Review of HSMM R and Python software

C. Bérard[1], M.-J. Cros[2], J.-B. Durand[3],
C. Lothodé[4], S. Plancade[2], R. Trépos[2], **N. Vergne**[5]

[1]*LITIS, Univ Rouen Normandie, France*
[2]*INRAE-MIAT, Univ Toulouse, France*
[3]*AMAP, CIRAD, Montpellier, France*
[4]*LAREMA, CNRS, SFR MATHSTIC, Univ Angers, France*
[5]*LMRS, CNRS, Univ Rouen Normandie, France*

MaSeMo : Markov, Semi-Markov Models and Associated Fields
(from Theory to Application and back), Paris, 1-4 July 2025

Plan

Plan

Review of Software for HSMMs

- Objective : Review of R and Python packages for Explicit Duration Hidden Markov Models (ED-HMMs).
- Focus on multichain models, coupled chains, and the integration of covariates.
- Identified a significant lack of dedicated packages for these specific themes.
- Comparison based on :
  - General characteristics ;
  - Sojourn duration distributions ;
  - Emission distributions.
- Two illustrative examples :
  - Squirrel toy model (discrete emissions, Python) ;
  - Deer toy model (continuous emissions, R).

Available Packages for HSMMs

Package	Language	Package	Language
CHMM	R	mhsmm	R
DNN-HSMM	Python	online_hmm	Python
edhsmm	Python	PHSMM	R
GENSCAN	C	pyhsmm	Python
hhsmm	R	rarhsmm	R
hmmTMB	R	SequenceAnalysis	Python
hsmm	R	signalHsmm	R
hsmmlearn	Python	smmR	R
LaMa	R	ziphsmm	R
mHMMBayes	R		

Inventory of HSMM and related software packages (R, Python, C)

Plan

Package : hhsmm

- **Last update :** May 8, 2024 (CRAN)
- **Authors :** M. Amini, A. Bayat, R. Salehian
- **Code Source :** `https://cran.r-project.org/package=hhsmm`, `https://github.com/mortamini/hhsmm`
- **Emission :** Custom (Mixture Gaussian, multinomial, B-splines, robust, additive)
- **Sojourn durations :** Shifted NB, log, Poisson, gamma, Weibull, log-normal, non-parametric
- **Language :** R, C
- **Features :** Hybrid Markov/semi-Markov, handles multiple sequences, EM init via K-means

Package : hsmm

- **Last update :** April 25, 2013 (archived May 9, 2022)
- **Authors :** J. Bulla, I. Bulla
- **Code Source :** `https://cran.r-project.org/package=hsmm`
- **Emission :** Bernoulli, Normal, Poisson, Student's t
- **Sojourn durations :** Non-parametric, geometric, NB, log, Poisson
- **Language :** R, C++
- **Features :** Single sequence, user-defined EM init, Viterbi + smoothing, simulation available

Package : LaMa

- **Last update :** January 29, 2025 (CRAN)
- **Author :** J.-O. Koslik
- **Code Source :** `https://janoleko.github.io/LaMa/`
- **Emission :** Any univariate R distribution
- **Sojourn durations :** HMM approximation of HSMMs
- **Language :** R, C++
- **Features :** Multiple sequences, covariates allowed, user provides init, optimization via external R solvers

Package : mhsmm

- **Last update :** August 23, 2023 (CRAN)
- **Authors :** J. O'Connell, S. Højsgaard
- **Code Source :** `https://cran.r-project.org/package=mhsmm`
- **Emission :** Gaussian, multivariate Gaussian, Poisson, custom
- **Sojourn durations :** Shifted Poisson, gamma, non-parametric
- **Language :** R, C
- **Features :** Multiple sequences, right-censored EM, missing values allowed

Package : PHSMM

- **Last update :** February 9, 2021
- **Author :** J. Pohle
- **Code Source :** `https://cran.r-project.org/package=PHSMM`
- **Emission :** Normal, gamma, Poisson, Bernoulli
- **Sojourn durations :** Non-parametric
- **Language :** R
- **Features :** Single sequence, penalized ML estimation, deterministic or stationary init

Package : rarhsmm

- **Last update :** October 18, 2017
- **Authors :** Z. J. Xu, Y. Liu
- **Code Source :** `https://github.com/cran/rarhsmm`
- **Emission :** Multivariate Gaussian
- **Sojourn durations :** Non-parametric
- **Language :** R
- **Features :** with/without autoregressive coefficients, with/without regularization, Supports multiple sequences, deterministic init, no longer maintained

Package : ziphsmm

- **Last update :** May 22, 2018 (archived Dec 12, 2022)
- **Authors :** Z. J. Xu, Y. Liu
- **Code Source :** `https://cran.r-project.org/package=ziphsmm`
- **Emission :** Zero-inflated Poisson
- **Sojourn durations :** Logarithmic, geometric, shifted Poisson
- **Language :** R, C++
- **Features :** Single sequence, gradient descent with multiple init, stationarity assumed

Plan

edhsmm

- **Last update :** May 26, 2024 (archived)
- **Author :** A. C. Lubang
- **Code Source :** `https://github.com/poypoyan/edhsmm`
- **Emissions :** Multivariate normal, multinomial
- **Sojourn durations :** Non-parametric
- **Language :** Python with Cython for heavy computations
- **Features :** Multiple sequences, initial law via forward-backward smoothing, EM-based fitting, EM init via K-means, Viterbi decoding, options for right and experimental left censoring

hsmmlearn

- **Last update :** August 21, 2021
- **Author :** J. Vankerschaver
- **Code Source :** `https://github.com/jvkersch/hsmmlearn`
- **Emissions :** Gaussian, categorical
- **Sojourn durations :** Non-parametric + custom
- **Language :** C++/Cython with Python interface
- **Features :** single sequence only, initial law via smoothed initial probabilities, EM init by user, wraps same C++ code as R package `hsmm`; API inspired by `HMMLearn`

online_hmm

- **Last update :** April 10, 2015 (unmaintained)
- **Author :** A. Bietti
- **Code Source :** `https://github.com/albietz/online_hmm`
- **Emissions :** Gaussian, categorical (multinomial)
- **Sojourn durations :** Binomial, Poisson
- **Language :** Python
- **Features :** single sequence only, initial law and EM init by user, Online inference and parameter estimation for (H)SMMs with streaming data

OpenAlea : SequenceAnalysis

- **Last update :** Jan 23, 2018 / Aug 30, 2024 (under maintenance)
- **Authors :** Y. Guédon, J. B. Durand, C. Pradal, T. Arsouze
- **Code Source :** `https://github.com/openalea/StructureAnalysis/tree/master/sequence_analysis`
- **Emissions :** Multivariate conditionally independent ; discrete (categorical, binomial, Poisson, negative binomial, geometric Poisson), continuous (Gamma, zero-inflated Gamma, Gaussian, Inverse Gaussian, Von Mises)
- **Sojourn durations :** Binomial, Poisson, negative binomial, geometric Poisson, uniform
- **Language :** C++ with Python interface
- **Features :** multiple sequences, initial law via stationary distribution or MLE, EM init user-defined or default geometric distribution for sojourn, focus on discrete sequence analysis, HSMM with explicit durations.

pyhsmm

- **Last update :** August 1, 2020
- **Authors :** M. Johnson, A. Wiltschko, Y. Katz, et al.
- **Code Source :** `https://github.com/mattjj/pyhsmm`
- **Emissions :** Multivariate Gaussian (customizable)
- **Sojourn durations :** Poisson (customizable)
- **Language :** C++/Cython with Python interface
- **Features :** multiple sequences in Bayesian framework, parameters treated as random variables, uses Bayesian priors and Gibbs sampling, Bayesian nonparametric inference via HDP-HMM and HDP-HSMM ; Gibbs sampling for all latent variables including transitions, durations, emissions, and states

Plan

CHMM : R package

- **Last update on CRAN :** September 29, 2017
- **Authors :** X. Wang, J. Aubert
- **Code source :**
  `https://cran.r-project.org/web/packages/CHMM/`
  `https://github.com/julieaubert/CHMM`
- **Emission :** Normal
- **Sojourn durations :** Not concerned
- **Language :** `R`
- **Features :** single sequence, initial law defined specifically by the model, EM init by clustering techniques (`mclust` or `kmeans`)

Dedicated to Coupled Hidden Markov Models (not semi-Markovian). Manages multiple hidden Markov chains with correlation between hidden states at a given time-step. Provides EM and VEM algorithms, suited for large state spaces or long chains. Main application : detecting CNVs in DNA sequences.

DNN-HSMM : Python library

- **Last update on GitHub :** March 14, 2021
- **Author :** S. Takaki
- **Code source :** `https://github.com/sp-nitech/DNN-HSMM`
- **Emission :** Gaussian
- **Sojourn durations :** Discrete Gaussian
- **Language :** Python (PyTorch)
- **Features :** single sequence, initial law fixed (first state), gradient descent with user-provided initial parameter values

Python implementation for speech synthesis using HSMM parameters encoded by a Deep Neural Network. Left-right deterministic transitions. State restoration not implemented. Parameters estimated by backpropagation of log-likelihood derivatives through PyTorch.

GENSCAN : C program

- **Last update :** February 18, 2003
- **Author :** C. Burge
- **Code source :** `http://hollywood.mit.edu/GENSCAN.html`
- **Emission :** Non-parametric
- **Sojourn durations :** Non-parametric
- **Language :** C
- **Features :** multiple sequences and initial law : not detailed

Program dedicated to gene structure identification in genomic DNA using HSMMs with non-parametric emission and sojourn duration distributions.

hmmTMB : R package

- **Last update on CRAN :** October 24, 2023 ; January 14, 2025 (Github)
- **Authors :** T. Michelot, R. Glennie
- **Code source :**
  `https://cran.r-project.org/web/packages/hmmTMB/`
  `https://github.com/TheoMichelot/hmmTMB/`
- **Emission :** Beta, binomial, categorical, Dirichlet, exponential, folded normal, Gamma, log-normal, multivariate normal,negative binomial, Poisson, Student's T, truncated normal, Tweedie, von Mises, Weibull, wrapped Cauchy
- **Sojourn durations :** HMM approximation of HSMMs via state-space augmentation
- **Language :** C++ and R
- **Features :** multiple sequences, initial law free or stationary, EM with deterministic or k-means methods

Implementation of HMMs with flexible emission distributions and HSMM approximations using Template Model Builder (TMB). Maximum likelihood and Bayesian estimation possible.

mHMMBayes : R package

- **Last update on CRAN :** April 4, 2024
- **Authors :** E. Aarts, S. Mildiner Moraga
- **Code source :**
  `https://cran.r-project.org/web/packages/mHMMbayes/`
  `https://github.com/emmekeaarts/mHMMbayes`
- **Emission :** Multivariate discrete (categorical), Poisson, Gaussian
- **Sojourn durations :** Geometric only (Markov)
- **Language :** C++ and R
- **Features :** multiple sequences, initial law fixed as stationary distribution, MCMC with user-provided initial values ; default priors are inverse Wishart (covariances) and multivariate Gaussian otherwise

Bayesian estimation of multilevel HMMs accommodating longitudinal multivariate data with individual heterogeneity. Includes covariates and missing data treatment (MAR assumption).

signalHsmm : R package

- **Last update on CRAN :** November 15, 2018 ; May 4, 2020 (Github)
- **Authors :** M. Burdukiewicz, P. Sobczyk, J. Chilimoniuk
- **Code source :**
  `https://cran.r-project.org/web/packages/signalHsmm/`
  `https://github.com/michbur/signalHsmm`
- **Emission :** Non-parametric
- **Sojourn durations :** Non-parametric
- **Language :** R
- **Features :** single sequence, initial law defined as free parameters

Package dedicated to predicting signal peptides in eukaryotic proteins using HSMMs with non-parametric emission and sojourn duration distributions.

smmR : R package

- **Last update on CRAN :** August 3, 2021
- **Authors :** V. Barbu, C. Bérard, D. Cellier, F. Lecocq, C. Lothodé, M. Sautreuil, N. Vergne
- **Code source :**
  `https://cran.r-project.org/web/packages/smmR/`
  `https://plmlab.math.cnrs.fr/lmrs/statistique/smmR`
- **Emission :** Not concerned
- **Sojourn durations :** Non-parametric, uniform, geometric, Poisson, discrete Weibull, negative binomial
- **Language :** `R`
- **Features :** multiple sequences, initial law deterministic or estimated, different type of sojourn duration, censoring, reliability.

Dedicated to estimation and simulation of semi-Markov models on discrete state spaces with various sojourn distributions. Computes reliability measures such as maintainability and failure rates.

Plan

Overview of HSMM Libraries

Hidden Semi-Markov Models (HSMMs) have many available software packages, each with distinct features. We compare them based on :

- **General characteristics** : multiple sequence support, handling missing data, initial distribution, initialization methods, model comparison (see first table).
- **Sojourn duration distributions** : modeling time spent in states (see second table).
- **Emission distributions** : modeling observations given states (see third table).

The following tables summarize these features across key Python and R libraries.

General Comparison

Software	Multiple sequences	Missing data	Initial distribution	Proposed automatic initialisation	Comparison criteria	Language	Last update
CHMM	-	-	E	x	-	R	2017
DNN-HSMM	-	-	F	-	-	Python	2021
edhsmm	x	-	F	x	-	Python	2023
GENSCAN	-	-	F	-	-	C	2003
hhsmm	x	x	F	x	x	R	2022
hmmTMB	x	-	E	-	-	R	2025
hsmm	-	-	F	-	-	R	2013
hsmmlearn	-	-	F	x	-	Python	2021
LaMa	x	-	E	-	-	R	2025
mHMMBayes	x	x	F, E	x	x	R	2024
mhsmm	x	x	F,E	-	-	R	2017
online_hmm	-	-	F	-	-	Python	2015
PHSMM	-	-	F, E	x	-	R	2021
pyhsmm	x	-	E	-	x	Python	2020
rarhsmm	x	-	-	-	-	R	2018
SequenceAnalysis	x	-	F, E	x	x	Python	2016
signalHsmm	-	-	F	-	-	R	2018
smmR	x	-	F,E	-	x	R	2021
ziphsmm	-	-	F	-	-	R	2018

Sojourn durations

Software	Non-parametric	Gamma	Geometric	Logarithmic	LogNormal	Negative Binomial	Poisson	Uniform	Discrete Weibull	Binomial	Custom
CHMM	-	-	-	-	-	-	-	-	-	-	-
DNN-HSMM	-	-	-	-	-	-	-	-	-	-	-
edhsmm	x	-	-	-	-	-	-	-	-	-	-
GENSCAN	x	-	x	-	-	-	-	-	-	-	-
hhsmm	x	x	x	shifted	x	shifted	shifted	-	x	-	-
hmmTMB	x*	-	x	-	-	-	-	-	-	-	-
hsmm	x	-	x	x	-	-	x	x	-	-	-
hsmmlearn	x	-	-	-	-	-	-	-	-	-	x
LaMa	x*	-	x	-	-	-	-	-	-	-	-
mHMMBayes	-	-	x	-	-	-	-	-	-	-	-
mhsmm	x	x	-	-	-	-	-	shifted	-	-	-
online_hmm	-	-	-	-	-	-	x	-	-	x	-
PHSMM	x	-	-	-	-	-	-	-	-	-	-
pyhsmm	-	-	-	-	-	-	x	-	-	-	x
rarhsmm	x	-	-	-	-	-	-	-	-	-	-
SequenceAnalysis	x	-	shifted	-	-	shifted	shifted	x	-	shifted	-
signalHsmm	x	-	-	-	-	-	-	-	-	-	-
smmR	-	-	x	-	-	x	x	x	x	-	-
ziphsmm	-	-	x	x	-	-	x	-	-	-	-

Observations

Software	Bernoulli	Custom	Normal	Multivariate normal	Poisson	Student's t	Nonparametric
CHMM	-	-	x	-	-	-	-
DNN-HSMM	-	-	x	-	-	-	-
edhsmm	-	-	x	x	-	-	x
GENSCAN	-	-	-	-	-	-	-
hhsmm	-	x	x	x	-	-	x
hmmTMB	x	x	x	x	x	x	x
hsmm	x	-	x	-	x	x	-
hsmmlearn	-	-	x	-	-	-	x
LaMa	x	x	x	-	x	x	x
mHMMBayes	-	-	x	-	x	-	x
mhsmm	-	x	x	x	x	-	-
online_hmm	-	-	x	-	-	-	x
PHSMM	x	-	x	-	x	x	-
pyhsmm	-	-	-	x	-	-	-
rarhsmm	-	-	-	x	-	-	-
SequenceAnalysis	x	-	x	-	shifted	-	x
signalHsmm	-	-	-	-	-	-	-
smmR	-	-	-	-	-	-	-
ziphsmm	-	-	-	-	x*	-	-

Plan

Docker Container for HSMM Libraries

- Docker bundles applications with dependencies for easy deployment.
- We provide a Docker image with snapshots of key HSMM libraries : **edhsmm** (0.2.2), **hhsmmlearn** (commit 69bc8aa), **hhsmm** (0.2.5), **PHSMM** (1.0).
- Image URL : `https://forgemia.inra.fr/inca-hsmm/software-review/container_registry`.
- Usage :

```
## Launch Jupyter Lab:
docker run --rm -it -p 8888:8888
registry.forgemia.inra.fr/inca-hsmm/software-review:2.0
## Launch bash shell:
docker run --rm -it -p 8888:8888
registry.forgemia.inra.fr/inca-hsmm/software-review:2.0 bash
```

- Example : Load `edhsmm` in Python or `hhsmm` in R inside container.

Plan

Squirrel Toy Model : HSMM Setup

- Model : Squirrel moves among 3 reserves over winter (500 days).
- Data : 300 sequences of observations (naturalist's guesses).
- Transition matrix (true model) :

$$\begin{pmatrix} 0.0 & 0.5 & 0.5 \\ 0.3 & 0.0 & 0.7 \\ 0.4 & 0.6 & 0.0 \end{pmatrix}$$

- Max duration in a reserve = 4 days, modeled with non-parametric sojourn duration distributions.
- Emission : Multinomial distribution modeling observation errors (naturalist guesses).

k	$\mathbb{P}(X_{n+1}=1 \mid J_n=k)$	$\mathbb{P}(X_{n+1}=2 \mid J_n=k)$	$\mathbb{P}(X_{n+1}=3 \mid J_n=k)$	$\mathbb{P}(X_{n+1}=4 \mid J_n=k)$
1	0.1	0.005	0.005	0.89
2	0.1	0.005	0.89	0.005
3	0.1	0.89	0.005	0.005

k	$\mathbb{P}(Y_t=1 \mid Z_t=k)$	$\mathbb{P}(Y_t=2 \mid Z_t=k)$	$\mathbb{P}(Y_t=3 \mid Z_t=k)$
1	0.8	0.1	0.1
2	0.1	0.8	0.1
3	0.1	0.1	0.8

HSMM Parameter Initialization and Simulation

- Define true HSMM parameters in Python (initial dist, duration, transition, emission).
- Simulate 300 trajectories of length 500 using the true model.

```
# Example: initialize true model parameters
hsmm.pi = np.array([2/3, 1/3, 0])
hsmm.dur = np.array([
   [0.1, 0.005, 0.005, 0.89],
   [0.1, 0.005, 0.89, 0.005],
   [0.1, 0.89, 0.005, 0.005]])
hsmm.tmat = np.array([...])
hsmm.emit = np.array([...])
```

EM Algorithm for Parameter Estimation

- Initialize HSMM with rough starting parameters.
- Run EM algorithm for 20 iterations to estimate parameters.
- After 7 loops, 30 seconds, we obtain :

$$\begin{pmatrix} 0.0 & 0.41 & 0.59 \\ 0.27 & 0.0 & 0.73 \\ 0.53 & 0.47 & 0.0 \end{pmatrix}$$

The non parametric sojourn duration distribution is the following :

k	$\mathbb{P}(X_{n+1}=1 \mid J_n=k)$	$\mathbb{P}(X_{n+1}=2 \mid J_n=k)$	$\mathbb{P}(X_{n+1}=3 \mid J_n=k)$	$\mathbb{P}(X_{n+1}=4 \mid J_n=k)$
1	5.2e-02	9.8e-05	6.8e-03	9.41e-01
2	2.3e-02	4.2e-02	8.9e-01	4.1e-02
3	3.3e-02	9.7e-01	8.3e-04	1.7e-04

The non parametric emission is the following:

k	$\mathbb{P}(Y_t=1 \mid Z_t=k)$	$\mathbb{P}(Y_t=2 \mid Z_t=k)$	$\mathbb{P}(Y_t=3 \mid Z_t=k)$
1	0.85	0.1	0.05
2	0.1	0.8	0.1
3	0.05	0.08	0.87

https://forgemia.inra.fr/inca-hsmm/software-review/-/blob/main/squirrel/squirrel.ipynb

Plan

Model and Application

- Aim : Link accelerometer data to cervid activities using an ED-HSMM.
- Observations : Scalar roll angle from 3D accelerometer data sampled at 8 Hz (0.125s timestep).
- Hidden states : 5 behaviours (1 : Foraging head-down, 2 : Grooming, 3 : Running, 4 : Unmoving, 5 : Walking head-up)
- Sojourn duration modeled by Poisson (approximated by Gamma in implementation).
- Emission distributions : Gaussian with state-dependent means and variances.

Model Parameters and Simulation

- Transition matrix M estimated nonparametrically from labeled data :

$$M = \begin{pmatrix} 0 & 0.025 & 0.028 & 0.744 & 0.203 \\ 0.028 & 0 & 0.014 & 0.888 & 0.070 \\ 0.080 & 0.028 & 0 & 0.256 & 0.636 \\ 0.279 & 0.269 & 0.032 & 0 & 0.420 \\ 0.236 & 0.043 & 0.071 & 0.650 & 0 \end{pmatrix}$$

- Sojourn and emission parameters (example) :

i	λ_i	m_i	σ_i^2
1	98.02	20.41	12.04
2	67.42	-14.98	14.45
3	37.69	-12.88	9.01
4	103.64	-24.60	14.61
5	46.13	-4.32	8.61

- Used Gamma distribution for sojourn times in fitting.
- Simulated sequences generated with these parameters.

Results and Estimation

- Initial parameter estimates from clustering show some deviation.
- Estimated transition matrix example :

$$M = \begin{pmatrix} 0.000 & 0.014 & 0.043 & 0.686 & 0.257 \\ 0.035 & 0.000 & 0.035 & 0.860 & 0.070 \\ 0.000 & 0.048 & 0.000 & 0.190 & 0.762 \\ 0.254 & 0.288 & 0.034 & 0.000 & 0.424 \\ 0.171 & 0.045 & 0.090 & 0.694 & 0.000 \end{pmatrix}$$

- Estimated sojourn and emission parameters :

i	η_i	m_i	σ_i^2
1	86.791	20.369	12.137
2	74.942	-14.974	13.762
3	40.699	-12.751	9.179
4	107.340	-24.578	14.680
5	42.915	-4.301	8.560

- Computation time : 1 to 5 minutes on standard machine.

Plan

Conclusion and Perspectives

- Not a lot of packages still maintained for HSMMs.
- Each package has its own particularities.
- A lack of packages for the specific theme of our projets : multichains, coupled chains, covariates.
- Most HSMM software focus on Explicit Duration HMMs (ED-HMMs), where sojourn durations depend only on the current state.
- Only `smmR` package supports more general sojourn duration distributions.
- We are developing `hsmmR`, a new R package for simulation and estimation of general HSMMs, soon available on CRAN :

  `https://cran.r-project.org/web/packages/hsmmR/index.html`

- Upcoming book :
  A comprehensive Guide to HSMM : Theory, Software, and Advanced Extensions, Nathalie Peyrard, Benoîte de Saporta.