = 0$, then the expected number of jumps from i to j will remain 0 through all iterations, i.e. all $\mathbf{Q}_k$ will belong to the boundary of $\mathcal{Q}$, where differentiability does not make sense, and where some of the above conditions do not hold. If it is desirable to choose $\mathbf{Q}_0$ such that some $(\mathbf{Q}_0)_{ij} = 0$, a convergence result similar to theorem 4 can be obtained by reducing the parameter space by the restriction $q_{ij} = 0$.

Use of the restricted parameter space $\mathcal{Q}_\varepsilon$ is a rather crude way to solve the problem at the boundary where $\det\{\exp(\mathbf{Q})\} \to 0$ and is mainly a technical device to prove theorem 4. A softer approach would be to use a likelihood function that is penalized near the critical boundary in such a way that the penalized likelihood goes to 0 as $\det\{\exp(\mathbf{Q})\} \to 0$. The EM algorithm can also be applied to maximum penalized likelihood estimation; see McLachlan and Krishnan (1997). An obvious way to penalize the likelihood is provided by the conjugate priors that are discussed in the next section, where an MCMC method is presented as an alternative to the EM algorithm.

4. Markov chain Monte Carlo estimation

In this section we present an MCMC approach to estimating the parameters of a discretely observed Markov jump process. The setting is slightly more general than that in the previous sections because this can be useful and does not essentially complicate the MCMC approach. Consider a Markov jump process $\{J(t)\}$ with $p = p_1 + p_2 + \ldots + p_m$ states and intensity matrix $\mathbf{Q}$. A new process $\{X(t)\}$ is defined in the following way:

$$X(t) = i \Leftrightarrow J(t) \in \{p_{i-1} + 1, \ldots, p_i\}, \qquad i = 1, 2, \ldots, m,$$

where $p_0 = 0$. Thus we have grouped the states of J, and X indicates which group the process J is in at any given time. The process $\{X(t)\}$ is in general not a time homogeneous Markov process, since the sojourn times in states $1, 2, \ldots, m$ are not necessarily exponentially distributed.

In this section we consider discrete time observations of X, and the purpose is to estimate the intensity matrix $\mathbf{Q}$ of the Markov jump process J underlying the non-Markovian process X to the extent that this is possible. If $p_i = 1$ for all $i \in E$, then we may estimate the parameters of $\mathbf{Q}$ whenever it is uniquely determined by the distribution of the discrete time process (see Section 2 for some necessary conditions). If some $p_i > 1$, then $\mathbf{Q}$ is no longer unique, and it is not possible to estimate all parameters of $\mathbf{Q}$ by MCMC sampling, which will be apparent from the following discussion. It will, however, be possible to estimate functions of $\mathbf{Q}$ that are invariant under the different representations of the distribution of the observed discrete time process. An example is the (time-dependent) rates of transitions between the different states $1, 2, \ldots, m$ of the process X.

Consider the discrete time observations $\mathbf{x} = (x_1, \ldots, x_n)$ of the continuous time jump process $\{X(t)\}_{t \geqslant 0}$ observed at times $t_1, \ldots, t_n$ up to time τ ($t_1 = 0$ and $t_n = \tau$). We choose a prior $\phi(\mathbf{Q})$ and are interested in the conditional distribution of $\mathbf{Q}$ given the data $\mathbf{x}$. We shall, however, study the slightly more general problem of finding the conditional distribution of $(\mathbf{Q}, \mathbf{J})$ given $\mathbf{x}$, where $\mathbf{J} = \{J(t)\}_{0 \leqslant t \leqslant \tau}$ denotes the continuous time sample path of J. For this we employ the Gibbs sampler with two sites $\mathbf{Q}$ and $\mathbf{J}$ and sample by alternately drawing $\mathbf{J}$ given $(\mathbf{Q}, \mathbf{x})$ and $\mathbf{Q}$ given $(\mathbf{J}, \mathbf{x})$($\mathbf{x}$ is of course of no importance when conditioning on $\mathbf{J}$). Iteration of the Gibbs sampler results in a sequence of variables $(\mathbf{Q}_n, \mathbf{J}_n)$. Under suitable conditions the Gibbs sampler will eventually produce a stationary and ergodic sequence, i.e., after discarding a certain burn-in period, say the first $K - 1$ iterations, the sequence $(\mathbf{Q}_n, \mathbf{J}_n)_{n \geqslant K}$ may be considered stationary, and the stationary distribution is exactly that of $(\mathbf{Q}, \mathbf{J})$ given $\mathbf{x}$.

If $p_i = 1$ for all i, then by ergodicity the empirical average

$$\frac{1}{N} \sum_{i=K}^{N+K} \mathbf{Q}_i$$

converges to the true mean of $\mathbf{Q}$ conditionally on $\mathbf{x}$. Also credibility intervals based on the empirical distribution of $(\mathbf{Q}_n, \mathbf{J}_n)_{n \geqslant K}$ may be constructed, and quantiles of the empirical distribution may be of interest as well.

In situations where $\mathbf{Q}$ is not uniquely determined by the distribution of the discrete time sample, the mean of the posterior distribution may not be a meaningful quantity, and credibility intervals hardly make sense for parameters that are not uniquely determined. The same is true if $p_i > 1$ for some i. However, a function of $\mathbf{Q}$ that depends only on the distribution of the discrete time sample can in both cases be estimated by averaging the simulated values of the function. As discussed in Section 2, the set of $\mathbf{Q}$s for which this problem occurs when $p_i = 1$ for all i is complicated, so it is important to study the posterior distribution carefully for indications that the problem has occurred, for instance by inspecting scatterplots. It might seem desirable to use a prior that is concentrated on the set of $\mathbf{Q}$s for which $\exp(\mathbf{Q}) \in \mathcal{P}_{00}$ but, since this set is a very complicated set, this idea would be very difficult to implement. An easier, but less satisfactory, solution is a prior concentrated on the set of $\mathbf{Q}$s for which $\exp(\mathbf{Q})$ satisfies inequality (2.12).

A proper choice of prior is usually essential to ensure good mixing properties and a posterior which is not dominated by the prior. Sometimes hyperparameters may have to be specified to ensure satisfactory mixing; experience shows, however, that this is not necessary in the present case. We choose the prior

$$\phi(\mathbf{Q}) \propto \prod_{i=1}^{n} \prod_{j \neq i} q_{ij}^{\alpha_{ij}-1} \exp(-q_{ij}\beta_i), \tag{4.1}$$

where $\alpha_{ij} > 0$, $i, j \in E$, and $\beta_i > 0$, $i \in E$, are known constants to be chosen conveniently (set to be all 1 in the following examples). Then $q_{ij} \sim \Gamma(1/\beta_i, \alpha_{ij})$. In this way parameters that are near the critical boundary are effectively penalized because there at least one of the q_{ij}s must go to ∞ (at least one eigenvalue goes to ∞).

This family of priors is conjugate for the model for continuous observation in the time interval $[0, \tau]$, which is an exponential family of processes; see Küchler and Sørensen (1997). Indeed, the posterior is

$$p^*(\mathbf{Q}) = L_\tau^{(c)}(\mathbf{Q})\,\phi(\mathbf{Q})$$
$$\propto \prod_{i=1}^{n} \prod_{j\neq i} q_{ij}^{N_{ij}(\tau)+\alpha_{ij}-1} \exp[-q_{ij}\{R_i(\tau)+\beta_i\}],$$

where the likelihood function $L_\tau^{(c)}(\mathbf{Q})$ is given by equation (2.1). Drawing a $\mathbf{J}$ given $(\mathbf{Q},\mathbf{x})$ is performed by simulating Markov jump processes step by step through the intervals $[t_k, t_{k+1}]$ initiating from some initial condition $X(t_k)=i$ such that the process will be in a state (or more generally in a group of substates) $X(t_{k+1})=j$, say, by time t_{k+1}. This can be done by simple rejection sampling, which has turned out to be quite efficient even in higher dimensions; see the next section. There may, however, be situations where more sophisticated methods such as importance sampling or Metropolis–Hastings algorithms may have to be applied if some transition from $X(t_k)$ to $X(t_{k+1})$ has a low probability.

5. Implementation and examples

From a computational point of view the crucial step in the EM algorithm is the E-step. Since the E-step is essentially given as a solution to the system of differential equations (3.5) and (3.8) of dimension m, the time complexity is necessarily exponentially increasing as a function of the number of states. If the sampling frequency is constant, i.e. $t_{k+1}-t_k=\Delta$ for all k, the computational burden reduces significantly. In this case the EM algorithm is essentially insensitive to the number of data points.

When solving for the E-step, we have applied the fourth-order Runge–Kutta method, which is a relatively fast and very reliable method. The step size specification is a crucial parameter for the Runge–Kutta method which should be chosen adequately depending on the increment $t_{k+1}-t_k$. For $t_{k+1}-t_k=1$ a step size of 0.2 or less was found to perform well. The execution time of the E-step is inversely proportional to the step size. One way of obtaining both speed and precision is to let the step size vary from coarser at the first iterations to finer at the last iterations. Convergence of the EM algorithm may on a contemporary personal computer be obtained in a central processor unit (CPU) time ranging from a few seconds for two states to about 45 min for 10 states (90 free variables).

The only time-consuming part of the MCMC algorithm is the (rejection) sampling of the trajectories through the observed data points. Even if rejection could be avoided, the execution time increases proportionally to the number of observed data points. There is also a linear increase in the execution time as the number of states increases. The MCMC algorithm runs effectively also for a large number of states (10 and upwards) with relatively short burn-in periods. For uniquely imbeddable data it seems that the MCMC algorithm is insensitive to the choice of the parameters in the gamma prior.

For either approach, the model state space must coincide with the observed states and states that can be inferred from the data. If some states are not observed in the data, a reduction of the state space must take place if it cannot be deduced from the data and structure of the intensity matrix that such unobserved states have been visited by the Markov jump process between the observation time points. For a birth-and-death process it can for instance be inferred that an unobserved state 4 has been visited if the data contain the observations 3 and 5.

In example 1 we saw what typically happens when the matrix of empirical transition probabilities (2.8) is not imbeddable. The following examples, most of them with simulated data, illustrate how the proposed estimation methods work in various situations.

5.1. Example 2

Consider the intensity matrix

$$\boldsymbol{\Lambda}=\begin{pmatrix} -10.81910 & 4.174788 & 0.6537399 & 2.405759 & 3.584813 \\ 4.946455 & -11.30472 & 0.0792565 & 2.065462 & 4.213545 \\ 2.059437 & 2.949915 & -9.921236 & 4.214654 & 0.6972306 \\ 3.353977 & 0.2849829 & 4.707894 & -8.927676 & 0.5808223 \\ 1.880896 & 2.227394 & 0.8062197 & 0.1347199 & -5.049230 \end{pmatrix}.$$

Suppose that we observe a Markov jump process with intensity matrix $\boldsymbol{\Lambda}$ at times $t=1,2,\ldots$. Then the transition matrix of the observed Markov chain is

$$\mathbf{P}=\begin{pmatrix} 0.2112767 & 0.1784938 & 0.1178301 & 0.1588798 & 0.3335184 \\ 0.2112521 & 0.1785134 & 0.1177149 & 0.1587348 & 0.3337837 \\ 0.2114293 & 0.1783682 & 0.1185720 & 0.1598076 & 0.3318222 \\ 0.2114681 & 0.1783316 & 0.1187727 & 0.1600574 & 0.3313693 \\ 0.2111336 & 0.1785908 & 0.1172314 & 0.1581089 & 0.3349343 \end{pmatrix}.$$

All rows of **P** are approximately equal to the row vector $\boldsymbol{\pi}$ where the co-ordinates equal the probabilities of the stationary distribution. This occurs in all cases where the observed Markov jump process has sufficiently high intensities compared with the sampling frequency that the process settles almost into stationary mode between two consecutive sampling times. One problem is that any stationary Markov jump process with stationary distribution $\boldsymbol{\pi}$ has a transition matrix that is close to **P** above, and hence with a large probability the solution to the imbedding problem given by the matrix of empirical transition probabilities (2.8) is not unique. Even worse, it is quite possible that the maximum likelihood estimator does not exist (e.g. because two rows of matrix (2.8) are identical so that the determinant is zero); see the discussion after theorem 1.

If we had instead sampled at the time points $t=0.2,0.4,\ldots$, we would have been in a situation of unique imbedding. In practice it may be difficult to judge whether a solution that is produced by either the EM or the MCMC algorithm is unique. An indication can be obtained by calculating the matrix of empirical transition probabilities **P**. If the rows of **P** are close to being identical, the solution is most probably not unique. For the MCMC algorithm non-uniqueness will result in a non-stationary sequence of intensity parameters (but of course drawn from the same distribution).

5.2. Example 3

Here we consider the computationally and statistically demanding problem of estimating the 90 free parameters in a 10-state Markov jump process from a series of 5000 simulated observations sampled at intervals of 0.5.

The MCMC estimation was based on 10000 iterations. Burn-in occurred for all parameters in less than 200 iterations. This was concluded by inspection of graphs of the time series of simulated parameters and their autocorrelation functions, which were found to decrease quickly. We used the last 9000 iterations for estimation. In the gamma prior all parameters were equal to 1. The EM algorithm converged to a precision of six decimal places in 620 iterations using 2885 s of CPU time. The 10000 iterations of the MCMC algorithm took 24434 s of CPU time. If we had used only 1000 simulated parameter values after the burn-in for our MCMC estimates instead of 9000, the CPU time would have been 4887 s, and the estimates would largely have been unchanged. In Table 1 the estimates of the 10 diagonal intensities are given, including the continuous time maximum likelihood estimates that the EM and MCMC estimates attempt to reconstruct from the incomplete discrete time data. It seems that the MCMC algorithm tends

Table 1. Estimates of the total jump rates of the 10 states†

Method	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}
TRUE	2.418	2.090	1.973	2.011	2.020	3.120	2.085	2.189	1.950	1.884
CONT	2.666	2.049	1.862	1.990	2.046	3.162	2.139	2.215	1.866	1.927
EM	2.662	2.065	1.869	1.893	2.000	3.391	2.190	2.137	1.877	1.920
MCMC	2.801	2.517	1.938	1.977	2.089	3.639	2.290	2.248	1.950	2.010
2.5%	2.421	1.879	1.717	1.748	1.853	3.112	2.013	1.968	1.734	1.763
97.5%	3.211	2.517	2.165	2.219	2.344	4.277	2.596	2.542	1.950	2.263

†TRUE, the parameter values that were used in the simulation of the data; CONT, maximum likelihood estimates based on continuous time data; EM, estimates when applying the EM algorithm to the discrete time data; MCMC, estimates when applying the MCMC method to discrete time data; the last two rows are the 2.5% and 97.5% percentiles of the data from the MCMC method.

Table 2. Comparison of the continuous time, EM and MCMC estimates of rates out of state 1 and the 95% credibility bounds of the MCMC method

Method	λ_{12}	λ_{13}	λ_{14}	λ_{15}	λ_{16}	λ_{17}	λ_{18}	λ_{19}	$\lambda_{1,10}$
TRUE	0.417	0.065	0.241	0.358	0.495	0.008	0.207	0.421	0.206
CONT	0.436	0.114	0.268	0.337	0.704	0.000	0.193	0.416	0.198
EM	0.470	0.117	0.237	0.443	0.585	0.000	0.077	0.530	0.203
MCMC	0.456	0.133	0.233	0.452	0.627	0.060	0.095	0.547	0.207
2.5%	0.277	0.010	0.068	0.246	0.342	0.002	0.005	0.312	0.051
97.5%	0.664	0.316	0.414	0.694	0.971	0.186	0.256	0.810	0.396

to overestimate these rates of transitions. This is probably due to our choice of the prior which does not allow for zero or very small rates. Table 2 shows the estimates of the intensities of the transitions out of state 1. Also here both methods work well.

MCMC estimation also worked well for a 20-state Markov jump process (380 parameters). Here each iteration took about 5 CPU s, which—as expected—is twice the CPU time that is used per iteration for the 10-state process.

5.3. Example 3

Considerable improvement in speed and precision may be obtained for a submodel of the full Markov jump process model with fewer parameters. An example is a birth-and-death process, for which the number of free parameters increases only linearly with the number of states as opposed to the quadratic increase for the full model. In the E-step of the EM algorithm we still must solve the system of differential equations, which from a numerical point of view simplifies only slightly by the parameter reduction, so the main improvement in speed and precision is due to the actual parameter reduction itself. To give a concrete example, 1000 observations sampled at equidistant times $t = 1, 2, \ldots$ were generated from two five-state Markov jump processes: one of the general type and one of the birth-and-death type. The full model has 20 free parameters, whereas the birth-and-death process has only 10. Convergence (to six decimal places) of the EM algorithm was obtained in 40 iterations for the birth-and-death process, whereas it took 2210 iterations for the full model. Whereas for the birth-and-death process there were 5.7 iterations per second of CPU time, there were 5.1 iterations per second of CPU time for the full model. Similar considerations apply to the MCMC approach.

Table 3. MCMC estimates with the 95% credibility bounds for certain parameters of the 50-state birth-and-death process

Method	$\lambda_{3,4}$	$\lambda_{4,3}$	$\lambda_{8,9}$	$\lambda_{9,8}$	$\lambda_{13,14}$	$\lambda_{14,13}$	$\lambda_{18,19}$	$\lambda_{19,18}$	$\lambda_{23,24}$	$\lambda_{24,23}$
True value	0.989	0.016	0.942	0.116	0.372	0.510	0.227	0.865	0.615	0.927
MCMC	1.891	0.199	1.951	0.144	0.463	0.659	0.247	0.899	0.749	1.110
2.5%	0.229	0.004	0.236	0.003	0.346	0.495	0.153	0.564	0.481	0.706
97.5%	5.379	0.813	5.315	0.595	0.605	0.857	0.376	1.351	1.180	1.726
	$\lambda_{28,29}$	$\lambda_{29,28}$	$\lambda_{33,34}$	$\lambda_{34,33}$	$\lambda_{38,39}$	$\lambda_{39,38}$	$\lambda_{43,44}$	$\lambda_{44,43}$	$\lambda_{48,49}$	$\lambda_{49,48}$
True value	0.625	0.451	0.378	0.661	0.824	0.339	0.359	0.513	0.960	0.567
MCMC	0.916	1.021	0.624	0.826	0.824	0.353	0.341	0.491	1.032	0.672
2.5%	0.165	0.103	0.280	0.349	0.657	0.278	0.213	0.305	0.798	0.519
97.5%	2.618	3.250	1.172	1.630	1.017	0.439	0.507	0.737	1.324	0.870

In view of these considerations, we can test whether the rejection sampling that we use to simulate a continuous time trajectory conditionally on the discrete time data causes problems for MCMC estimation when the number of states is very large by considering estimation for a birth-and-death process. A simulated birth-and-death process with 50 states was observed at 5000 equidistant time points $t = 1, 2, \ldots$. The execution time to run 10 000 MCMC iterations was 41 032 CPU s or 4.1 CPU s per iteration, so the rejection sampling works even for this extreme number of states. The burn-in time appeared to be less than 20 iterations and the autocorrelation function decreases very quickly to 0, but we discarded the first 1000 iterations. Estimates of some of the parameters that are the average of 9000 simulated parameter values are given in Table 3. In some cases the 95% credibility bounds are rather wide because the corresponding transitions were relatively rarely observed.

6. Concluding remarks

We have demonstrated that maximum likelihood estimation of the intensity matrix of a Markov jump process with finite state space is practically feasible by means of the EM algorithm or an MCMC procedure. When one or more of the intensities are large, the maximum likelihood estimator may not exist. Essentially the problem of non-existence occurs when the process moves too quickly compared with the sampling frequency, which implies that much happens between the sampling times that we do not obtain information about. Therefore non-existence of the maximum likelihood estimator should perhaps be taken as a sign that there is not enough information in the data to estimate the intensity matrix properly. If the process is such that it moves quickly between the states within one or more groups, but more slowly between the groups and other states, it might be a good idea to join each of the groups into a new single state, and then to estimate only the transition intensities between the states in this new process with reduced state space. In this way the information in the data is used to estimate the parameters about which the data actually contain information. It is, of course, not possible that both the original process and the new process are Markovian, so the results that are obtained by means of the new process must be interpreted with care.

As we have seen, another way around the non-existence problem is to use a penalized likelihood function or the MCMC estimator with a suitable prior. Then an estimator will always be obtained, but it is likely that, at least in extreme cases, the estimator will depend heavily on the prior. A more serious problem is that the MCMC approach may hide problems of nonexistence or non-uniqueness of the maximum likelihood estimator. In the first case, it might not be noticed that the data contain very little information on certain parameters or that the model is perhaps not appropriate. In the second case, nonsensical results may be obtained. Again care is required.

Acknowledgements

We are grateful to the referees and the Joint Editor for their constructive comments. Our collaboration was partly financed by the Centre for Mathematical Physics and Stochastics funded by the Danish National Research Foundation. The research of Michael Sørensen was also supported by the European Community's 'Human potential programme' under contract HPRN-CT-2000-00100, 'Statistical methods for dynamical stochastic models', and by the Centre for Analytical Finance and the Danish Mathematical Finance Network, both financed by the Danish Social Science Research Council.

References

Aït-Sahalia, Y. (2002) Maximum likelihood estimation of discretely sampled diffusions: a closed-form approximation approach. *Econometrica*, **70**, 223–262.

Bibby, B. M., Jacobsen, M. and Sørensen, M. (2004) Estimating functions for discretely sampled diffusion-type models. In *Handbook of Financial Econometrics* (eds Y. Aït-Sahalia and L. P. Hansen). Amsterdam: North-Holland. To be published.

Billingsley, P. (1961) *Statistical Inference for Markov Processes*. Chicago: University of Chicago Press.

Bladt, M., Neuts, M. F., Meini, B. and Sericola, B. (2002) Distributions of reward functions on continuous-time, Markov chains. In *Matrix-analytic Methods: Theory and Applications* (ed. G. Latouche). Singapore: World Scientific Publishing Company.

Culver, W. J. (1966) On the existence and uniqueness of the real logarithm of a matrix. *Proc. Am. Math. Soc.*, **17**, 1146–1151.

Cuthbert, J. R. (1972) On uniqueness of the logarithm for Markov semi-groups. *J. Lond. Math. Soc.*, **4**, 623–630.

Cuthbert, J. R. (1973) The logarithm function for finite-state Markov semi-groups. *J. Lond. Math. Soc.*, **6**, 524–532.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc.* B, **39**, 1–38.

Elerian, O., Chib, S. and Shepard, N. (2001) Likelihood inference for discretely observed non-linear diffusions. *Econometrica*, **69**, 959–993.

Elfving, G. (1937) Zur Theorie der Markoffschen Ketten. *Acta Soc. Sci. Finn.* A, **2**, 1–17.

Hoffmann, M. (1999) L_p-estimation of the diffusion coefficient. *Bernoulli*, **5**, 447–481.

Israel, R. B., Rosenthal, J. S. and Wei, J. Z. (1997) Finding generators for Markov chains via empirical transition matrices, with applications to credit ratings. *Rev. Finan. Stud.*, **10**, 481–523.

Jacobsen, M. (1982) Statistical analysis of counting processes. *Lect. Notes Statist.*, **12**.

Jarrow, R. A., Lando, D. and Turnbull, S. M. (1997) A Markov model for the term structure of credit risk spreads. *Rev. Finan. Stud.*, **10**, 481–523.

Johansen, S. (1974) Some results on the imbedding problem for finite Markov chains. *J. Lond. Math. Soc.*, **8**, 345–351.

Keiding, N. (1974) Estimation in the birth process. *Biometrika*, **61**, 71–80.

Keiding, N. (1975) Maximum likelihood estimation in the birth-and-death process. *Ann. Statist.*, **3**, 363–372.

Kessler, M. and Sørensen, M. (1999) Estimating equations based on eigenfunctions for a discretely observed diffusion process. *Bernoulli*, **5**, 299–314.

Kingman, J. F. C. (1962) The imbedding problem for finite Markov chains. *Z. Wahrsch.*, **1**, 14–24.

Küchler, U. and Sørensen, M. (1997) *Exponential Families of Stochastic Processes*. New York: Springer.

McLachlan, G. J. and Krishnan, T. (1997) *The EM Algorithm and Extensions*. New York: Wiley.

Neuts, M. F. (1995) *Algorithmic Probability: a Collection of Problems*. London: Chapman and Hall.

Oakes, D. (1999) Direct calculation of the information matrix via the EM algorithm. *J. R. Statist. Soc.* B, **61**, 479–482.

Roberts, G. O. and Stramer, O. (2001) On inference for partially observed nonlinear diffusion models using Metropolis-Hastings algorithms. *Biometrika*, **88**, 603–621.

Wu, C. J. F. (1983) On the convergence properties of the EM algorithm. *Ann. Statist.*, **11**, 95–103.