Maximum likelihood estimation in the two-state Markovian arrival process

Emilio Carrizosa\textsuperscript{a} and Pepa Ramírez-Cobo\textsuperscript{b}

\textsuperscript{a}Departamento de Estadística e Investigación Operativa, Universidad de Sevilla (Spain),
\textsuperscript{b} IMUS Instituto de Matemáticas de la Universidad de Sevilla (Spain)

January 15, 2014

Abstract

The Markovian arrival process (MAP) has proven a versatile model for fitting dependent and non-exponential interarrival times, with a number of applications to queueing, teletraffic, reliability or finance. Despite theoretical properties of MAPs and models involving MAPs are well studied, their estimation remains less explored. This paper examines maximum likelihood estimation of the second-order MAP using a recently obtained parameterization of the two-state MAPs.

Key words: Maximum likelihood estimation; Markovian arrival processes; Hidden Markov models; Kullback-Leibler divergence

1 Introduction

Since Neuts (1979) described the Markovian arrival processes (MAPs for short) for the first time, a number of works have dealt with theoretical properties and applications of such point processes. In particular, because of their versatility, many uses in queueing, teletraffic, reliability or finance have been suggested. For a recent account of the literature on MAPs applications, we refer the reader to [Kim & Kim (2010); Wu et al (2011); Okamura et al (2009); Casale et al (2010); Montoro-Cazorla et al (2009); Badescu et al (2007); Cheung & Landriault (2010)]

The versatile character of MAPs is due to two main properties; on the one hand, the interarrival times (i.e, the times between epochs of occurrence of a certain event) in a MAP have a phase-type distribution, which is a rather convenient and flexible framework for fitting realworld data, see for example
On the other hand, the MAP allows for correlated interarrival times, a feature increasingly present in a number of real data traces. While performance analysis for models incorporating MAPs is a well-developed area, less progress has been made on statistical estimation for such models. The MAP is a complex model which includes transitions to hidden states between real arrivals. In practice, only inter-arrival time data are usually observed and therefore, in this context, the observed data can be viewed as being generated from a hidden Markov process. See e.g. Ephraim & Merhav (2002).

The simplest MAP is the two-state MAP, called hereafter MAP$_2$. The MAP$_2$ is usually represented in terms of six parameters, see for instance Eum et al. (2007), Bodrog et al. (2008) or Ramírez-Cobo & Lillo (2012). However, such representation in terms of 6 parameters overparameterizes the process, making it unidentifiable: different MAP$_2$ parameterizations produce the very same joint density for any sequence of inter-arrival times. In the context of statistical inference, this implies that it is not sensible to estimate the individual parameters of the MAP$_2$ given a sample of inter-arrival time data, since different parameters represent the same process. Several papers have investigated a moments matching approach for parameters inference, as is the case of Horváth & Telek (2002), Telek & Horváth (2007), Eum et al. (2007), Bodrog et al. (2008) or Casale et al. (2010). However, in these references, the issue of identifiability of the model has not been taken into account (being Telek & Horváth (2007) and Bodrog et al. (2008) an exception). Maximum likelihood estimation has been proposed in Breuer (2002), Klemm et al. (2003) and Okamura et al. (2009), the EM algorithm being the tool suggested in such papers. The non-identifiability of the representation used in terms of 6 parameters has serious negative consequences: the likelihood function has infinitely many global maxima, and, on top of this, the likelihood function may be highly multimodal, implying that standard methods such as the suggested EM algorithm, will be strongly dependent on the starting values for these algorithms, and they run the risk of getting stuck at a poor local maximum.

Recently Bodrog et al. (2008) solve the identifiability problem for the MAP$_2$ by providing a canonical/unique representation of the process, so that the infinitely many equivalent parameterizations are reduced to a single one. This work is intended as an attempt to gain insight into the maximum likelihood estimation of the MAP$_2$. Unlike previous studies, we do not use the EM algorithm, which calls for very time-consuming simulations in the “F” phase. Instead, our analysis is based on the direct maximization of the likelihood function.
This paper is organized as follows. After a brief review of the second-order MAP in Section 2, we discuss in Section 3 how to compare estimators in the MAP\(^2\). Then we describe in Section 4 the optimization problem consisting of maximizing the likelihood function. Such maximization is not trivial, since technical problems appear for evaluating the objective and, needless to say, to optimize it. The encountered numerical difficulties and the way to avoid them are pointed out in detail, and numerical illustrations are shown.

Finally, Section 5 discusses the findings and delineate some possible directions for future research.

2 Preliminaries on MAP\(^2\)s

The MAP\(^2\) is a doubly stochastic process \(\{J(t), N(t)\}\), where \(J(t)\) represents an irreducible, continuous, Markov process with state space \(S = \{1, 2\}\) and \(N(t)\) is a counting process. See \cite{Neuts, 1979, Lucantoni, 1993, Ramírez-Cobo et al., 2010, Ramírez-Cobo & Lillo, 2012}.

The MAP\(^2\) behaves as follows: the initial state \(i_0 \in S\) is generated according to the initial probability vector \(\theta = (\theta, 1 - \theta)\) and at the end of an exponentially distributed sojourn time in state \(i\), with mean \(1/\lambda_i\), two possible state transitions can occur. First, with probability \(0 \leq p_{ij} \leq 1\) a single arrival occurs and the MAP\(^2\) enters a state \(j \in S\), which may be the same \((j = i)\) or different to \((j \neq i)\) the previous state. On the other hand, with probability \(0 \leq p_{ij} \leq 1\), no arrival occurs and the MAP\(^2\) enters a different state \(j \neq i\).

A stationary MAP\(^2\) can thus be expressed in terms of the parameters \(\{\lambda, P_0, P_1\}\), where \(\lambda = (\lambda_1, \lambda_2)\), and \(P_0\) and \(P_1\) are 2×2 transition probability matrices with elements \(p_{ij} (i \neq j)\) and \(p_{ij1}\), respectively. Instead of transition probability matrices, any MAP\(^2\) can also be characterized by \(\{D_0, D_1\}\), in terms of the rate matrices,

\[
D_0 = \begin{pmatrix}
-\lambda_1 & \lambda_1 p_{120} \\
\lambda_2 p_{210} & -\lambda_2
\end{pmatrix}, \quad D_1 = \begin{pmatrix}
\lambda_1 p_{111} & \lambda_1 (1 - p_{120} - p_{111}) \\
\lambda_2 p_{211} & \lambda_2 (1 - p_{210} - p_{211})
\end{pmatrix}.
\]

The matrix \(D_0\) is assumed to be stable, and as a consequence, it is nonsingular and the sojourn times are finite with probability 1. The definition of \(D_0\) and \(D_1\) implies that \(D = D_0 + D_1\) is the infinitesimal generator of the underlying Markov process, with stationary probability vector \(\pi = (\pi, 1 - \pi)\), computed as \(\pi D = 0\).

The MAP\(^2\) can be viewed as a Markov renewal process. Indeed, let \(X_n\) denote the state of the MAP\(^2\) at the time of the \(n\)th arrival, and let \(T_n\) denote the time between the \((n-1)\)st and \(n\)th arrival. Then \(\{X_{n-1}, T_n\}_{n=1}^\infty\)
is a Markov renewal process, and in particular, \( \{X_n\}_{n=1}^{\infty} \) is a Markov chain whose transition matrix \( P^* \) is given by

\[
P^* = (-D_0)^{-1}D_1. \tag{2}
\]

In practice only partial information of the MAP\(_2\) is observed. It is assumed that the sequence of interarrival times \( \{T_n\}_{n=1}^{\infty} \) is observed, but the states where arrivals occur \( \{X_n\}_{n=1}^{\infty} \) are not.

Special attention deserves the analysis of the random variable \( T \), the time between two successive arrivals in the stationary version of a MAP\(_2\). Its moments are computed as

\[
\mu_n = E(T^n) = n! \phi (-D_0)^{-n} e, \tag{3}
\]

where \( \phi = (\phi, 1 - \phi) \) is the probability distribution satisfying \( \phi P^* = \phi \), and \( e \) is a vector with all its coordinates equal to one.

The likelihood function for a sequence of interarrival times in the stationary version of the MAP\(_2\) is given by

\[
f(t_1, t_2, \ldots, t_n | D_0, D_1) = \phi e^{D_0 t_1} D_1 e^{D_0 t_2} D_1 \ldots e^{D_0 t_n} D_1 e. \tag{4}
\]

Observe that the MAP allows for correlated inter-arrival times, thus the likelihood function in (4) does not decompose into the product of the marginal likelihoods of the different terms. The coefficient \( \rho_k \) of autocorrelation of lag \( k \) is given by

\[
\rho_k = \gamma^k \frac{\mu_2}{\mu_2 - \mu_1^2}, \quad \text{for } k > 0, \tag{5}
\]

where \( 0 \leq \gamma < 1 \) is one of the two eigenvalues of the transition matrix \( P^* \) (since \( P^* \) is stochastic, then necessarily the other eigenvalue is equal to 1), \( \text{(Bodrog et al., 2008).} \)

The expression (1) for the MAP\(_2\) in terms of 6 parameters is known to be overparameterized, \( \text{Ramírez-Cobo et al. (2010).} \) However, \( \text{Bodrog et al. (2008)} \) provide a unique, canonical representation for the MAP\(_2\) in terms of just four parameters. Such canonical representation is the one we are using in this paper. Specifically, if the correlation parameter \( \gamma \) in (5) is positive, then the canonical form of the MAP\(_2\) is given by

\[
D_0 = \begin{pmatrix} x & y \\ 0 & u \end{pmatrix}, \quad D_1 = \begin{pmatrix} -x - y & 0 \\ v & -u - v \end{pmatrix}. \tag{6}
\]

On the other hand, for those MAP\(_2\)s such that \( \gamma \leq 0 \), then their canonical form is

\[
D_0 = \begin{pmatrix} x & y \\ 0 & u \end{pmatrix}, \quad D_1 = \begin{pmatrix} 0 & -x - y \\ -u - v & v \end{pmatrix}, \tag{7}
\]

where, \( x, u \leq 0, y, v \geq 0, x + y \leq 0, u + v \leq 0. \)
3 Comparing estimators

Our aim is to derive (maximum likelihood) estimates of the parameters of the \( \text{MAP}_2 \)'s. This would allow one, for instance, to make inference on the distribution function of the random variable \( T \), or to properly simulate the process.

A remarkable issue is that \( \text{MAP}_2 \)'s may have very similar behavior, despite being represented by rather different parameters. This is not as trivial as traditional overparameterization suggests, since the \( \text{MAP}_2 \) has been analyzed using the overparameterized form \([1]\): pretty different parameters sets are fully equivalent, in the sense that they represent exactly the same \( \text{MAP}_2 \). Even if the canonical form \([6]-[7]\) is used, and thus no identifiability problems exist, different parameters may yield very similar \( \text{MAP}_2 \)'s. In other words, closeness of two \( \text{MAP}_2 \)'s is not correctly measured in terms of the (euclidean) distance between the parameters identifying them. In order to adequately compare different estimators we may use an appropriate similarity measure between the processes they represent. In particular, we measure closeness between parameters representing two \( \text{MAP}_2 \)'s by an empirical Kullback-Leibler divergence \((\text{KL})\) of their interarrival times joint density functions: Given two \( \text{MAP}_2 \)'s, with associated matrices \( \{D_0, D_1\} \) and \( \{\hat{D}_0, \hat{D}_1\} \), given the length \( n \) of the observed sequences and the number \( N \) of runs the experiment is repeated, we will measure the closeness between two \( \text{MAP}_2 \)'s by means of the empirical KL divergence \( D_{KL}\)

\[
D_{KL}\left(\{D_0, D_1\}||\{\hat{D}_0, \hat{D}_1\}\right) := \frac{1}{N} \sum_{i=1}^{N} \log \frac{f(t^{(i)}|\{D_0, D_1\})}{f(t^{(i)}|\{D_0, D_1\})},
\]

where, for \( i = 1, 2, \ldots, N \), \( t^{(i)} = (t_1^{(i)}, \ldots, t_n^{(i)}) \) is a sequence of interarrival times generated from \( \{D_0, D_1\} \).

**Example 1.** As an example, we consider a sample of \( n = 500 \) interarrival times simulated from the \( \text{MAP}_2 \) with canonical form

\[
D_0 = \begin{pmatrix} -20 & 6 \\ 0 & -0.5 \end{pmatrix}, \quad D_1 = \begin{pmatrix} 14 & 0 \\ 0.0426 & 0.4574 \end{pmatrix}.
\]

(8)

We want to compare estimates as obtained from the method of moments, as discussed in Section 4.1 below. The theoretical and empirical moments are given respectively by

\[
(\rho_1, \mu_1, \rho_2, \mu_3) = (0.0864, 1.6802, 6.6887, 40.1276),
(\hat{\rho}_1, \hat{\mu}_1, \hat{\rho}_2, \hat{\mu}_3) = (0.0643, 1.6494, 7.0219, 44.1291).
\]

(9)
The estimate is given by
\[
\hat{D}^{(1)}_0(0) = \begin{pmatrix} -999.9998 & 500.5033 \\ 0 & -0.4735 \end{pmatrix}, \quad \hat{D}^{(1)}_1(0) = \begin{pmatrix} 499.4965 & 0 \\ 0.1315 & 0.3420 \end{pmatrix},
\]
with moments given by
\[
(\hat{\rho}_1, \hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3) = (0.0643, 1.6538, 6.9842, 4.2471).
\]
The superscript \(^{(1)}\) in (10) implies that the MAP\(^2\) is expressed in the first canonical form. On the other hand, the notation \(\hat{D}^{(1)}_0(0)\) and \(\hat{D}^{(1)}_1(0)\) in (10) refers to the initial solution to the ML problem (see Section 4). If instead, a sample of size \(n = 1000\) is considered, the empirical moments,
\[
(\bar{\rho}_1, \bar{\mu}_1, \bar{\mu}_2, \bar{\mu}_3) = (0.0804, 1.6877, 6.7973, 43.0030),
\]
are closer to the theoretical ones, and the estimate is given by
\[
\hat{D}^{(1)}_0(0) = \begin{pmatrix} -2.1562 & 0.6346 \\ 0 & -0.4679 \end{pmatrix}, \quad \hat{D}^{(1)}_1(0) = \begin{pmatrix} 1.5216 & 0 \\ 0.0852 & 0.3827 \end{pmatrix},
\]
whose moments are
\[
(\hat{\rho}_1, \hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3) = (0.0804, 1.6877, 6.7973, 43.0034).
\]
It is interesting to note that, despite the estimated moments are close to the empirical and theoretical values, the elements of the matrices \(\{\hat{D}^{(1)}_0(0), \hat{D}^{(1)}_1(0)\}\) for \(n = 500\) and \(n = 1000\) differ pretty much from those of the theoretical \(\{D_0, D_1\}\) (with exception of parameter \(u\)). If instead, the empirical moments in the objective function are replaced by the real, theoretical ones, then the estimated matrices become
\[
\hat{D}^{(1)}_0(0) = \begin{pmatrix} -21.9163 & 6.5879 \\ 0 & -0.5001 \end{pmatrix}, \quad \hat{D}^{(1)}_1(0) = \begin{pmatrix} 15.3284 & 0 \\ 0.0425 & 0.4576 \end{pmatrix},
\]
more similar to the theoretical \(\{D_0, D_1\}\).

The empirical KL divergences give us a more informative image on how far the estimated processes are from the original one:
\[
D_{KL}\left(\{D_0, D_1\}, \{\hat{D}^{(1)}_0(0), \hat{D}^{(1)}_1(0)\}\right) = 46.0405 \ (n = 500),
\]
\[
D_{KL}\left(\{D_0, D_1\}, \{\hat{D}^{(1)}_0(0), \hat{D}^{(1)}_1(0)\}\right) = 9.6855 \ (n = 1000),
\]
\[
D_{KL}\left(\{D_0, D_1\}, \{\hat{D}^{(1)}_0(0), \hat{D}^{(1)}_1(0)\}\right) = 0.0430 \ (\text{theoretical moments}).
\]
From the above results, we can assert that estimate in the case where the empirical moments are exactly the theoretical ones is closer to \( D_0, D_1 \), than the estimate when \( n = 1000 \), which is closer to \( D_0, D_1 \) than the estimate in the case that \( n = 500 \). However, since the DK divergence is not upper bounded, the value 46.0405 is not conclusive enough of how similar \( D_0, D_1 \) and its estimate are. In order to get a clearer idea of this, a random different MAP2 from (8) was simulated

\[
\begin{align*}
D_0^\ast &= \begin{pmatrix}
-1 & 0.001 \\
0 & -0.005 \\
\end{pmatrix}, \\
D_1^\ast &= \begin{pmatrix}
0.999 \\
10^{-5} \\
0 \\
-10^{-5} + 0.005 \\
\end{pmatrix}, \\
\end{align*}
\]

with theoretical moments

\[
(\hat{\rho}(1), \hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3) = (0.3963, 67.3783, 2.6686 \times 10^4, 1.6011 \times 10^7). \\
\]

Then, we obtained

\[
D_{KL}(\{D_0, D_1\}, \{D_0^\ast, D_1^\ast\}) = 74.9794,
\]

which is clearly larger than the divergences in (11).

Although the previous results are preliminary, they shed some light on the complexity when comparing two given MAP2 representations. Since the topic exceeds the scope of this paper we do not look into it in greater depth and aim to address it in the future. □

4 Maximum likelihood estimate

In this section we look closely at the problem of estimating the parameters in the MAP2 by maximizing the likelihood function, given by (4). We will make use of the canonical representation of the process, and this way we avoid the typical switching problems of nonidentifiability. Specifically, given a sequence of interarrival times \( t = (t_1, t_2, \ldots, t_n) \) we aim to solve the following
optimization problem, concerning the first canonical form:

\[
\begin{align*}
\max & \quad \phi e^{D_0^1} D_1 e^{D_0^2} D_1 \ldots e^{D_0^n} D_1 e \\
\text{s.t.} & \quad D_0 = \begin{pmatrix} x & y \\ 0 & u \end{pmatrix}, \\
& \quad D_1 = \begin{pmatrix} -x - y & 0 \\ v & -u - v \end{pmatrix}, \\
& \quad x, u \leq 0, \\
& \quad y, v \geq 0, \\
& \quad x + y \leq 0, \\
& \quad u + v \leq 0, \\
& \quad \phi(-D_0)^{-1} D_1 = \phi.
\end{align*}
\]

With regard to the second canonical form, we formulate \((P2)\) as \((P1)\), where matrices \(D_0\) and \(D_1\) are given by \([6]\). To obtain the MAP2 estimate, we proceed as follows. First, the solutions to \((P1)\) and \((P2)\), \(\{\hat{D}_0^{(1)}, \hat{D}_1^{(1)}\}\) and \(\{\hat{D}_0^{(2)}, \hat{D}_1^{(2)}\}\) are computed. Finally, the selected estimate will be the MAP2 \(\{\hat{D}_0^{(1)}, \hat{D}_1^{(1)}\}\) or \(\{\hat{D}_0^{(2)}, \hat{D}_1^{(2)}\}\) that maximizes the likelihood.

Textbook models usually simplify maximum likelihood estimation problems by taking logs, and then simplifying the objective, which is given as a summation of \(n\) terms. This is not possible in our model: the objective function \((4)\) does not admit such a factorization due to the fact that the interarrival times are not independent, and thus the joint density is not expressed as the product of marginal likelihoods. This makes even the evaluation of the objective cumbersome. Other technical difficulties also appear. These, as well as ways to overcome such difficulties, are discussed in what follows.

4.1 Finding a starting solution

The choice of a good starting solution is always crucial to attain convergence of the ML algorithm to a good estimate. This is particularly relevant in our case, since an inadequate choice of the parameters may lead the algorithm to diverge, or even to be unable to provide an output, because of the presence of too big numbers.

We have found that a good starting point is obtained if one uses the moments matching estimate. The procedure to derive it is described below. The canonical representation of the MAP2 in terms of four parameters leads Bodrog et al. (2008) to show that any MAP2 is completely characterized by
its first three moments, $\mu_1$, $\mu_2$, $\mu_3$ and lag-one autocorrelation coefficient $\rho_1$.

As a consequence, given a sequence of interarrival times $t = (t_1, t_2, \ldots, t_n)$ with sample values $\bar{\mu}_i$, for $i = 1, 2, 3$ and $\bar{\rho}(1)$, the method of moments would allow one to estimate the parameters $(x, y, u, v)$ in the canonical form of the $MAP_2$ by solving the nonlinear system of equations

$$\begin{align*}
\mu_i(x, y, u, v) & = \bar{\mu}_i, \text{ for } i = 1, 2, 3, \\
\rho_1(x, y, u, v), & = \bar{\rho}_1.
\end{align*}$$

(14)

However, in real-world data, (14) may have no feasible solution. In order to obtain an estimate, we seek instead the parameters $(x, y, u, v)$ fulfilling as much as possible (14). Given $\tau > 0$, define the function

$$\delta_\tau(x, y, u, v) = \left\{ \rho_1(x, y, u, v) - \bar{\rho}_1 \right\}^2 + \tau \left\{ \left( \frac{\mu_1(x, y, u, v) - \bar{\mu}_1}{\bar{\mu}_1} \right)^2 + \left( \frac{\mu_2(x, y, u, v) - \bar{\mu}_2}{\bar{\mu}_2} \right)^2 + \left( \frac{\mu_3(x, y, u, v) - \bar{\mu}_3}{\bar{\mu}_3} \right)^2 \right\}.$$  

We propose to solve the following optimization problem:

$$(P0) \begin{cases}
\min & \delta_\tau(x, y, u, v) \\
\text{s.t.} & x, u \leq 0, \\
& y, v \geq 0, \\
& x + y \leq 0, \\
& u + v \leq 0.
\end{cases}$$

The penalty parameter $\tau$ needs to be tuned. In our experiments it has been set to $\tau = 1$, which seems to perform well in practice. Obviously $(x, y, u, v)$ solves (14) iff it is an optimal solution of (P0), whose optimal value is 0.

In order to solve the multimodal Problem (P0), we have used the MATLAB© routine fmincon. Numerical inaccuracies were found, and then the range of the parameters was slightly reduced, by adding to (P0) the constraints

$$\begin{align*}
x, u & \in [-1000, -2 \times 10^{-16}] \\
y, v & \in [0.00001, 100].
\end{align*}$$

A multistart was then executed with 100 randomly chosen starting points and found to yield satisfactory results. The solution to (P0), noted $\{\hat{D}_0(0), \hat{D}_1(0)\}$ will be used as starting point of the algorithm that maximizes the likelihood function.

It is worth pointing out here that other initial values could have been chosen, for example random starting $MAP_2$s; however we have found that
the use of the moments matching estimate reduces the numerical problems in
practice. A total of one thousand random MAPs were estimated via the ML
method described in Section 4.2 where the starting values were (1) randomly
generated versus (2) the moments matching estimates. In the first case, in a
32% of the generated MAP, the solution given by the computer possessed
a likelihood function equal to 0 or to infinite. This percentage decreased to
14% in the case of the moments matching estimates. When the objective
function was evaluated using the final ML estimates, a 35% of times it was
equal to 0 or to infinite in the first case (that is, when a random seed was
selected), against a 1% when the moments matching estimate was used as
starting value. Additionally, for those cases where the objective function did
not present any numerical inconsistency using a random starting point, the
61.53% of times the objective function was larger using a moments method
estimate as starting point than when a random MAP was used.

4.2 Evaluation of the likelihood function
In principle, (P1)-(P2) can be solved using standard optimization routines,
and, as discussed above, the moments method estimate, obtained solving
(P0) with a multistart, is a recommended starting point.

However we have found serious difficulties in carrying out the numerical
evaluation of the likelihood function (4), which turns out problematic in
practice when the variability in the sample \( t \) is large. This section is devoted
to analyze such a problem.

As a motivational example, consider the MAP given by (12). Note that
the theoretical variance of the interarrival times is \( 2.2146 \times 10^4 \). A sample
of 500 observations was generated from this MAP with a sample variance
equal to \( 3.4521 \times 10^4 \). That is why some extreme values, of the order of
10^3 were obtained. When evaluating the likelihood (4), it was found that
\( f(t_1, \ldots, t_n|D_0, D_1) \approx 0 \). An explanation for this phenomenon is as follows.
Given a MAP with canonical form as in (6), the term \( e^D_1 \) in (4) satisfies
\[
e^{D_1} = \begin{pmatrix}
-x - y & e^{tx} + y \frac{e^{tx} - e^{tu}}{(x - u)v} \\
 & \frac{e^{tx} - e^{tu}}{(x - u)(-u - v)} \\
 & e^{tu} \\
\end{pmatrix}
\]
Since \( x, u < 0 \), it follows immediately that
\[
\lim_{t \to \infty} e^{D_1} = 0,
\]
no matter which values the parameters \( x, u, y, v \) take. Here \( 0 \) denotes a \( 2 \times 2 
zero matrix. The same phenomenon happens when the second canonical form
is considered. This result implies that, in practice, in the presence of large interarrival times, the numerical evaluation of (4) is rather difficult. For instance, in the considered sample, $t_1 = 18.12$, $t_2 = 465.49$, $t_3 = 120.70$ and
\[
\begin{pmatrix}
  0.0000 & 0.0000 \\
  0.0000 & 0.0046
\end{pmatrix},
\]
\[
10^{-5} \times \begin{pmatrix}
  0.0000 & 0.0002 \\
  0.0004 & 0.2218
\end{pmatrix},
\]
\[
10^{-8} \times \begin{pmatrix}
  0.0000 & 0.0006 \\
  0.00012 & 0.6054
\end{pmatrix}.
\]
The factors $e^{D_0 t_1} D_1 \ldots e^{D_0 t_k} D_1$ become smaller as $k$ increases, and indeed the computer (MATLAB© software) returns $e^{D_0 t_1} D_1 \ldots e^{D_0 t_k} D_1 = 0$ for $k = 118$. This example is not an isolated case. Indeed, we experienced that it is more a rule than an exception that large interarrival times appear in the simulated samples. From Figure 1 which depicts the theoretical variance versus the mean of the inter-arrival times of 100,000 randomly simulated MAP$_2$s, it can be seen that the variance $V(T)$ increases considerably with the mean $E(T)$.

![Figure 1: $E(T)$ vs. $V(T)$ for 100000 simulated random MAP$_2$s.](image)

We have found rather convenient to re-scale the sample, thus re-scaling the likelihood function to a more tractable range. This is possible since the likelihood function (4) satisfies
\[
f(t|D_0, D_1) = c^{-n} f \left( \frac{1}{c} t | cD_0, cD_1 \right) \quad \forall c > 0,
\]
where $n$ is the length of $\mathbf{t}$. In other words, the ML estimates obtained for interarrival times $\mathbf{t}$ is a re-scaled by $c$ version of that obtained for interarrival times $\frac{1}{c}\mathbf{t}$, for any positive $c$. In our numerical experience we have found good results setting $c$ as the standard deviation of the data, so that the new sample variance is equal to 1 and therefore, less extreme values are expected to appear in the sample. Specifically, the algorithm to follow is:

1. Set $c := \text{std}(\mathbf{t})$, the standard deviation of $\mathbf{t} = (t_1, \ldots, t_n)$.
2. Consider the new sample $\mathbf{t}^* = \left(\frac{1}{c}\mathbf{t}\right)$.
3. Compute the ML estimates of $D_0^*$ and $D_1^*$, noted $\hat{D}_0^*$ and $\hat{D}_1^*$, by maximizing $f(\mathbf{t}^*|D_0^*, D_1^*)$ via a standard optimization algorithm.
4. Calculate the estimate of $D_0$ and $D_1$ as $\hat{D}_0 = \frac{1}{c}\hat{D}_0^*$ and $\hat{D}_1 = \frac{1}{c}\hat{D}_1^*$.

Next section illustrates the approach for a pair of simulated data sets.

### 4.3 Numerical illustration

**Example 2.** Consider the sequence of interarrival times, the MAP\(^2\) defined by (8) and its moments matching estimate (10) in Example 1. It can be checked that $f(\mathbf{t}|\hat{D}_0^{(1)}(0), \hat{D}_1^{(1)}(0)) \approx 0$. We set $c = \text{std}(\mathbf{t}) = 2.076$ and we compute $\mathbf{t}^* = \mathbf{t}/c$. Now, it can be seen that

$$\log f(\mathbf{t}^*|c\hat{D}_0^{(1)}(0), c\hat{D}_1^{(1)}(0)) = -431.3554.$$  

From \([15]\), it can be concluded that

$$\log f(\mathbf{t}|\hat{D}_0^{(1)}(0), \hat{D}_1^{(1)}(0)) = -500 \times \log(c) - 431.3554 = -796.5823.$$  

The MATLAB\(^\text{©}\) routine fmincon is used to obtain the solution to (P1), the ML estimates $\{\hat{D}_0^*, \hat{D}_1^*\}$. In this case, it was found that

$$\hat{D}_0^* = \begin{pmatrix} -30.7238 \\ 0 \end{pmatrix}, \quad \hat{D}_1^* = \begin{pmatrix} 6.7257 \\ -1.0069 \end{pmatrix}, \quad \hat{D}_0^* = \begin{pmatrix} 23.9981 \\ 0 \end{pmatrix}, \quad \hat{D}_1^* = \begin{pmatrix} 0.0735 \\ 0.9334 \end{pmatrix},$$
and then, dividing $\hat{D}_0^\star$ and $\hat{D}_1^\star$ by $c$ yields

$$
\hat{D}_0^{(1)} = \begin{pmatrix} -14.7994 & 3.2397 \\ 0 & -0.4850 \end{pmatrix}, \quad \hat{D}_1^{(1)} = \begin{pmatrix} 11.5596 & 0 \\ 0.0354 & 0.4496 \end{pmatrix},
$$

(16)

whose moments $\{\hat{\rho}, \hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3\}$ are obtained as

$$(\hat{\rho}, \hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3) = (0.1163, 1.6537, 6.7643, 41.8285).$$

In addition, the log-likelihood has increased with respect to the one provided by the moments matching estimate, i.e., the one obtained by solving (P0):

$$
\log f(t^*|\hat{D}_0^\star, \hat{D}_1^\star) = -248.5386,
$$

which implies that

$$
\log f(t|\hat{D}_0^{(1)}, \hat{D}_1^{(1)}) = -613.7655.
$$

(17)

There has been also an improvement in terms of the DK divergence:

$$
D_{KL}(\{D_0, D_1\}, \{\hat{D}_0^{(1)}, \hat{D}_1^{(1)}\}) = 0.7938,
$$

(18)

considerably smaller than 46.0405 in (11).

Next, we consider the estimate of the MAP2 in the second canonical form. The solution to (P0) is found

$$
\hat{D}_0^{(2)}(0) = \begin{pmatrix} -77.6722 & 23.4148 \\ 0 & -0.4861 \end{pmatrix}, \quad \hat{D}_1^{(2)}(0) = \begin{pmatrix} 0 & 54.2573 \\ 0.1406 & 0.3455 \end{pmatrix},
$$

with estimated moments

$$(\hat{\rho}(1), \hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3) = (-0.0287, 1.7144, 7.0451, 43.4798).$$

Note how the estimated moments are close to the empirical ones given by (9), with the exception of the autocorrelation coefficient, which in this case is negative. The algorithm to solve (P2) was implemented with starting solution given by $\{\hat{D}_0^{(2)}(0), \hat{D}_1^{(2)}(0)\}$ and yielded

$$
\hat{D}_0^{(2)} = \begin{pmatrix} -16.8292 & 6.4688 \\ 0 & -0.5343 \end{pmatrix}, \quad \hat{D}_1^{(2)} = \begin{pmatrix} 0 & 10.3606 \\ 0.1236 & 0.4107 \end{pmatrix},
$$

whose moments are

$$(\rho_1, \mu_1, \mu_2, \mu_3) = (-0.0146, 1.6505, 6.1523, 34.5420).$$

13
The KL divergence is
\[
D_{KL}\left(\{D_0, D_1\}, \{\hat{D}_0^{(2)}, \hat{D}_1^{(2)}\}\right) = 10.0508,
\]
larger than (18). Finally, the log-likelihood function is
\[
\log f\left(t|\hat{D}_0^{(2)}, \hat{D}_1^{(2)}\right) = -707.3972. \tag{19}
\]

To select the final estimate, the log-likelihoods (17) and (21) are compared. In this case the estimate of the MAP2 in its first form is chosen. □

Example 3. In this example, a MAP2 with a large variance of the interarrival times is estimated. Consider the MAP2 defined by (12), whose theoretical moments are given by (13). Note that the variance of the interarrival time is \(2.2146 \times 10^4\). Let \(t = (t_1, \ldots, t_n)\) be a sample of size \(n = 500\) of interarrival times simulated from (12) whose sample moments are
\[
(\bar{\rho}_1, \bar{\mu}_1, \bar{\mu}_2, \bar{\mu}_3) = (0.0709, 180.6187, 6.9675 \times 10^4, 3.8681 \times 10^7).
\]

In this case, \(c = \text{std}(t) = 192.6803\). The estimate obtained by solving (P0) in the first canonical form is given by
\[
\hat{D}_0^{(1)}(0) = \begin{pmatrix}
-122.6681 & 24.7206 \\
0 & -0.0052
\end{pmatrix}, \quad \hat{D}_1^{(1)}(0) = \begin{pmatrix}
97.9475 & 0 \\
0.0001 & 0.0051
\end{pmatrix}.
\]

The moments of the MAP2 defined by \(\{\hat{D}_0^{(1)}(0), \hat{D}_1^{(1)}(0)\}\) are
\[
(\hat{\rho}_1, \hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3) = (0.0516, 170.2164, 6.8348 \times 10^4, 3.9318 \times 10^7).
\]

As in Example 2, it can be checked that \(f\left(t|\hat{D}_0^{(1)}(0), \hat{D}_1^{(1)}(0)\right) \approx 0\). However,
\[
\log f\left(t^*|c\hat{D}_0^{(1)}(0), c\hat{D}_1^{(1)}(0)\right) = -466.9192,
\]
which, from (15), implies
\[
\log f\left(t|\hat{D}_0^{(1)}(0), \hat{D}_1^{(1)}(0)\right) = -3097.4.
\]

The solution to (P1) was
\[
\hat{D}_0^* = \begin{pmatrix}
-200.0788 \\
0
\end{pmatrix}, \quad \hat{D}_1^* = \begin{pmatrix}
198.7370 & 0 \\
0.0019 & 0.9925
\end{pmatrix}.
\]
and then, dividing $\hat{D}_0$ and $\hat{D}_1$ by $c$ leads to

\[
\hat{D}_0^{(1)} = \begin{pmatrix} -1.0384 & 0.0070 \\ 0 & 0.0052 \end{pmatrix}, \quad \hat{D}_1^{(1)} = \begin{pmatrix} 1.0314 & 0 \\ 0.0000 & 0.0052 \end{pmatrix}.
\]

It can be seen that the log-likelihood function has increased to

\[
\log f (t^*|\hat{D}_0^{(1)}, \hat{D}_1^{(1)}) = -392.8464,
\]

or equivalently,

\[
\log f (t|\hat{D}_0^{(1)}, \hat{D}_1^{(1)}) = -3023.4.
\]

The estimated moments are

\[
(\hat{\rho}_1, \hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3) = (0.1765, 151.6006, 5.8668 \times 10^4, 3.4103 \times 10^7).
\]

Finally, the DK divergence with respect to the estimates are

\[
D_{KL} (\{D_0, D_1\}, \{\hat{D}_0^{(1)}, \hat{D}_1^{(1)}\}) = 0.6973, \quad (20)
\]

much smaller than that obtained from (P0):

\[
D_{KL} (\{D_0, D_1\}, \{\hat{D}_0^{(1)(0)}, \hat{D}_1^{(1)(0)}\}) = 151.6021.
\]

The solution to (P0), in second canonical form was found as

\[
\hat{D}_0^{(2)(0)} = \begin{pmatrix} -0.0052 & 0.0052 \\ 0 & -142.8445 \end{pmatrix}, \quad \hat{D}_1^{(2)(0)} = \begin{pmatrix} 0 & 0 \\ 137.4806 & 5.3639 \end{pmatrix},
\]

with estimated moments

\[
(\hat{\rho}_1, \hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3) = (0, 181.8407, 6.8710 \times 10^4, 3.8943 \times 10^7).
\]

Note the incapability of the estimate to capture the strictly positive lag-one autocorrelation coefficient. Then, the solution to (P2), where \{\hat{D}_0^{(2)(0)}, \hat{D}_1^{(2)(0)}\} is used as starting solution is given by

\[
\hat{D}_0^{(2)} = \begin{pmatrix} -0.0052 & 0.0052 \\ 0 & -1.3115 \end{pmatrix}, \quad \hat{D}_1^{(2)} = \begin{pmatrix} 0 & 0 \\ 1.2240 & 0.0875 \end{pmatrix},
\]

with moments

\[
(\hat{\rho}_1, \hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3) = (0, 181.0040, 6.99896 \times 10^4, 4.0497 \times 10^7).
\]

The KL divergence is

\[
D_{KL} (\{D_0, D_1\}, \{\hat{D}_0^{(2)}, \hat{D}_1^{(2)}\}) = 84.0645,
\]

clearly larger than (20). Finally, the log-likelihood function is

\[
\log f (t|\hat{D}_0^{(2)}, \hat{D}_1^{(2)}) = -3051.7, \quad (21)
\]

which is smaller than $-3023.4$, therefore the estimate in first canonical form is selected. □
4.4 Canonical versus redundant representation

The $MAP_2$ can be expressed via either the redundant representation (1) or the canonical forms (6) or (7). In principle, the only difference between the two representations is that the canonical one allows for a unique estimate of the model parameters, while the lack of identifiability of representation (1) implies possibly infinite estimates. However, the elements of interest associated with the $MAP_2$, namely, the distributional properties of the variable $T$, are the same under equivalent representations. Also, if the interest is in the estimation of the $MAP_2/G/1$ queueing system, Ramírez-Cobo et al. (2012) recently proved that the steady-state distributions coincide under equivalent arrival processes. Therefore, it is natural to wonder which are the benefits of using the estimates in canonical representation instead of the redundant ones.

To look more closely at this problem a hundred of random $MAP_2$s in redundant representation were simulated and estimated via a ML approach equivalent to that described in Section 4.2, where the objective function is written in terms of the redundant variables $\{\lambda_1, \lambda_2, p_{120}, p_{110}, p_{210}, p_{211}\}$. Here too the starting point was calculated as the solution of the equivalent problem to (P0), where the moments are expressed in terms of the 6 variables. Once the estimates were obtained, the DK divergences between the real parameters and the estimated ones in redundant version, were calculated. On the other hand, the canonical estimates of the random $MAP_2$s and their DK divergences were computed using the ML method of Section 4.2. Figure 2 depicts the histogram of the ratio between the DK divergences of the redundant over the canonical estimates. It can be seen that the DK divergence of the redundant forms are considerable larger than those from the canonical versions and in consequence, the canonical estimates are closer to the true parameters than the redundant ones.

It should be also pointed out that in eighteen out of the hundred of simulated $MAP_2$s, it was not possible to obtain the ML estimate in redundant version. Apparently, the evaluation of the likelihood function in terms of six parameters presents more numerical problems than that in the canonical version, and in all these cases numerical inconsistencies were found.

5 Discussion

In this paper we deepen our understanding of the maximum likelihood estimation of the second-order $MAP$, a suitable stochastic process for many statistical modeling applications. Despite the apparent straightforwardness
of the problem, the matrix notation as well as the intrinsic dependence structure of the process turn the evaluation and maximization of the likelihood function into a complicated task in practice. These difficulties are overcome by the use of the canonical representation of the process, a proper re-scaling of the objective function and a choice of a particular starting solution of the algorithm. A method to compare between different estimates is also delineated.

Prospects regarding this work may concern inference for higher order MAP, which are expected to show more versatility for modeling purposes. We are aware of the complexity of such a problem due to the lack of unique representations and the increasing number of parameters. These complications present a challenging problem that we hope to address in the future.

In the spirit of a reproducible research the codes utilized in this paper to estimate the $MAP_2$ are available at

http://personal.us.es/jrcobo/www/Software.html

as a stand-alone MATLAB® toolbox.

**Acknowledgements**

Research partially supported by research grants and projects MTM2009-14039 (Ministerio de Ciencia e Innovación, Spain) and FQM329 (Junta de
Andalucía, Spain), both with EU ERDF funds. The corresponding author is supported by Consolider ”Ingenio Mathematica” through her post-doc contract.

References

Aalen, O. (1995). Phase-type distributions in survival analysis. *Scandinavian journal of statistics, 22*, 145–157.

Asmussen, S., & Olsson, M. (1998). Phase-type distributions. In *In Kotz, S., Read, C.B. and Banks, D.L., editors, Encyclopedia of Statistical Science Update, 2* (pp. 525–530).

Badescu, A., Drekic, S., & Landriault, D. (2007). Analysis of a threshold dividend strategy for a MAP risk model. *Scandinavian Actuarial journal, 4*, 227–247.

Bodrog, L., Heindlb, A., Horvátha, G., & Telek, M. (2008). A Markovian canonical form of second-order matrix-exponential processes. *European Journal of Operational Research, 190*, 459–477.

Breuer, L. (2002). An EM algorithm for batch Markovian arrival processes and its comparison to a simpler estimation procedure. *Ann. Operations Research, 112*, 123–138.

Casale, G., Z. Zhang, E., & Simirni, E. (2010). Trace data characterization and fitting for Markov modeling. *Performance Evaluation, 67*, 61–79.

Cheung, E., & Landriault, D. (2010). A generalized penalty function with the maximum surplus prior to ruin in a MAP risk model. *Insurance: Mathematics and Economics, 46*, 127–134.

Ephraim, Y., & Merhav, N. (2002). Hidden Markov Processes. *IEEE Transactions on information theory, 48*, 1518–1569.

Eum, S., Harris, R., & Atov, I. (2007). A matching model for MAP-2 using moments of the counting process. In *Proceedings of the International Network Optimization Conference, INOC 2007*. Spa, Belgium.

Horváth, M., & Telek, M. (2002). Markovian modeling of real data traffic: Heuristic phase type and MAP fitting of heavy tailed and fractal like samples. In *Performance evaluation of complex systems: Techniques and Tools, IFIP Performance 2002, in: LNCS Tutorial Series, vol. 2459* (pp. 405–434).
Kim, B., & Kim, J. (2010). Queue size distribution in a discrete-time D-BMAP/G/1 retrial queue. *Computers and Operations research, 37*, 1220–1227.

Klemm, A., Lindemann, C., & Lohmann, M. (2003). Modeling IP traffic using Batch Markovian Arrival Process. *Perform. Eval., 54*, 149–173.

Lucantoni, D. (1993). The BMAP/G/1 queue: A tutorial. In L. Donatiello, & R. Nelson (Eds.), *Models and Techniques for Performance Evaluation of Computer and Communication Systems* (pp. 330–358). New York: Springer.

Lucantoni, D., Meier-Hellstern, K., & Neuts, M. (1990). A single-server queue with server vacations and a class of nonrenewal arrival processes. *Advances in Applied Probability, 22*, 676–705.

Montoro-Cazorla, D., Pérez-Ocón, R., & Segovia, M. (2009). Replacement policy in a system under shocks following a Markovian arrival process. *Reliability Engineering and System Safety, 94*, 497–502.

Neuts, M. F. (1979). A versatile markovian point process. *Journal of Applied Probability, 16*, 764–779.

O’Cinneide, C. (1989). On non-uniqueness of representations of phase-type distributions. *Stochastic Models, 5*, 247–259.

Okamura, H., Dohi, T., & Trivedi, K. (2009). Markovian arrival process parameter estimation with group data. *IEEE/ACM Trans. Networking, 17*, 1326–1339.

Ramírez-Cobo, P., & Lillo, R. (2012). New results about weakly equivalent MAP_{2} and MAP_{3} processes. *Methodology and Computing in Applied Probability, doi: 10.1007/s11009-011-9227-x*.

Ramírez-Cobo, P., Lillo, R., & Wiper, M. (2010). Nonidentifiability of the two-state Markovian arrival process. *Journal of applied probability, 47*, 630–649.

Ramírez-Cobo, P., Lillo, R., & Wiper, M. (2012). Identifiability of the MAP_{2}/G/1 queueing system. To appear in *TOP*.

Telek, M., & Horváth, G. (2007). A minimal representation of markov arrival processes and a moments matching method. *Performance evaluation, 64*, 1153–1168.
Wu, J., Z., L., & Yang, G. (2011). Analysis of the finite source MAP/PH/N retrial G-queue operating in a random environment. *Applied Mathematical Modelling, 35*, 1184–1193.