Statistical inference for discretely observed Markov jump processes.

Mogens Bladt
IIMAS–UNAM
A.P. 20-726
01000 Mexico, D.F.
Mexico
bladt@sigma.iimas.unam.mx

Michael Sørensen
Department of Applied Mathematics and Statistics
University of Copenhagen
Universitetsparken 5
DK-2100 Copenhagen Ø
Denmark
michael@math.ku.dk

Abstract

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 MCMC-procedure with a suitable prior. The theory is illustrated by a simulation study.

Key words: EM-algorithm, imbedding problem, likelihood inference, Markov chain Monte Carlo.

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 e.g. Billingsley (1961), Jacobsen (1982), and Küchler & Sørensen (1997). If a Markov jump process is only observed 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 & Sørensen (1999), Hoffmann (1999), Aït-Sahalia (2002), Elerian, Chib & Shepard (2001), and Bibby, Jacobsen & Sørensen (2003). 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) and Keiding (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, Lando & Turnbull (1997). This lead Israel, Rosenthal & Wei (1997) to propose a method of estimating the jump intensities from discrete time observations. Their method is, however, not efficient and does only after an ad hoc modification of the estimator yield an intensity matrix.

In this paper we discuss the problems 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 procedure. It is possible that the maximum likelihood estimator 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 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 Markov chain Monte Carlo procedure presented in Section 4. In fact, a Gibbs sampler with a conjugate prior turns out to be sufficient to solve the problem. A numerical study in Section 5 indicate that when the maximum likelihood estimator exists, the two methods do an equally good job.

2 The likelihood function

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

Küchler & Sørensen (1997)). The likelihood function is given by

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

The process $N_{ij}(t)$ is the number of transitions from state i to state j in the time interval $[0, t]$, while

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

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

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

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 $\boldsymbol{Q}$ 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 $P^{\Delta_i}(\boldsymbol{Q})$, where $\Delta_i = t_{i+1} - t_i$ and

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

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

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

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 zero. 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 $P^\Delta(\boldsymbol{Q})$, so the likelihood function simplifies somewhat:

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

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

$$L(\boldsymbol{P}) = \prod_{i=1}^{m} \prod_{j=1}^{m} \boldsymbol{P}_{ij}^{K_{ij}(n)}, \quad \boldsymbol{P} \in \mathcal{P} \tag{2.7}$$

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 one. This likelihood function is identical to the one for m independent multinomial distributions, so the maximum likelihood estimator of the parameter $\boldsymbol{P}$ is

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

where

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

Define

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

the set of transition matrices that correspond to discrete time observation of a continuous time Markov jump process. Now suppose we calculate $\hat{\boldsymbol{P}}$ by (2.8) based on our discrete time observations of a continuous time Markov jump process. If $\hat{\boldsymbol{P}} \in \mathcal{P}_0$, there exists a $\hat{\boldsymbol{Q}} \in \mathcal{Q}$ such that $P^{\Delta}(\hat{\boldsymbol{Q}}) = \hat{\boldsymbol{P}}$, and the likelihood function (2.6) attains its maximal value at $\hat{\boldsymbol{Q}}$, 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 $\hat{\boldsymbol{Q}}$ needs not be unique. When $\hat{\boldsymbol{P}} \notin \mathcal{P}_0$ the situation is not clear due to the complicated structure of $\mathcal{P}_0$, but it seems not to be an uncommon occurrence 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 2.1 below. In particular, the probability that $\hat{\boldsymbol{P}} \in \mathcal{P}_0$ goes to one as n tends to infinity. We shall give a complete discussion of the case $m = 2$, where the maximum likelihood estimator does not exist when $\hat{\boldsymbol{P}} \notin \mathcal{P}_0$.

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

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

and derived the following general results about $\mathcal{P}_0$. For $m \geq 3$, $\mathcal{P}_0$ 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 involved.

The second problem is whether there are two or more intensity matrices, $\boldsymbol{Q}$, for which the corresponding transition matrix, $\exp(\Delta\boldsymbol{Q})$, 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 whether the parametrization of the distribution of the data $X(\Delta), \ldots, X(n\Delta)$ by $\boldsymbol{Q}$ is identifiable. Let $\mathcal{P}_{00}$ denote the subset of $\mathcal{P}_0$ of transition matrices $\boldsymbol{P} \in \mathcal{P}_0$, for which $\boldsymbol{Q}$ is uniquely determined by $\boldsymbol{P} = \exp(\boldsymbol{Q})$. 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 $\boldsymbol{P} \in \mathcal{P}_0$, for which all eigenvalues of $\boldsymbol{P}$ are positive and no elementary divisor (Jordan block) of $\boldsymbol{P}$ belonging to any eigenvalue appears more than once. Thus once $\hat{\boldsymbol{P}}$ has been calculated from (2.8), it is in principle easy to check whether it determines an estimator of the intensity matrix uniquely (provided that $\hat{\boldsymbol{P}} \in \mathcal{P}_0$). If $\boldsymbol{P} \notin \mathcal{P}_{00}$, there are infinitely many solutions $\boldsymbol{X}$ to the equation $\boldsymbol{P} = \exp(\boldsymbol{X})$, not all of which belong to $\mathcal{P}_0$. The set of solutions is countable if all real eigenvalues of $\boldsymbol{P}$ are positive with their Jordan blocks appearing only once and any complex eigenvalue belongs to only one Jordan block. Otherwise the are uncountably many solutions. Cuthbert (1973) showed that in the countable case only a finite subset of the solutions are in $\mathcal{P}_0$.

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

$$\inf_i \boldsymbol{P}_{ii} \geq \tfrac{1}{2}. \tag{2.11}$$

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

$$\left(\inf_i \boldsymbol{P}_{ii}\right) \cdot \det(\boldsymbol{P}) > e^{-\pi} \prod_i \boldsymbol{P}_{ii}, \tag{2.12}$$

see Cuthbert (1973) ($e^{-\pi} \simeq 0.0432$).

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

Theorem 2.1 *If $\hat{\boldsymbol{P}}$ given by (2.8) belongs to $\mathcal{P}_0$, then the maximum likelihood estimator of the intensity matrix $\hat{\boldsymbol{Q}}$ exists and is the solution to $\hat{\boldsymbol{P}} = \exp(\Delta\hat{\boldsymbol{Q}})$. If $\hat{\boldsymbol{P}} \notin \mathcal{P}_0$, then either the maximum likelihood estimator $\hat{\boldsymbol{Q}}$ exists and satisfies that $\exp(\Delta\hat{\boldsymbol{Q}}) \in \delta\mathcal{P}_0$ (given by (2.10)), or the likelihood function (2.6) has no maximum in $\mathcal{Q}$. If the true transition matrix $\boldsymbol{Q}_0$ satisfies that $\exp(\Delta\boldsymbol{Q}_0) \in \operatorname{int}\mathcal{P}_0$, and if and the Markov process is ergodic, then the probability that the maximum likelihood estimator exists goes to one as $n \to \infty$, and $\exp(\Delta\hat{\boldsymbol{Q}}) \to \exp(\Delta\boldsymbol{Q}_0)$ almost surely. Moreover, if $\boldsymbol{Q}_0$ satisfies that $\exp(\Delta\boldsymbol{Q}_0) \in \operatorname{int}\mathcal{P}_{00}$, then the probability that the maximum likelihood estimator is unique goes to one and $\hat{\boldsymbol{Q}} \to \boldsymbol{Q}_0$ almost surely as $n \to \infty$. The condition $\exp(\Delta\boldsymbol{Q}_0) \in \operatorname{int}\mathcal{P}_{00}$ is satisfied when Δ is sufficiently small.*

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

$$\mathcal{P}_c = \{\boldsymbol{P} \in \mathcal{P} \mid \log L(\boldsymbol{P}) \geq -c\},$$

where $L(\boldsymbol{P})$ is the likelihood function for the full class of Markov chains given by (2.7) and $c > 0$. Consider the compact set $\mathcal{P}_c \cap \bar{\mathcal{P}}_0$ for a $c > 0$ sufficiently large that $\mathcal{P}_c \cap \bar{\mathcal{P}}_0$ is not empty. Here $\bar{\mathcal{P}}_0$ denotes the set $\bar{\mathcal{P}}_0 = \mathcal{P}_0 \cup \{\boldsymbol{P} \in \mathcal{P} \mid \det(\boldsymbol{P}) = 0\}$. The continuous function

$L(\boldsymbol{P})$ has a maximum $\tilde{\boldsymbol{P}}$ in $\mathcal{P}_c \cap \bar{\mathcal{P}}_0$, and since $L(\boldsymbol{P})$ increases whenever $\boldsymbol{P}$ is moved in the direction of $\hat{\boldsymbol{P}}$, $\tilde{\boldsymbol{P}}$ is on the boundary of $\mathcal{P}_c \cap \bar{\mathcal{P}}_0$. Thus either $\tilde{\boldsymbol{P}} \in \delta\mathcal{P}_0$, in which case there exists a $\boldsymbol{Q}$ such that $\exp(\Delta\hat{\boldsymbol{Q}}) = \tilde{\boldsymbol{P}}$ (remember that $\mathcal{P}_0$ is closed relative to $\mathcal{P}_+$), or $\det(\tilde{\boldsymbol{P}}) = 0$, in which case the likelihood function does not have a maximum in $\mathcal{Q}$.

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

The situation that $\det(\tilde{\boldsymbol{P}}) = 0$, where the maximum likelihood estimator does not exist, is more likely to happen when the determinant of $\exp(\Delta\boldsymbol{Q}_0)$ is close to zero. When the Markov process is ergodic, $\exp(\Delta\boldsymbol{Q}_0)$ converges as $\Delta \to \infty$ to the singular matrix, where all rows are equal to the row vector $\boldsymbol{\pi}$ given by $\boldsymbol{\pi}\boldsymbol{Q}_0 = 0$ (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 the only thing we can say for sure about uniqueness is that the maximum likelihood estimator is unique when $\hat{\boldsymbol{P}} \in \mathcal{P}_{00}$ and that the maximum likelihood estimator is not unique when $\hat{\boldsymbol{P}} \in \mathcal{P}_0 \backslash \mathcal{P}_{00}$. If $\hat{\boldsymbol{P}} \notin \mathcal{P}_0$, we cannot be sure that the maximum likelihood estimator is unique, even when $\tilde{\boldsymbol{P}} \in \mathcal{P}_0$, because of the complicated geometric structure of the set $\mathcal{P}_0$.

Example 2.2 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 $\det(\boldsymbol{P}) = 0$, so the two state example is instructive.

For an intensity matrix

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

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

$$P^{\Delta}(\boldsymbol{Q}) = \frac{1}{\alpha+\beta} \begin{pmatrix} \beta + \alpha e^{-\Delta(\alpha+\beta)} & \alpha(1 - e^{-\Delta(\alpha+\beta)}) \\ \beta(1 - e^{-\Delta(\alpha+\beta)}) & \alpha + \beta e^{-\Delta(\alpha+\beta)} \end{pmatrix}$$

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

$$\pi_{11} = P^{\Delta}(\boldsymbol{Q})_{11} = 1 - (1-\rho)\alpha/(\alpha+\beta) \quad \text{and} \quad \pi_{21} = P^{\Delta}(\boldsymbol{Q})_{21} = (1-\rho)\beta/(\alpha+\beta).$$

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

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

Note that Π_0 is a parametrization of $\mathcal{P}_0$, while $\mathcal{P} = [0,1]^2$. The determinant of $P^{\Delta}(\boldsymbol{Q})$ equals $\pi_{11} - \pi_{21}$, so the diagonal $\pi_{11} = \pi_{21}$ corresponds to the problematic boundary of $\mathcal{P}_0$, where $\det(\boldsymbol{P}) = 0$. 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 $\hat{\boldsymbol{P}} \in \mathcal{P}_0$, 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 $0 \leq \pi_{21} < \hat{\pi}_{11}$, 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 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 get around the problem by penalizing the likelihood with a prior, for instance

$$\phi(\alpha, \beta) \propto \alpha^a e^{-b\alpha} \beta^c e^{-d\beta},$$

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 $K_{21}(n)/K_{2\cdot}(n) \geq K_{11}(n)/K_{1\cdot}(n)$, i.e. when $\hat{\boldsymbol{P}} \notin \mathcal{P}_0$. This estimator is not explicit, but must be found numerically. □

The eigenvalues of $\exp(\Delta \boldsymbol{Q})$ are $e^{\Delta\lambda_i}$, $i = 1, \ldots, m$, where $\{\lambda_i\}$ are the eigenvalues of $\boldsymbol{Q}$. Therefore, as $\exp(\Delta \boldsymbol{Q})$ goes to the critical boundary, where $\det(\exp(\Delta \boldsymbol{Q})) \to 0$, one or more of the eigenvalues of $\boldsymbol{Q}$ must go to minus infinity (Δ is fixed). Therefore the idea presented in Example 2.2 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 $\exp(\Delta \boldsymbol{Q}_0) \in \operatorname{int} \mathcal{P}_{00}$, that $(\boldsymbol{Q}_0)_{ij} > 0$ for $i \neq j$, and that the process is ergodic. As earlier $\boldsymbol{Q}_0$ 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 discussed in the following section, the Fisher information matrix can be calculated by means of a formula given by Oakes (1999). If $(\boldsymbol{Q}_0)_{ij} = 0$ 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 EM algorithm

The EM algorithm is a broadly applicable method for optimizing the likelihood function in cases where only partial information is available. The discretely observed Markov jump process is such an example, where maximum likelihood estimation would be an easy task if complete data $X = \{X(t) | 0 \leq t \leq \tau\}$ where observed, but where only data $Y_i = X(t_i), i = 1, ..., n$ are available. Here $t_1 = 0$ and $t_n = \tau$. If $Y = \{Y_i | i = 1, ..., n\}$, then $Y = u(X)$ for a many–to–one map, and the EM–algorithm estimates the intensity matrix $\boldsymbol{Q}$ essentially by

iterating the following two steps: (E–step) replacing the unobserved parts by their respective conditional expected values given data $Y = y$ and (M–step) performing maximum likelihood on the complete data. To be more precise, let $\boldsymbol{Q}_0$ denote any intensity matrix (initial value). Then the EM algorithm works as follows.

(1) (E–step) Calculate the function

$$g : \boldsymbol{Q} \to \mathbb{E}_{\boldsymbol{Q}_0}(\log L_\tau^{(c)}(\boldsymbol{Q})|Y = y) \tag{3.1}$$

(2) (M–step) $\boldsymbol{Q}_0 = \text{argmax}_{\boldsymbol{Q}} g(\boldsymbol{Q})$.

(3) GO TO (1).

From (2.1) we see that

$$\begin{aligned}\mathbb{E}_{\boldsymbol{Q}_0}(\log L_\tau^{(c)}(\boldsymbol{Q})|Y = y) &= \sum_{i=1}^{m}\sum_{j\neq i}\log(q_{ij})\mathbb{E}_{\boldsymbol{Q}_0}(N_{ij}(\tau)|Y = y) \\ &\quad - \sum_{i=1}^{m}\sum_{j\neq i} q_{ij}\mathbb{E}_{\boldsymbol{Q}_0}(R_i(\tau)|Y = y).\end{aligned}$$

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

$$\tilde{M}_{ij}^k(t) = \mathbb{E}_{\boldsymbol{Q}_0}[R_k(t)|X(t) = j, X(0) = i] \tag{3.2}$$

and

$$\tilde{f}_{ij}^{k\ell}(t) = \mathbb{E}_{\boldsymbol{Q}_0}(N_{k\ell}(t)|X(t) = j, X(0) = i) \tag{3.3}$$

because

$$\mathbb{E}_{\boldsymbol{Q}_0}(N_{ij}(\tau)|Y = y) = \sum_{k=1}^{n-1}\tilde{f}_{y_k,y_{k+1}}^{ij}(t_{k+1} - t_k) \tag{3.4}$$

$$\mathbb{E}_{\boldsymbol{Q}_0}(R_\ell(\tau)|Y = y) = \sum_{k=1}^{n-1}\tilde{M}_{y_k,y_{k+1}}^{\ell}(t_{k+1} - t_k). \tag{3.5}$$

In order to calculate (3.2), it turns out to be convenient to study the related functional (we drop the index $\boldsymbol{Q}_0$ 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).

Theorem 3.1 *The function M_{ij}^k solves the differential equation*

$$\frac{d}{dt}M_{ij}^k(t) = \sum_\ell M_{i\ell}^k(t)q_{\ell j} + \exp(t\boldsymbol{Q})_{ij}\delta_{jk} \tag{3.6}$$

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

Define $\boldsymbol{M}^k_{i\cdot}(t) = (M^k_{i1}(t), ..., M^k_{im}(t))$ (row vector). Then (3.6) may be written as

$$\frac{d}{dt}\boldsymbol{M}^k_{i\cdot}(t) = \boldsymbol{M}^k_{i\cdot}(t)\boldsymbol{Q} + \boldsymbol{A}^k_i(t),$$

where $\boldsymbol{A}^k_i(t) = \boldsymbol{e}_i' \exp(\boldsymbol{Q}t)\boldsymbol{e}_k\boldsymbol{e}_k'$ with $\boldsymbol{e}_i$ denoting the unit vector with the ith coordinate equal to 1 and with $\boldsymbol{e}_i'$ denoting its transpose. This is a system of inhomogeneous linear differential equations, and from the initial condition $\boldsymbol{M}^k_{i\cdot}(0) = 0$ it is clear that the solution to the system is

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

Thus on matrix form $\boldsymbol{M}^k = \{M^k_{ij}\}_{ij\in E}$ we have that

$$\boldsymbol{M}^k(t) = \int_0^\infty \exp(s\boldsymbol{Q})\,(\boldsymbol{e}_k\boldsymbol{e}_k')\exp((t-s)\boldsymbol{Q})ds.$$

Now choose $\lambda \geq \max_{i=1,..,m}(-Q_{ii})$ and define $\boldsymbol{B} = \boldsymbol{I} + \frac{1}{\lambda}\boldsymbol{Q} = \frac{1}{\lambda}(\lambda\boldsymbol{I} + \boldsymbol{Q})$. It is clear that $\boldsymbol{B}$ is a stochastic matrix (transition matrix) and

$$\exp(\boldsymbol{Q}t) = \exp(-\lambda t\boldsymbol{I} + \lambda t\boldsymbol{B}) = \sum_{n=0}^\infty e^{-\lambda t}\frac{(\lambda t)^n}{n!}\boldsymbol{B}^n.$$

Calculating matrix–exponentials in this way is referred to as the uniformization method (see Neuts (1995) p. 232) and is known to be very efficient. Then we obtain

$$\begin{aligned}\boldsymbol{M}^k(t) &= \int_0^t \exp(s\boldsymbol{Q})\,(\boldsymbol{e}_k\boldsymbol{e}_k')\exp((t-s)\boldsymbol{Q})ds \\ &= \int_0^t \sum_{i=0}^\infty e^{-\lambda s}\frac{(\lambda s)^i}{i!}\boldsymbol{B}^i\,(\boldsymbol{e}_k\boldsymbol{e}_k')\sum_{j=0}^\infty e^{-\lambda(t-s)}\frac{(\lambda(t-s))^j}{j!}\boldsymbol{B}^j ds \\ &= e^{-\lambda t}\sum_{i,j=0}^\infty \int_0^t \frac{(\lambda s)^i(\lambda(t-s))^j}{i!j!}ds\;\boldsymbol{B}^i\,(\boldsymbol{e}_k\boldsymbol{e}_k')\,\boldsymbol{B}^j \\ &= e^{-\lambda t}\sum_{i,j=0}^\infty \frac{1}{\lambda}\frac{(\lambda t)^{i+j+1}}{(i+j+1)!}\;\boldsymbol{B}^i\,(\boldsymbol{e}_k\boldsymbol{e}_k')\,\boldsymbol{B}^j \\ &= e^{-\lambda t}\sum_{n=0}^\infty \frac{1}{\lambda}\frac{(\lambda t)^{n+1}}{(n+1)!}\sum_{\ell=0}^n \boldsymbol{B}^\ell\,(\boldsymbol{e}_k\boldsymbol{e}_k')\,\boldsymbol{B}^{n-\ell}.\end{aligned}$$

Now we can calculate the quantity (3.2) by

$$\tilde{M}^k_{ij}(t) = M^k_{ij}(t)/\boldsymbol{e}_i\exp(\boldsymbol{Q}t)\boldsymbol{e}_j. \tag{3.7}$$

In order to calculate the quantity (3.3), 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

$$f^{k\ell}_{ij}(t) = \mathbb{E}\left(N_{k\ell}(t)I\{X(t) = j\}|X(0) = i\right). \tag{3.8}$$

for fixed k, ℓ.

Theorem 3.2 *The function $f_{ij}^{k\ell}$ given by (3.8) solves the differential equation*

$$\frac{\partial}{\partial t} f_{ij}^{k\ell}(t) = \sum_{h=1}^{m} f_{ih}^{k\ell} q_{hj} + q_{k\ell} \exp(\boldsymbol{Q}t)_{ik} \delta_{j\ell}, \tag{3.9}$$

with boundary condition $f_{ij}^{k\ell}(0) = 0$ for all i, j.

Proof: In Bladt et al. (2002) the joint transform $V^*(\boldsymbol{s}, Z; t) = \{V_{ij}^*(\boldsymbol{s}, Z; t)\}$ of holding times $R_i(t)$ in state i and number of transitions from state k to ℓ, $N_{k\ell}(t)$, is defined by

$$V_{ij}^*(\boldsymbol{s}, Z; t) = \mathbb{E}\left(\exp\left(-\sum_{h=1}^{m} s_h R_h(t) \right) \prod_{a,b} z_{ab}^{N_{ab}} I\{X(t) = j\} \middle| X(0) = i \right),$$

where $\boldsymbol{s} = (s_1, ..., s_m)$ and $\boldsymbol{Z} = \{z_{ab}\}_{a,b=1,...,m}$ are variables. (Notice that the setting of Bladt et al. (2002) is slightly more general but specialize to the above setting). Thus the transform under consideration is a joint Laplace and generating function type of transform. In order to get hold of the $f_{ij}^{k\ell}(t)$ we set $\boldsymbol{s} = \mathbf{0}$, $z_{ab} = 1$ if $(a, b) \neq (k, \ell)$ and $z_{k\ell} = z$. In Bladt et al. (2002) it is shown that

$$V^*(\boldsymbol{s}, \boldsymbol{Z}; t) = \exp\left((\boldsymbol{Q} \bullet Z + \Delta(\boldsymbol{s})\boldsymbol{I})t \right),$$

where $\boldsymbol{I}$ denotes the identity matrix, $\bullet$ denotes the Schur product (defined as a product between two matrices $\boldsymbol{A} = \{a_{ij}\}$ and $\boldsymbol{B} = \{b_{ij}\}$ by $\boldsymbol{A} \bullet \boldsymbol{B} = \{a_{ij} b_{ij}\}$), and $\Delta(\boldsymbol{s})$ is the diagonal matrix with the numbers $s_1, \ldots, s_m$ as its diagonal. Specializing to our case we get that

$$V^*(\mathbf{0}, \boldsymbol{Z}; \mathbf{t}) = \exp\left(\boldsymbol{Q} \bullet \boldsymbol{Z}\mathbf{t} \right).$$

It is clear from the exponential form that

$$\frac{\partial}{\partial t} V^*(\mathbf{0}, \boldsymbol{Z}; \mathbf{t}) = \mathbf{V}^*(\mathbf{0}, \boldsymbol{Z}; \mathbf{t}) \left[\boldsymbol{Q} \bullet \boldsymbol{Z} \right]$$

or

$$\frac{\partial}{\partial t} V_{ij}^*(\mathbf{0}, \boldsymbol{Z}; \mathbf{t}) \;=\; \sum_{h=1}^{m} V_{ih}^*(\mathbf{0}, \boldsymbol{Z}; \mathbf{t}) [\boldsymbol{Q} \bullet \boldsymbol{Z}]_{\mathbf{hj}}.$$

Differentiating this equation with respect to z yields

$$\begin{aligned} \frac{\partial}{\partial t} \frac{\partial}{\partial z} V_{ij}^*(\mathbf{0}, \boldsymbol{Z}; \mathbf{t}) \;=\; & \sum_{h=1}^{m} \frac{\partial}{\partial z} V_{ih}^*(\mathbf{0}, \boldsymbol{Z}; \mathbf{t}) \left[\boldsymbol{Q} \bullet \boldsymbol{Z} \right]_{\mathbf{hj}} \\ & + \sum_{h=1}^{m} V_{ih}^*(\mathbf{0}, \boldsymbol{Z}; \mathbf{t}) \frac{\partial}{\partial \mathbf{Z}} \left[\boldsymbol{Q} \bullet \boldsymbol{Z} \right]_{\mathbf{hj}}, \end{aligned}$$

and by evaluating this equation at $z = 1$, we obtain the desired equation (3.9). The boundary conditions $f_{ij}^{k\ell}(0) = 0$ for all i, j are obvious. □

Applying the same arguments as when solving for $M_{ij}(t)$, we may write the matrix $\boldsymbol{f}^{k\ell}(t) = \{f_{ij}^{k\ell}(t)\}_{i,j \in E}$ as

$$\boldsymbol{f}^{k\ell}(t) = q_{k\ell} \int_0^t e^{\boldsymbol{Q}s} \left(\boldsymbol{e}_k \boldsymbol{e}_\ell' \right) e^{\boldsymbol{Q}(t-s)} ds.$$

By uniformization we then obtain that

$$\boldsymbol{f}^{k\ell}(t) \;=\; q_{k\ell}e^{-\lambda t}\sum_{n=0}^{\infty}\frac{1}{\lambda}\frac{(\lambda t)^{n+1}}{(n+1)!}\sum_{\ell=0}^{n}\boldsymbol{B}^{\ell}\left(\boldsymbol{e}_k\boldsymbol{e}_\ell'\right)\boldsymbol{B}^{n-\ell}.$$

We can now calculate the quantity $\tilde{f}_{ij}^{k\ell}(s,t)$ defined by (3.3):

$$\tilde{f}_{ij}^{k\ell}(s,t) = f_{ij}^{k\ell}(t-s)/\boldsymbol{e}_i\exp(\boldsymbol{Q}(t-s))\boldsymbol{e}_j. \tag{3.10}$$

We can now sum up the EM algorithm for maximum likelihood estimation of $\hat{\boldsymbol{Q}}$ as follows:

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

1. *Calculate $\tilde{M}_{y_i,y_{i+1}}^{k}(t_{i+1}-t_i)$ and $\tilde{f}_{y_i,y_{i+1}}^{k\ell}(t_{i+1}-t_i)$ for all k,ℓ under the model with intensity matrix $\boldsymbol{Q}$ by (3.7) and (3.10).*

2. *Calculate $\mathbb{E}_{\boldsymbol{Q}}\left(R_i(\tau)|Y=y\right)$ and $\mathbb{E}_{\boldsymbol{Q}}\left(N_{ij}|Y=y\right)$ by (3.4) and (3.5).*

3. *Calculate $\hat{\boldsymbol{Q}}$ by $\hat{\boldsymbol{Q}}_{ij} = \mathbb{E}_{\boldsymbol{Q}}\left(N_{ij}|Y=y\right)/\mathbb{E}_{\boldsymbol{Q}}\left(R_i(\tau)|Y=y\right)$ for all i,j.*

4. *$\boldsymbol{Q} := \hat{\boldsymbol{Q}}$. GOTO 1.*

Let $\boldsymbol{Q}_0, \boldsymbol{Q}_1, \boldsymbol{Q}_2, ...$ be a sequence of intensity matrices obtained by the EM–algorithm. Then certainly $L_n(\boldsymbol{Q}_{k+1}) \geq L_n(\boldsymbol{Q}_k)$ for $k = 0, 1, 2, \ldots$, where L_n is the discrete time likelihood function (2.5), see Dempster, Laird & Rubin (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 & Krishnan (1997). Unfortunately, one of Wu's conditions, condition (3.19) in McLachlan & Krishnan (1997), is not satisfied by the model treated here. In the 2-state case considered in Example 2.2 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 $\det(\exp(\boldsymbol{Q})) \to 0$. One way around this problem is to use the slightly smaller parameter space

$$\mathcal{Q}_\epsilon = \{\boldsymbol{Q} \in \mathcal{Q} \,|\, \det(\exp(\boldsymbol{Q})) \geq \epsilon\}$$

for some small $\epsilon > 0$. With this restricted parameter set, it is clear that condition (3.19) in McLachlan & 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 & Krishnan (1997), which by Theorem 3.2 in that book would imply the convergence of the sequence $\{\boldsymbol{Q}_k\}$. Condition (3.18) with $d = m(m-1)$ is trivial, and condition (3.20) that the function $\boldsymbol{Q} \mapsto L_n(\boldsymbol{Q})$ is continuous and differentiable in the interior of the parameter space follows from the fact that the function $\boldsymbol{Q} \mapsto \exp(\boldsymbol{Q})$ 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 e.g. Neuts (1995). The continuity of the function

$$(\boldsymbol{Q}, \boldsymbol{Q}_0) \to \mathbb{E}_{\boldsymbol{Q}_0}(\log L_\tau^{(c)}(\boldsymbol{Q})|Y=y),$$

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

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

is satisfied for the full parameter space $\mathcal{Q}$, provided that the initial matrix $\boldsymbol{Q}_0$ is chosen in the interior of $\mathcal{Q}$. To see this, note that for any $\boldsymbol{Q}_0$ 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 obtained by using these expected values as the statistics in L_n have strictly positive off-diagonal elements (cf. 2.3), and hence $\boldsymbol{Q}_1$ belongs to the interior of $\mathcal{Q}$. Iteration of this argument shows that $\boldsymbol{Q}_k$ belongs to the interior of $\mathcal{Q}$ for all k. (Note that some $(\boldsymbol{Q}_k)_{ij}$ may well converge to zero as $k \to \infty$). However for the restricted parameter space $\mathcal{Q}_\epsilon$, it may happen that the sequence $\boldsymbol{Q}_k$ converges to the boundary where $\det(\exp(\boldsymbol{Q})) = \epsilon$ and that $\det(\exp(\boldsymbol{Q})_k) = \epsilon$ for some k. Then condition (3.21) in McLachlan & Krishnan (1997) will typically not be satisfied. In view of Theorem 3.2 in McLachlan & Krishnan (1997) we can summarize the discussion as follows.

Theorem 3.3 *Suppose the initial matrix $\boldsymbol{Q}_0$ belongs to the interior of the parameter space $\mathcal{Q}$, i.e. that $(\boldsymbol{Q}_0)_{ij} > 0$ for all $i \neq j$. Then the sequence $\{\boldsymbol{Q}_k\}$ will either converge to a stationary point of the likelihood function L_n or* $\det(\exp(\boldsymbol{Q}_k)) \to 0$.

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 discussed in the previous section. Obviously, it is a good idea to choose the initial matrix $\boldsymbol{Q}_0$ in such a way that $\det(\exp(\boldsymbol{Q}_k))$ is far from zero. If $\boldsymbol{Q}_0$ is chosen such that some $(\boldsymbol{Q}_0)_{ij} = 0$, then the expected number of jumps from i to j will remain zero through all iterations, i.e. all $\boldsymbol{Q}_k$ 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 $\boldsymbol{Q}_0$ such that some $(\boldsymbol{Q}_0)_{ij} = 0$, a convergence result similar to Theorem 3.3 can be obtained by reducing the parameter space by the restriction $q_{ij} = 0$.

Use of the restricted parameter space, $\mathcal{Q}_\epsilon$, is a rather crude way to solve the problem at the boundary where $\det(\exp(\boldsymbol{Q})) \to 0$ and is mainly a technical device to prove Theorem 3.3. 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 zero as $\det(\exp(\boldsymbol{Q})) \to 0$. The EM algorithm can also be applied to maximum penalized likelihood estimation, see McLachlan & Krishnan (1997). An obvious way to penalize the likelihood is provided by the conjugate priors discussed in the next section, where a Markov chain Monte Carlo method is presented as an alternative to the EM algorithm.

4 Markov chain Monte Carlo estimation

In this section we present a second approach to estimating the parameters of a discretely observed Markov jump processes which uses the methodology of Markov chain Monte Carlo. We present this approach in a slightly more general setting than the one 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 $p = p_1+p_2+...+p_m$ states and intensity matrix $\boldsymbol{Q}$. A new process $\{X(t)\}$ is defined in the following way:

$$X(t) = i \quad \Longleftrightarrow \quad J(t) \in \{p_{i-1}+1, ..., p_i\}, \ i = 1, 2, ...m,$$

where $p_0 = 0$. Thus we have grouped the states of J, and X indicate which group the process J is in at any given time. The process $\{X(t)\}$ is in general not a Markov process, since the sojourn times in states $1, 2, \ldots, m$ are not necessarily exponentially distributed. In fact, the time spent in a state from entrance into the state until X jumps away again is phase–type distributed (see e.g. Neuts (1981) or Asmussen (2003)). In this section we consider discrete time observations of X, and the purpose is to estimate the intensity matrix $\boldsymbol{Q}$ of the Markov jump process J underlying the non–Markovian process X to the extent this is possible. If $p_i = 1$ for all $i \in E$, then we may estimate the parameters of $\boldsymbol{Q}$ 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 $\boldsymbol{Q}$ is no longer unique (phase–type representations are not unique), and it is not possible to estimate all parameters of $\boldsymbol{Q}$ by Markov chain Monte Carlo, which will be apparent from the following discussion. It will, however, be possible to estimate functionals that are invariant under the different representations. An example is the (time–dependent) rates of transitions between the different states $1, 2, \ldots, m$ of the process X.

We decompose the intensity matrix $\boldsymbol{Q}$ in the following way:

$$\boldsymbol{Q} = \left(\begin{array}{c|c|c|c|c} \boldsymbol{Q}_{11} & \boldsymbol{Q}_{12} & \boldsymbol{Q}_{13} & ... & \boldsymbol{Q}_{1m} \\ \hline \boldsymbol{Q}_{21} & \boldsymbol{Q}_{22} & \boldsymbol{Q}_{23} & ... & \boldsymbol{Q}_{2m} \\ \hline \boldsymbol{Q}_{31} & \boldsymbol{Q}_{32} & \boldsymbol{Q}_{33} & ... & \boldsymbol{Q}_{3m} \\ \hline ... & ... & ... & ... & ... \\ \hline \boldsymbol{Q}_{m1} & \boldsymbol{Q}_{m2} & \boldsymbol{Q}_{m3} & ... & \boldsymbol{Q}_{mm} \end{array} \right).$$

Here $\boldsymbol{Q}_{ij}$ is a $p_i \times p_j$–matrix. If $i = j$ it is the subintensity matrix generating the phase–type distribution of the time until J first leaves the set of states $\{p_{i-1}, ..., p_i\}$. If $i < j$ the k'th row of $\boldsymbol{Q}_{ij}$ is proportional to the initial distribution for the phase–type distribution initiating at state j when the previous state of the process X was i with J exiting the ith group from the sub–state k. Similarly for $i > j$.

Consider the discrete time observations $\boldsymbol{x} = (x_1, ..., x_n)$ of the continuous time jump process $\{X(t)\}_{t\geq 0}$ observed at times $t_1, ..., t_n$ up to time τ ($t_1 = 0$ and $t_n = \tau$). The framework is essentially Bayesian. We choose a prior $\phi(\boldsymbol{Q})$ and are interested in the conditional distribution of $\boldsymbol{Q}$ given the data $\boldsymbol{x}$. We shall, however, study the slightly more general problem of finding the conditional distribution of $(\boldsymbol{Q}, \boldsymbol{J})$ given $\boldsymbol{x}$, where $\boldsymbol{J} = \{J(t)\}_{0\leq t\leq \tau}$ denotes the continuous time sample path of J.

As earlier, let $\mathcal{Q}$ denote the space of intensity matrices, and let $E_{\boldsymbol{J}}$ be the space of continuous time sample paths of a Markov jump processes with p states observed up to time τ. We must construct a Markov chain taking values in $\mathcal{Q} \times E_{\boldsymbol{J}}$ the stationary distribution of which is equal to the conditional distribution of $(\boldsymbol{Q}, \boldsymbol{J})$ given $\boldsymbol{x}$.

There are several ways in which one may construct such a chain, the simplest being the Gibbs sampler. More generally one could apply a Metropolis–Hastings algorithm, which has the Gibbs sampler as a special case. The Gibbs sampler is suitable for our purpose: we can sample from the posterior distribution using a Gibbs sampler with two sites, $\boldsymbol{Q}$ and $\boldsymbol{J}$.

We sample by alternately drawing $\boldsymbol{J}$ given $(\boldsymbol{Q}, \boldsymbol{x})$ and $\boldsymbol{Q}$ given $(\boldsymbol{J}, \boldsymbol{x})$ ($\boldsymbol{x}$ is of course of no importance when conditioning on $\boldsymbol{J}$). Iteration of the Gibbs sampler results in a sequence of variables $(\boldsymbol{Q}_n, \boldsymbol{J}_n)$. Under suitable conditions the Gibbs sampler will eventually produce a stationary and ergodic sequence, that is, after discarding a certain burn–in period, say the first K-1 iterations, the sequence $(\boldsymbol{Q}_n, \boldsymbol{J}_n)_{n \geq K}$ may be considered stationary.

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

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

converges to the true mean of $\boldsymbol{Q}$ conditionally on $\boldsymbol{x}$. Also credibility intervals based on the empirical distribution of $(\boldsymbol{Q}_n, \boldsymbol{J}_n)_{n \geq K}$ may be constructed, and quantiles of the empirical distribution may be of interest too. In situations where $\boldsymbol{Q}$ is not uniquely determine by the distribution of the discrete time sample, the mean of the posterior distribution may not be a meaningful quantity, but functionals of the type discussed in the case where some $p_i > 1$ below can still be estimated. As discussed in Section 2 the set of $\boldsymbol{Q}$s for which this happens is complicated, so it is important to study the posterior distribution carefully for indications that this problem has occurred, for instance by inspecting scatter plots like those in Section 5. It might seem desirable to use a prior that is concentrated on the set of $\boldsymbol{Q}$s for which $\exp(\boldsymbol{Q}) \in \mathcal{P}_{00}$, but since this set is very complicated, this idea would be very difficult to implement. An easier, but less satisfactory, solution is a prior concentrated on the set of $\boldsymbol{Q}$s for which $\exp(\boldsymbol{Q})$ satisfies (2.12).

If some $p_i > 1$ and the phase-type representations are no longer unique, it is not possible to estimate $\boldsymbol{Q}$ through simple averaging of the $\boldsymbol{Q}_i$'s since the representations may switch through the iterations. Even more, credibility intervals hardly makes sense for parameters which are not uniquely determined. Functionals invariant under different representations may, however, conveniently be calculated using this method. Specifically, let $F(\cdot)$ be some functional which depends on the distribution of the process $X(t)$ and is invariant under changes of the representation $\boldsymbol{Q}$ (i.e. if $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$ are two representations resulting in the same distribution of the process $X(t)$, then $F(\boldsymbol{Q}_1) = F(\boldsymbol{Q}_2)$). Then we can estimate $F(\boldsymbol{Q})$ by

$$\frac{1}{N} \sum_{i=K}^{N+K} F(\boldsymbol{Q}_i).$$

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

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

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

(1997). Indeed, the posterior is

$$\begin{aligned} p^*(\boldsymbol{Q}) &= L_\tau^{(c)}(\boldsymbol{Q})\ \phi(\boldsymbol{Q}) \\ &\propto \prod_{i=1}^{n}\prod_{j\neq i} q_{ij}^{N_{ij}(\tau)+\alpha_{ij}-1} e^{-q_{ij}(R_i(\tau)+\beta_i)}, \end{aligned}$$

where the likelihood function $L_\tau^{(c)}(\boldsymbol{Q})$ is given by (2.1). The Gibbs sampler now works as follows.

1. *Draw initial* $\boldsymbol{Q}$ *from the prior.*
2. *Simulate a Markov jump process* J_t *with intensity matrix* $\boldsymbol{Q}$ *up to time* τ *such that* $X(t_i) = x_i$.
3. *Calculate the statistics* $N_{ij}(\tau)$ *and* $R_i(\tau)$ *from* $\{J(t)\}_{0\leq t\leq\tau}$.
4. *Draw a new* $\boldsymbol{Q}$ *from the posterior distribution.*
5. *GO TO 2.*

Remark 4.1 Simulating Markov jump processes $J(t)$ such that $J(t_i) \in \{p_{x_i-1}, ..., p_{x_i}\}$ can be done in several ways. Since we do not have have detailed information about the sub–states, we may initiate in any sub–state state $k_1 \in \{p_{x_1-1}+1, ..., p_{x_1}\}$. There are several ways in which to proceed. The simplest is to simulate Markov jump processes J_t up to time t_2 such that $J(t_2) \in \{p_{x_2-1}+1, ..., p_{x_2}\}$. This can be done by simple rejection if the criterion is not met and acceptance otherwise. Observe the state $k_2 = J(t_2)$. Simulate Markov processes initiating from k_2 until $J(t_3) \in \{p_{x_3-1}+1, ..., p_{3_2}\}$ and so on.

An efficient way to simulate Markov jump processes which have to pass through sets of possible states is to choose one particular trajectory among many possibles according to current transition rates $\boldsymbol{Q}$. Hence from the incomplete discrete data $x_1, ..., x_N$ we construct complete discrete data $y_1, .., y_N$ in the following way. Suppose that we have chosen some state i at time t_ℓ. The probability of $J(t)$ being in state j at time $t_{\ell+1}$ is $\boldsymbol{e}_i' \exp(\boldsymbol{Q}(t_{\ell+1} - t_\ell))\boldsymbol{e}_j$. Since j must belong to $\{p_{\ell-1}+1, ..., p_\ell\}$, we simply choose among the states $\{p_{\ell-1}+1, ..., p_\ell\}$ from the corresponding conditional distribution.

5 Example

In this section we present a simulation study that compare the two methods of estimating a discretely observed Markov jump process. We have simulated a sample path of a Markov jump process with four states and with intensity matrix

$$\boldsymbol{Q} = \begin{pmatrix} -1.00 & 0.25 & 0.25 & 0.50 \\ 0.20 & -1.50 & 0.30 & 1.00 \\ 0.80 & 0.80 & -2.0 & 0.40 \\ 1.00 & 0.60 & 0.90 & -2.50 \end{pmatrix}$$

in the time interval $[0, 250]$. The data are the states of the process at 500 time points, equidistantly displaced by 0.5. The maximum likelihood estimator of the transition matrix of the discrete time Markov chain, given by (2.8), is

$$\hat{\boldsymbol{P}}_{\text{obs}} = \begin{pmatrix} 0.6546 & 0.0928 & 0.1186 & 0.1287 \\ 0.1466 & 0.5172 & 0.1034 & 0.2328 \\ 0.1905 & 0.2500 & 0.4048 & 0.1548 \\ 0.3113 & 0.1604 & 0.1415 & 0.3868 \end{pmatrix}.$$

Since we have the full continuous time sample path, we can also calculate the maximum likelihood estimate of the intensity matrix based on the continuous–time likelihood, cf. (2.3),

$$\hat{\boldsymbol{Q}}_{\text{cont}} = \begin{pmatrix} -1.084 & 0.2329 & 0.2734 & 0.5773 \\ 0.1352 & -1.5544 & 0.2872 & 1.1320 \\ 0.7162 & 0.8891 & -2.1734 & 0.5680 \\ 1.3502 & 0.6365 & 0.8487 & -2.8353 \end{pmatrix}.$$

Next we compare the EM–algorithm and the MCMC approach on the discrete time data. The EM–algorithm converged in less than 500 iterations to the intensity matrix

$$\hat{\boldsymbol{Q}}_{\text{EM}} = \begin{pmatrix} -1.0258 & 0.1771 & 0.4067 & 0.4420 \\ 0.2037 & -1.5694 & 0.2922 & 1.0734 \\ 0.5429 & 1.0672 & -2.0759 & 0.4659 \\ 1.2145 & 0.5492 & 0.5848 & -2.3487 \end{pmatrix}.$$

The one–step transition probabilities of the discrete time Markov chain with step length 0.5 corresponding to $\hat{\boldsymbol{Q}}_{\text{EM}}$, $\hat{\boldsymbol{P}}^{0.5}_{\text{EM}} = \exp(0.5\boldsymbol{\Lambda}_{\text{EM}})$, is

$$\hat{\boldsymbol{P}}^{0.5}_{\text{EM}} = \begin{pmatrix} 0.6580 & 0.0933 & 0.1192 & 0.1295 \\ 0.1466 & 0.5172 & 0.1034 & 0.2328 \\ 0.1905 & 0.2500 & 0.4048 & 0.1548 \\ 0.3113 & 0.1604 & 0.1415 & 0.3868 \end{pmatrix}.$$

We note that $\hat{\boldsymbol{P}}^{0.5}_{\text{EM}}$ and $\hat{\boldsymbol{P}}_{\text{obs}}$ are practically identical, the discrepancy being due to numerical errors. Thus the EM–algorithm has found the maximum likelihood estimator of $\boldsymbol{Q}$. To check whether $\hat{\boldsymbol{Q}}_{\text{EM}}$ is uniquely determined by $\hat{\boldsymbol{P}}_{\text{obs}}$, we can apply the criteria in Section 2. It turns out that (2.11) is too weak in this case, but the criterion (2.12) shows that $\hat{\boldsymbol{Q}}_{\text{EM}}$ is indeed unique.

Concerning the MCMC, we performed a larger experiment drawing 10,000 intensity matrices including an initial burn–in of 1,000 iterations. The values of the parameters in the prior were simply set at $\alpha_{ij} = \beta_i = 1$. The average of the 9,000 intensity matrices is

$$\hat{\boldsymbol{Q}}_{\text{MCMC}} = \begin{pmatrix} -1.0873 & 0.1905 & 0.4266 & 0.4701 \\ 0.2493 & -1.6756 & 0.3262 & 1.1001 \\ 0.5690 & 1.1218 & -2.2384 & 0.5476 \\ 1.2424 & 0.5962 & 0.6602 & -2.4988 \end{pmatrix},$$

which is close to the maximum likelihood estimator $\hat{\boldsymbol{Q}}_{\text{EM}}$ as one would expect.

Quantiles of the empirical distribution of the 9,000 simulated intensity matrices were calculated and are listed in Table 5.1.

These quantiles give a good impression of how well the parameters are determined by the data marginally. To get an idea of how dependent the MCMC-estimates of the entries in the intensity matrix are, we made scatter plots of the values of $(q_{i_1,j_1}, q_{i_2,j_2})$ for all the 9000 matrices we have generated for each of the 66 combinations of $(i_1, j_1), (i_2, j_2)$, $i_2 \neq j_2$. Most of these plots are similar, so we have chosen to present 6 typical examples in Figure 5.1. Of the 66 scatter plots, 28 are similar to the two plots in the first row, 21 are similar to the plots in the second row, 9 are similar to the first plot in the last row, and 8 are similar to the last plot. The estimates are not very dependent.

Transition	2.5 %	5 %	50 %	95 %	97.5 %
1-2	0.021	0.038	0.1832	0.3700	0.4095
1-3	0.2043	0.2356	0.4174	0.6563	0.7097
1-4	0.2247	0.2589	0.4587	0.7239	0.7864
2-1	0.0165	0.0333	0.2323	0.5272	0.5840
2-3	0.0382	0.0658	0.3066	0.6665	0.7658
2-4	0.6431	0.7033	1.0828	1.5558	1.6593
3-1	0.1493	0.2115	0.5482	0.9960	1.1045
3-2	0.5741	0.6461	1.0937	1.6863	1.8219
3-4	0.0809	0.1245	0.5045	1.1061	1.2530
4-1	0.7463	0.8137	1.2255	1.7297	1.8357
4-2	0.1544	0.2196	0.5768	1.0456	1.1444
4-3	0.1930	0.2516	0.6311	1.1623	1.3153

Table 5.1: Quantiles of the posterior distribution of the entries q_{ij} of the intensity matrix.

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 a 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 fast compared to the sampling frequency, which implies that a lot 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 fast 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 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,

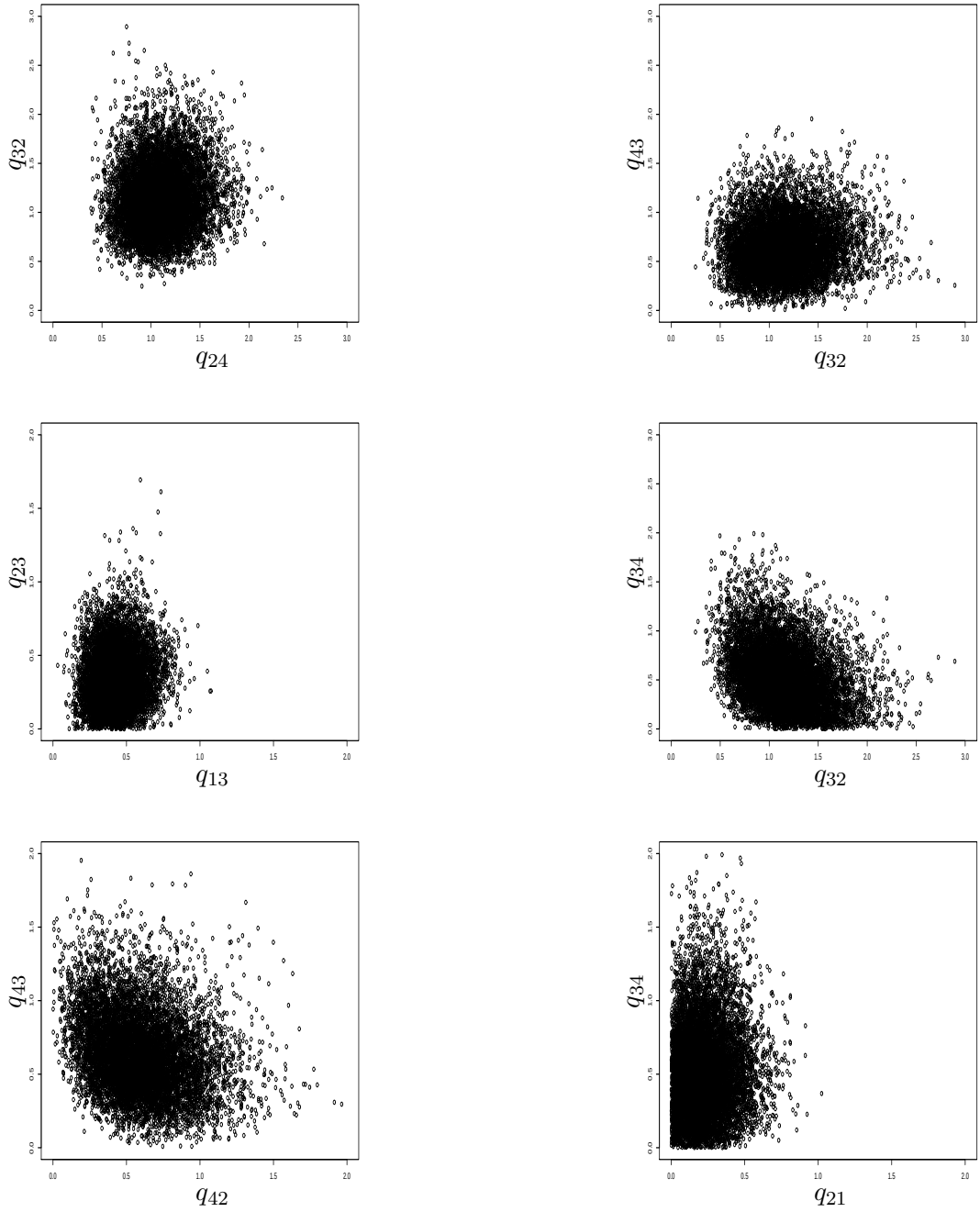

Figure 5.1: Scatter plots of the MCMC-simulations from the posterior distribution of the pairs $(q_{i_1,j_1}, q_{i_2,j_2})$ for values of i_1, j_1, i_2, j_2 that show typical patterns.

of course, not possible that both the original process and the new process are Markovian, so the results 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 depends quite a bit on the prior. A more serious problem is that the MCMC approach may hide

problems of non-existence 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

Most of the research presented here was done while Mogens Bladt visited the Department of Applied Mathematics and Statistics, University of Copenhagen, a visit that was partly financed by the Centre for Mathematical Physics and Stochastics funded by The Danish National Research Foundation. The research of Michael Sørensen was supported by the European Commission through the Research Training Network DYNSTOCH under the Human Potential Programme, by MaPhySto – A Network in Mathematical Physics and Stochastics funded by The Danish National Research Foundation, 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.

Asmussen, S. (2003). *Applied Probability and Queues*. Springer Verlag, Heidelberg-New York.

Bibby, B. M.; Jacobsen, M. & Sørensen, M. (2003). "Estimating functions for discretely sampled diffusion-type models". In Aït-Sahalia, Y. & Hansen, L. P., editors, *Handbook of Financial Econometrics*. Amsterdam: North-Holland. Forthcoming.

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

Bladt, M.; Neuts, M. F.; Meini, B. & Sericola, B. (2002). "Distributions of Reward Functions on Continuous-time, Markov chains". In G. Latouche, P. T., editor, *Matrix-Analytic Methods: Theory and Applications*. 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. London Math. Soc.*, 4:623 – 630.

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

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

Elerian, O.; Chib, S. & 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. Finnicae A*, 2.

Hoffmann, M. (1999). "L_p-estimation of the Diffusion Coefficient". *Bernoulli*, 5:447–481.

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

Jacobsen, M. (1982). *Statistical Analysis of Counting Processes*. Springer-Verlag, New York. Lecture Notes in Statistics 12.

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

Johansen, S. (1974). "Some results on the imbedding problem for finite Markov chains". *J. London 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". *Annals of Statistics*, 3:363 – 372.

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

Kingman, J. F. C. (1962). "The imbeding problem for finite Markov chains". *Z. Wahrscheinlichkeitstheorie*, 1:14–24.

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

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

Neuts, M. F. (1981). *Matrix-Geometric Solutions in Stochastic Models*. John Hopkins University Press.

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

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

Wu, C. J. F. (1983). "On the convergence properties of the EM algorithm". *Annals of Statistics*, 11:95–103.