

dsmmR: Estimation and Simulation of Drifting Semi-Markov Models

Vlad Stefan Barbu¹, Ioannis Mavrogiannis¹, and Nicolas Vergne¹

¹ Univ Rouen Normandie, CNRS, LMRS UMR 6085, F-76000 Rouen, France ¶ Corresponding author

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

Software

- [Review](#)
- [Repository](#)
- [Archive](#)

Editor: Hugo Gruson

Reviewers:

- [@Athene-ai](#)
- [@RomainAzais](#)

Submitted: 22 January 2023

Published: unpublished

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).

Summary

Markov models are a common approach in the researcher's toolbox in order to deal with the modeling of many real-life phenomena presented as a sequence under discrete time. However, this imposes the assumption that the sequence under inspection is homogeneous with respect to time and furthermore that the sojourn times follow the Geometric distribution. This is not always true in practice when modeling, for example, DNA sequences.

Drifting semi-Markov models (DSMM) formulate the combination of semi-Markov models, which allow arbitrary distributions for the sojourn times, together with drifting Markov models, which describe the non-homogeneity of a sequence through a smooth, known shape that is gradually evolving, expressed through a linear (polynomial) function. As a result, DSMM are best suited to capture the non-homogeneity that occurs from the intrinsic evolution of the system or from the interactions between the system and the environment. For a detailed introduction to semi-Markov models see ([Barbu & Limnios, 2009](#)). Drifting Markov models were first introduced in ([Vergne, 2008](#)).

dsmmR is an R package which allows the user to perform parametric and non-parametric estimation and simulation of drifting semi-Markov processes. The user can also define their own parametric and non-parametric DSMM specifications, allowing for a necessary degree of freedom when dealing with a research question. Furthermore, three different types of DSMM are considered. These three models differ in the way they characterize the drifting semi-Markov kernel. Specifically, the first model allows both the transition matrix and the sojourn time distribution to vary (i.e. to "drift"), the second model allows only the transition matrix to drift while the third model allows only the sojourn time distribution to drift.

The main functions of dsmmR are the following:

- `fit_dsmm()`: estimates a DSMM given a sequence.
- `parametric_dsmm()`: defines a parametric DSMM.
- `nonparametric_dsmm()`: defines a non-parametric DSMM.
- `simulate()`: simulates a sequence.
- `get_kernel()`: returns the drifting semi-Markov kernel.

The estimation of the DSMM kernel is parametric or non-parametric and can be defined through the function `fit_dsmm()`, which returns an object of the S3 class (`fit_dsmm_parametric`, `dsmm`) or (`fit_dsmm_nonparametric`, `dsmm`). In the parametric estimation case, several discrete sojourn time distributions are considered for the sojourn times: Uniform, Geometric, Poisson, Discrete Weibull (of type 1) and Negative Binomial. The parametric DSMM specification can be defined through the function `parametric_dsmm()`, which returns an object of the S3 class (`dsmm_parametric`, `dsmm`). In the parametric definition case, we have the same discrete sojourn time distributions. The non-parametric DSMM specification can be defined through the function `nonparametric_dsmm()`, which returns an object of the S3 class (`dsmm_nonparametric`, `dsmm`). It allows for the sojourn time distributions to be of an arbitrary shape.

- 43 The `dsmm` class acts like a wrapper class, enabling the handling of all three models and
44 parametric or non-parametric cases for any degree when it comes to the following actions:
- 45 ▪ Simulating a sequence of states under a drifting semi-Markov kernel through the S3
46 method `simulate.dsmm()`.
 - 47 ▪ Calculating the drifting semi-Markov kernel through the generic function `get_kernel()`.

48 Statement of need

49 Drifting semi-Markov processes represent a versatile tool that can be applied in many fields of
50 science like reliability, survival analysis, bioinformatics, engineering, finance and more. The
51 present R package consists of a novel approach to tackle gradually evolving non-homogeneity
52 in a polynomial way and also allows for arbitrary sojourn time distributions. Instead, existing
53 R packages deal with semi-Markov models, hidden semi-Markov models and drifting Markov
54 models.

55 For semi-Markov models we have the R package `SemiMarkov` (Król & Saint-Pierre, 2015)
56 that performs maximum likelihood estimation for parametric continuous-time semi-Markov
57 processes, where the distribution can be chosen between Exponential, Weibull or exponentiated
58 Weibull. That package computes associated hazard rates; covariates can also be taken into
59 account through the Cox proportional hazard model. The R package `smmR` (Barbu, Lecocq, et
60 al., 2021) deals with discrete-time multi-state semi-Markov models, performing parametric and
61 non-parametric estimation and simulation, with k -th order Markov chains also considered. For
62 the parametric estimation the following sojourn times were considered: Uniform, Geometric,
63 Poisson, Discrete Weibull and Negative Binomial. One or more sample paths were considered,
64 with or without censoring at the beginning or the end of the sample paths. Few R packages are
65 also dedicated to hidden semi-Markov models, implementing estimation and prediction methods.
66 Among them, we can cite the HMM R package (Himmelman, 2022), the `HiddenMarkov` R
67 package (Harte, 2021) and the `mhsmm` R package (O'Connell & Højsgaard, 2011).

68 Furthermore, for drifting Markov models the R package `drimmR` (Barbu, Lothode, et al., 2021)
69 was developed, performing non-parametric estimation and simulation, also allowing for the
70 exact computation of the associated reliability measures and for several frameworks with
71 regards to how many samples were used, considering if they were complete or incomplete and
72 if they were of the same length. In this package, a linear (polynomial) function was used to
73 capture the non-homogeneity.

74 Acknowledgements

75 We acknowledge the DATALAB Project (financed by the European Union with the European
76 Regional Development fund (ERDF) and by the Normandy Region) and the HSMM-INCA
77 Project (financed by the French Agence Nationale de la Recherche (ANR) under grant ANR-
78 21-CE40-0005).

79 References

- 80 Barbu, V. S., Lecocq, F., Lothode, C., & Vergne, N. (2021). *smmR: Simulation, estimation
81 and reliability of semi-Markov models*. <https://CRAN.R-project.org/package=smmR>
- 82 Barbu, V. S., & Limnios, N. (2009). *Semi-Markov chains and hidden semi-Markov models
83 toward applications: Their use in reliability and DNA analysis* (Vol. 191). Springer Science
84 & Business Media.
- 85 Barbu, V. S., Lothode, C., & Vergne, N. (2021). *drimmR: Estimation, simulation and reliability
86 of drifting Markov models*. <https://CRAN.R-project.org/package=drimmR>

- 87 Harte, D. (2021). *HiddenMarkov: Hidden markov models*. Statistics Research Associates.
88 <https://www.statsresearch.co.nz/dsh/sslib/>
- 89 Himmelmann, L. (2022). HMM: Hidden markov models. In *Scientific Software*. [https://](https://CRAN.R-project.org/package=HMM)
90 CRAN.R-project.org/package=HMM
- 91 Król, A., & Saint-Pierre, P. (2015). SemiMarkov: An R package for parametric estimation
92 in multi-state semi-markov models. *Journal of Statistical Software*, 66(6), 1–16. [http:](http://www.jstatsoft.org/v66/i06/)
93 www.jstatsoft.org/v66/i06/
- 94 O'Connell, J., & Højsgaard, S. (2011). Hidden semi markov models for multiple observation
95 sequences: The mhsmm package for R. *Journal of Statistical Software*, 39(4), 1–22.
96 <http://www.jstatsoft.org/v39/i04/>
- 97 Vergne, N. (2008). Drifting Markov models with polynomial drift and applications to DNA
98 sequences. *Statistical Applications in Genetics and Molecular Biology*, 7(1). [https:](https://doi.org/10.2202/1544-6115.1326)
99 doi.org/10.2202/1544-6115.1326

DRAFT