drimmR : R Package for Estimation, Simulation and Reliability of Drifting Markov Models

Vlad Stefan Barbu
University of Rouen

Geoffray Brelurut
University of Rouen

Annthomy Gilles
University of Rouen

Arnaud Lefebvre
University of Rouen

Victor Mataigne
University of Rouen

Alexandre Seiller
University of Rouen

Nicolas Vergne
University of Rouen

Abstract

The aim of this paper is to present the R package drimmR devoted to the estimation, simulation and associated reliability measures of drifting Markov models (DMMs). These are particular non-homogeneous Markov chains introduced in **?** and further developed in **?**, for which the Markov transition matrix is a linear/polynomial function of two/several Markov transition matrices. Several statistical frameworks are taken into account (one or several samples, complete or incomplete samples, models of the same length or not) and two types of estimations are proposed when starting from several samples. We also compute the probabilities and expectations of appearance of a word along a sequence and the p-value of a word occurrence under a given model.

Keywords: Markov models, Drifting Markov models, Non-parametric estimation, Simulation, Reliability, AIC, BIC.

1. Introduction

In this work we focus on multi-state systems modeled by means of a particular class of non-homogeneous Markov processes introduced in **?**, called drifting Markov processes. Most of the estimation methods and reliability results are developed in **?**.

Note that in many mathematical models it is assumed the homogeneity with respect to time, which is inappropriate in most of the applications. But, considering general non-homogeneous processes could be unrealistic from a practical point of view. For this reason, the drifting Markov chains introduced in **?** assume that the Markov transition matrix is a linear/polynomial function of two/several Markov transition matrices. Thus we obtain a "smooth" non-homogeneity, with sensibly less parameters than in the general case.

Few R packages have been developed to handle similar types of models of Markov, semi-Markov or hidden semi-Markov type, useful in reliability or DNA analysis. For semi-Markov models we have `semiMarkov` `R` package **?** that performs maximum likelihood estimation for parametric continuous-time semi-Markov processes, `smm` `R` package (**?**) which performs parametric and non-parametric estimation and simulation for multi-state discrete-time semi-Markov processes. Two `R` packages are also dedicated to hidden semi-Markov models, imple-

menting estimation and prediction methods: the `hsmm` `R` package **?** and the `mhsmm` `R` package **?**.

Note that there is no `R` package developed for drifting Markov models (DMMs). Thus the purpose of this paper is to present an `R` package that we have developed, called `DRIMM'R`, which performs estimation and simulation for such models, as well as the estimation of associated reliability measures. The aim of this paper is to describe the different possibilities of this package. To summarize, the package `DRIMM'R` that we present deals with different problems:

- We consider one or several sample paths; for several sample paths, two estimation methods are proposed: one is the usual LSE, the other one is the average of LSEs obtained on each sample;
- The samples paths are complete or incomplete;
- The sample paths come from drifting Markov chains that are of the same length or of different lengths (between the same Markov transition probability matrices);
- We derive exact computations of reliability/survival analysis measures (reliability or survival function, availability, maintainability, failure rates);
- We compute the probabilities and expectations of appearance of a word in a given sequence; we also compute the p-value of a word occurrence under a given model.

We would like to mention that a web interface called `WebDRIMM` has been also developed (cf. **?**) for simulating and estimating drifting Markov models, as well as associated reliability indicators; it is available at `http://bioinfo.univ-rouen.fr/WebDRIMM`

The paper is organized as follows. Section 2 describes the drifting Markov models used in this package, present associated reliability indicators and corresponding estimation results and techniques. Section 3 illustrates the different functions of the `DRIMM'R` package and we end the paper by presenting some concluding remarks on this `R` package in Section **??**.

2. Drifting Markov models, estimation and associated reliability measures

Let us consider a random system with finite state space $E = \{1, \ldots, s\}$, $s < \infty$. We assume that the time evolution of a system is governed by a discrete-time stochastic process with values in E. In the following we will introduce a class of processes according to **?** in Section 2.1, we will briefly present associated reliability indicators according to **?** in Section 2.2 and estimation of the parameters in Section 2.3, according to **??**.

2.1. Drifting Markov models

Let $\Pi_0 = (\Pi_0(u,v))_{u,v\in E}$ and $\Pi_1 = (\Pi_1(u,v))_{u,v\in E}$ be two Markov transition matrices of order 1 over the state space E.

Definition 1 (linear drifting Markov chain of order 1 and of length n) *A sequence X_0, $X_1, \ldots, X_n$ with state space $E = \{1, 2, \ldots, s\}$ is said to be a* linear drifting Markov chain (of

order 1) *of length n between the Markov transition matrices* Π_0 *and* Π_1 *if the distribution of* X_t, $t = 1, \ldots, n$, *is defined by*

$$\mathbb{P}(X_t = v \mid X_{t-1} = u, X_{t-2}, \ldots) = \Pi_{\frac{t}{n}}(u, v),\ u, v \in E, \tag{1}$$

where

$$\Pi_{\frac{t}{n}}(u, v) = \left(1 - \frac{t}{n}\right) \Pi_0(u, v) + \frac{t}{n} \Pi_1(u, v),\ u, v \in E. \tag{2}$$

Let us denote by $\alpha = (\alpha(1), \ldots, \alpha(s))$ the *initial distribution of the chain*, that is the distribution of X_0, $\alpha(u) = \mathbb{P}(X_0 = u)$ for any state $u \in E$.

The *linear drifting Markov model* of order 1 can be generalized to *polynomial drifting Markov model* of order k and degree d. Let $\Pi_{\frac{i}{d}} = (\Pi_{\frac{i}{d}}(u_1, \ldots, u_k, v))_{u_1, \ldots, u_k, v \in E}$ be d Markov transition matrices (of order k) over a state space E.

Definition 2 (polynomial drifting Markov chain of order k and of length n) *A sequence* X_0, X_1, …, X_n *with state space* $E = \{1, 2, \ldots, s\}$ *is said to be a* polynomial drifting Markov chain of order k *and of length n if the distribution of* X_t, $t = 1, \ldots, n$, *is defined by*

$$\mathbb{P}(X_t = v \mid X_{t-1} = u_k, X_{t-2} = u_{k-1}, \ldots) = \Pi_{\frac{t}{n}}(u_1, \ldots, u_k, v),\ u_1, \ldots, u_k, v \in E, \tag{3}$$

where

$$\Pi_{\frac{t}{n}}(u_1, \ldots, u_k, v) = \sum_{i=0}^{d} A_i(t) \Pi_{\frac{i}{d}}(u_1, \ldots, u_k, v),\ u_1, \ldots, u_k, v \in E, \tag{4}$$

with A_i *polynomials of degree* d *such as, for any* $i, j \in \{0, 1, \ldots, d\}$, $A_i(\frac{nj}{d}) =_{\{i=j\}}$.

We would like to stress that the coherence between notations implies the choice of the notation $\Pi_{\frac{i}{d}}$ and that, in fact, A_i are Lagrange polynomials; see **?** for more details on these two points.

2.2. Reliability of drifting Markov models

In order to undertake a reliability analysis of a system modeled by a DMM, let us assume that the state space of the system is partitioned into working and failure states, $E = U \cup D$, with $U \cap D = \emptyset$, where $U = \{1, \ldots, s_1\}$ represents the working states and $D = \{s_1 + 1, \ldots, s\}$ the failure states of the system. According to this partition of the state space we partition any matrix of vector we are working with and we denote the corresponding partitions accordingly (e.g., Π_0^{UU}, Π_0^{DU}, α^U etc.).

For a linear drifting Markov chain of order 1 $(X_t)_{0 \leq t \leq n}$, the reliability at time l, $l \in \mathbb{N}$, is given by

$$R(l) = \alpha^U \prod_{t=1}^{l} \left(\left(1 - \frac{t}{n}\right) \Pi_0^{UU} + \frac{t}{n} \Pi_1^{UU} \right) 1^U, \text{ where } 1^U = (\underbrace{1, \cdots, 1}_{s_1})^\top. \tag{5}$$

For a linear drifting Markov chain of order 1 $(X_t)_{0 \leq t \leq n}$, the pointwise (or instantaneous) availability at time l, $l \in \mathbb{N}$, is given by

$$A(l) = \alpha \prod_{t=1}^{l} \left(\left(1 - \frac{t}{n}\right) \Pi_0 + \frac{t}{n} \Pi_1 \right) 1^{E,U}, \text{ where } 1^{E,U} = (\underbrace{1, \cdots, 1}_{s_1}, \underbrace{0, \cdots, 0}_{s - s_1})^\top. \tag{6}$$

For a linear drifting Markov chain of order 1 $(X_t)_{0\leq t\leq n}$, the maintainability at time l, $l \in \mathbb{N}$, is given by

$$M(l) = 1 - \alpha^D \prod_{t=1}^{l} \left(\left(1 - \frac{t}{n} \right) \Pi_0^{DD} + \frac{t}{n} \Pi_1^{DD} \right) 1^D, \text{ where } 1^D = (\underbrace{1, \cdots, 1}_{s-s_1})^\top. \tag{7}$$

Similar formulas hold true for failure rates; see **?** for more details and **??** for similar questions in a discrete-time semi-Markov framework.

2.3. Estimation of drifting Markov models

In this section we will consider different types of data for which the estimators of the characteristics of a drifting Markov chain and of the associated reliability indicators will be derived.

One can observe one sample path, that will be denoted by $\mathcal{H}(m,n) := (X_0, X_1, \ldots, X_m)$, where m denotes the length of the sample path and n the length of the drifting Markov chain. Two cases can be considered: (a1) $m = n$ (a complete sample path); (a2) $m < n$ (an incomplete sample path).

One can also observe H i.i.d. sample paths, $\mathcal{H}_i(m_i, n_i), i = 1, \ldots, H$. Four cases can be considered here: (b1) $m_i = n_i = n$ for all $i = 1, \ldots, H$ (complete sample paths of drifting Markov chains of the same length) ; (b2) $n_i = n$ for all $i = 1, \ldots, H$ (incomplete sample paths of drifting Markov chains of the same length); (b3) $m_i = n_i$ for all $i = 1, \ldots, H$ (complete sample paths of drifting Markov chains of different lengths) ; (b4) $m_i \leq n_i$ for all $i = 1, \ldots, H$ (incomplete sample paths of drifting Markov chains of different lengths).

We have developed (cf. **??**) mean square estimators starting from data under these frameworks; we present here only the case of a linear drifting Markov chain of order 1 under the sample framework (b1) (complete sample paths of drifting Markov chains of the same length).

Under the setting (b1), starting with H complete sample paths of drifting Markov chains of the same length of a linear drifting Markov chain between two Markov transition matrices (of order 1) Π_0 and Π_1, for any states $u, v \in E$, the estimators of $\Pi_0(u,v)$ and $\Pi_1(u,v)$ are given by:

$$\widehat{\Pi}_{0;(n,H)}(u,v) = \frac{P_1(H,m,n)P_2(H,m,n) - P_3(H,m,n)P_4(H,m,n)}{P_5(H,m,n)P_1(H,m,n) - P_3(H,m,n)^2}$$
$$\widehat{\Pi}_{1;(n,H)}(u,v) = \frac{P_5(H,m,n)P_4(H,m,n) - P_3(H,m,n)P_2(H,m,n)}{P_5(H,m,n)P_1(H,m,n) - P_3(H,m,n)^2},$$

where we have introduced the following notation:

$$P_1(H,m,n) = \sum_{t=1}^{m}\sum_{h=1}^{H} {}_{\{X_{t-1}^h = u\}}(\frac{t}{n})^2,\ P_2(H,m,n) = \sum_{t=1}^{m}\sum_{h=1}^{H} {}_{\{X_{t-1}^h = u, X_t^h = v\}}(1 - \frac{t}{n}),$$
$$P_3(H,m,n) = \sum_{t=1}^{m}\sum_{h=1}^{H} {}_{\{X_{t-1}^h = u\}}(1 - \frac{t}{n})(\frac{t}{n}),\ P_4(H,m,n) = \sum_{t=1}^{m}\sum_{h=1}^{H} {}_{\{X_{t-1}^h = u, X_t^h = v\}}(\frac{t}{n}),$$
$$P_5(H,m,n) = \sum_{t=1}^{m}\sum_{h=1}^{H} {}_{\{X_{t-1}^h = u\}}(1 - \frac{t}{n})^2, \text{ with } m = (m_1, \ldots, m_H) \text{ and } n = (n_1, \ldots, n_H).$$

Note we can adapt the estimation procedures that we have previously obtained in order to get estimators of the drifting Markov models in the other cases; one can see **?** for more details. Using the expression of the reliability indicators of a drifting Markov chain previously obtained and the estimators of the characteristics of a drifting Markov chain, one immediately obtains the associated plug-in estimators of the reliability metrics.

3. The drimmR package

The drimmR package is principally devoted to the simulation and estimation of drifting Markov models, as well as to the estimation of associated reliability measures and to the computation of the probabilities, expectations and p-value of a word occurrence under a given model. All the different possibilities of the package are illustrated in Figure 1.

3.1. Estimation of drifting Markov models

The estimation of DMMs is carried out by the functions `dmmsum`, when starting from one (several) sample path (and obtain LSE). We will describe this function in the sequel.

1. The function `dmmsum`

The different **arguments** of this function are:

- `sequences`: A list of character vector(s) representing one (several) sequence(s) from which the estimation is carried out
- `order`: Order of the Markov chain
- `degree`: Degree of the polynomials (e.g., linear drifting if degree=1, etc.)
- `states`: Vector of states space of length s > 1
- `init.estim`: Method used to estimate the initial law.

Here we have an example of an estimation of a drifting Markov model of order 1 and degree 3 using the function `dmmsum`, starting from a DNA sequence called `lambda`.

```
data("lambda")
states <- c("a","c","g","t")
dmm <- dmmsum(lambda, 1, 3, states, init.estim="prod")
```

```
$states
[1] "a" "c" "g" "t"

$order
[1] 1

$degree
[1] 1

$Polynomials
    t^0 t^1
```

```
A_0   1  -1
A_1   0   1

$length
[1] 48502

$matrices
$matrices$Pi0
          a         c         g         t
a 0.2548330 0.2495270 0.2593907 0.2362493
c 0.2353305 0.2465529 0.3360221 0.1820945
g 0.2313610 0.3008693 0.2879023 0.1798674
t 0.1356776 0.2174431 0.4129533 0.2339260

$matrices$Pi1
          a         c         g         t
a 0.3398025 0.1715507 0.1870320 0.3016148
c 0.3339305 0.1911400 0.2075919 0.2673376
g 0.2801379 0.2600478 0.2021377 0.2576765
t 0.2218149 0.2283648 0.2305139 0.3193065


$init.estim
        a         c         g         t
0.2543400 0.2342171 0.2642778 0.2471651

attr(,"class")
[1] "dmm"    "dmmsum"
```

Estimation of corresponding model characteristics

Once a Markov drifting model has been estimated/constructed as described in the previous subsection, various characteristics of the model can be computed. These are: the log-likelihood of the sequence(s), the AIC and BIC information criteria, the stationary distribution on the entire sequence or only on a part of it, the distribution of the chain on the entire sequence or only on a part of it.

The functions `loglik`, `aic` and `bic` have the following **arguments**:

- `x`: Object of class `dmm`
- `sequences`: A list of character vector(s) representing one (several) sequence(s)

They return the numerical values of the log-likelihood, the AIC, the BIC of the sequence, respectively.

```
data(lambda, package = "drimmR")
sequence <- c("a","g","g","t","c","g","a","t","a","a","a")
dmm <-dmmsum(lambda, 1, 1, c('a','c','g','t'), init.estim = "freq")

loglik(dmm, sequence)
[[1]]
[1] -16.08783

aic(dmm, sequence)
[[1]]
[1] 56.17567

bic(dmm, sequence)
[[1]]
[1] 60.95041
```

The function `getTransitionMatrix` evaluates the transition matrix at a given position. The function has the following arguments :

- `x`: Object of class `dmm`
- `pos`: position along the sequence (integer)

```
data(lambda)
dmm <- dmmsum(lambda, 1, 1, c('a','c','g','t'),init.estim = "freq")
t <- 10
getTransitionMatrix(dmm,t)
```

```
          a         c         g         t
a 0.2548505 0.2495109 0.2593757 0.2362628
c 0.2353509 0.2465415 0.3359956 0.1821121
g 0.2313710 0.3008609 0.2878846 0.1798834
t 0.1356954 0.2174453 0.4129157 0.2339436
```

The function `getStationaryLaw` evaluates the stationary law at a given position or for all positions along the list of sequence(s).

- `x`: Object of class `dmm`
- `pos`: position along the sequence (integer)
- `all.pos`: FALSE (default, evaluation at pos index) ; TRUE (evaluation for all pos index)
- `internal`: FALSE (default) ; TRUE (for internal use of dmmsum initial law)

```
data(lambda)
sequence <- sample(lambda, 30, replace=TRUE )
dmm <- dmmsum(sequence, 1, 1, c('a','c','g','t'), init.estim = "freq")
t <- 10
getStationaryLaw(dmm,pos=t)
```

```
        a         c         g         t
0.2641443 0.2794671 0.2091486 0.2472401
```

```
getStationaryLaw(dmm,all.pos=TRUE)
```

```
                a         c         g         t
pos 1  0.2529924 0.2551165 0.3350840 0.1568071
pos 2  0.2497715 0.2615672 0.3302433 0.1584180
pos 3  0.2462983 0.2677077 0.3259821 0.1600119
pos 4  0.2426070 0.2735575 0.3222392 0.1615964
pos 5  0.2387271 0.2791341 0.3189611 0.1631778
pos 6  0.2346847 0.2844522 0.3161009 0.1647623
pos 7  0.2305025 0.2895249 0.3136173 0.1663552
pos 8  0.2262009 0.2943637 0.3114738 0.1679617
pos 9  0.2217974 0.2989783 0.3096378 0.1695864
pos 10 0.2173081 0.3033775 0.3080805 0.1712339
pos 11 0.2127468 0.3075688 0.3067761 0.1729082
pos 12 0.2081263 0.3115587 0.3057015 0.1746135
pos 13 0.2034578 0.3153529 0.3048358 0.1763536
pos 14 0.1987514 0.3189560 0.3041603 0.1781322
pos 15 0.1940163 0.3223724 0.3036581 0.1799531
pos 16 0.1892609 0.3256053 0.3033139 0.1818199
pos 17 0.1844925 0.3286576 0.3031137 0.1837362
pos 18 0.1797180 0.3315315 0.3030449 0.1857055
pos 19 0.1749438 0.3342288 0.3030958 0.1877316
pos 20 0.1701755 0.3367505 0.3032560 0.1898180
pos 21 0.1654185 0.3390975 0.3035156 0.1919684
pos 22 0.1606776 0.3412698 0.3038658 0.1941868
pos 23 0.1559573 0.3432675 0.3042984 0.1964769
pos 24 0.1512618 0.3450896 0.3048058 0.1988428
pos 25 0.1465952 0.3467353 0.3053808 0.2012887
pos 26 0.1419612 0.3482029 0.3060171 0.2038188
pos 27 0.1373632 0.3494905 0.3067086 0.2064377
pos 28 0.1328047 0.3505956 0.3074495 0.2091502
pos 29 0.1282888 0.3515155 0.3082347 0.2119610
pos 30 0.1238188 0.3522467 0.3090591 0.2148755
```

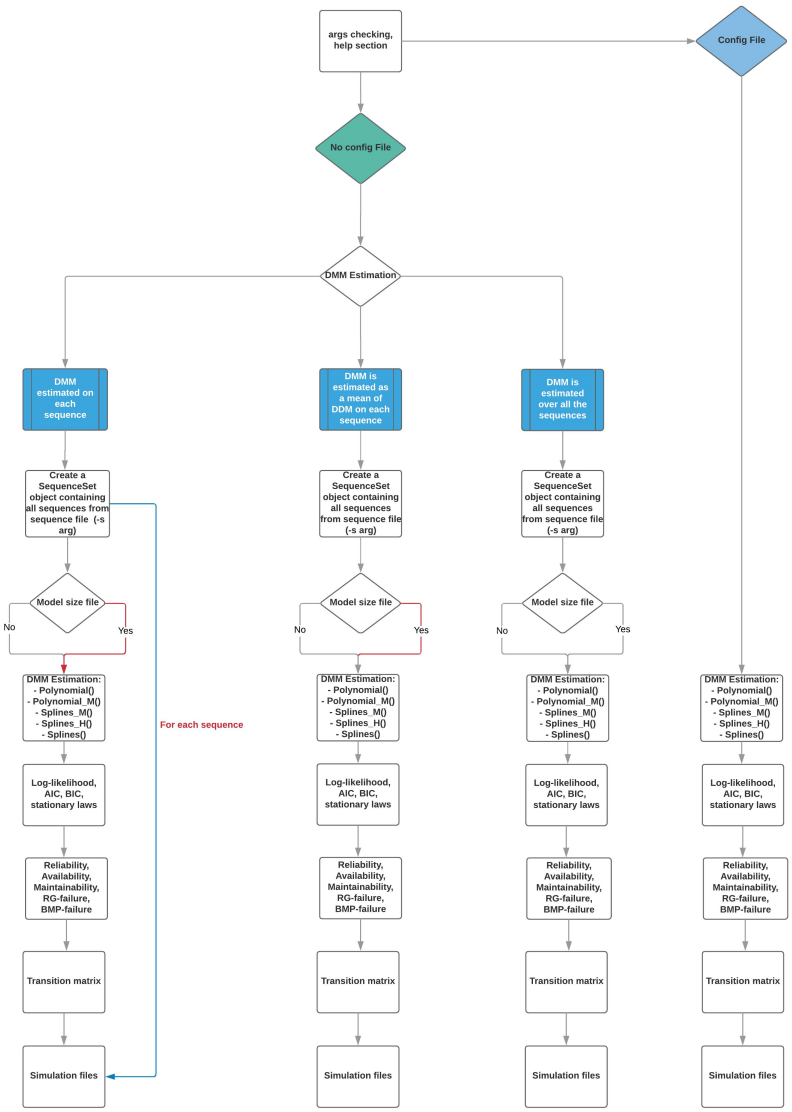

Figure 1: Schema of the `DRIMM'R package`.

Affiliation:

Firstname Lastname
Affiliation
Address, Country
E-mail: name@address
URL: http://link/to/webpage/