HaMMLET(1)

NAME

HaMMLET - Fast Bayesian Inference for Hidden Markov Models using Dynamic Haar Wavelet Compression.

DESCRIPTION

HaMMLET is a fast Forward-Backward Gibbs (FBG) sampler for Bayesian HMM. It also implements alternative sampling schemes (currently supported: Mixture Model sampling). Given numerical input data and prior parameters, it outputs a full distribution of latent HMM states for each position, integrating over the entire parameter space. In modern applications, such as the detection of copy-number variants (CNV) using whole-genome sequencing data, the input sizes are on the order of millions to billions of data points. To avoid prohibitively long running times and slow convergence, HaMMLET uses the Haar wavelet transform to dynamically compress the data into blocks of sufficient statistics, based on the lowest noise estimate in each iteration of the Gibbs sampler.

When using HaMMLET, please cite the following paper (a BibTeX file is provided in doc/hammlet.bib):

> Wiedenhoeft, J., Brugel, E., & Schliep, A. (2016). "Fast Bayesian Inference of Copy Number Variants using Hidden Markov Models with Wavelet Compression". PLOS Computational Biology, 12(5), e1004871. http://doi.org/10.1371/journal.pcbi.1004871. This paper was selected for oral presentation at RECOMB 2016.

USAGE EXAMPLE

hammlet -f data.csv -s 3 -a -R 0

Run HaMMLET on input file **data.csv**, using a **3**-state model with **automatic priors** and random-number seed **0** for reproducibility. This outputs the state marginals to **hammlet-marginals.csv**.

cat data.csv | hammlet -a -R 32 -s 8 -i M 100 0 F 200 5 -t 1 10 -O blocks compression marginals -o result- .csv

This reads data from STDIN. Using a model with **automatic** emission priors with standard parameters and a random seed of **32** for reproducibility, HaMMLET inferes an **8**-state segmentation. Sampling of state sequences is done by first running **mixture sampling** for **100** iterations, **none** of which is recorded, followed by **Forward-Backward Gibbs sampling (FBG)** for **200** iterations, recording every **fifth** of them. This is done using prior **transition** weights of **10** for self-transitions, and **1** for transition between different states. The output consists of the sizes of **blocks** for each iteration, the **compression** ratios as well as the state **marginals**. The output files are **result-blocks.csv**, **result.compression.csv**, and **result-marginals.csv**, respectively (notice the space **-o** takes two arguments, prefix and suffix).

OPTIONS

-h | -help Display a friendly help message.

-v | -verbose Print information to *STDOUT* during run-time.

-g | -arguments Print arguments. For each flag, print an asterisk if it was set by the user, as well as the parameters being used. If the flag was not set, these are the default parameters.

INPUT AND OUTPUT

HaMMLET reads the observed emission values as a stream of whitespace-separated numeric values from *STDIN*, unless **-f** is specified. Linebreaks are treated as whitespace, and no formatting such as CSV is interpreted. For multivariate models, the dimensions of each position are filled in increasing order, after which dimensions for the next position are filled. If the number of input values is not a multiple of the number of dimensions (see **-s** option), an exception is thrown.

The output consists of a CSV file representing a run-length encoded version of the state marginals. The first column represents the length of a segment (number of input positions), and subsequent columns represent the recorded counts for each state, in increasing order of state number.

-f *FILE* | -input-file *FILE* Read input data from *FILE* instead of *STDIN*.

-o *PREFIX SUFFIX* | -output-prefix *PREFIX SUFFIX* The prefix and suffix for the output file paths. Output files names are created by adding a short descriptor, e.g. *PREFIX*marginals*SUFFIX* for the file containing the marginal state distribution; for additional files, see the **-O** option for details. If this option is not set, the behavior depends on the -c/-f flags: If -c is set, -o **hammlet .csv** is used; if -f *FILENAME.EXT* is provided, -o **FILENAME- .EXT** is used instead.

-O *TYPE ...* | -output-data *TYPE ...* Specify a list of data types to be output in addition to the marginals. This only applies to recorded iterations as specified using **-i**. It may contain any of the following:

B | blocks Output the block structure (sizes of all blocks) created by dynamic compression, separated by whitespace.

C | compression Output the compression ratio for each iteration.

D | mapping Lines represent states in ascending order. Columns represent data dimensions. Entries represent the index of emission distributions for each state and data dimension.

G | segments Output the number of segments in each iteration, as well as the number of values used to store the compressed marginals (for diagnostic purposes).

P | parameters Output the emission parameters for each state in increasing order of state number, separated by tabs.

S | sequences Output each state sequence individually, one per line, separated by whitespace, using run-length encoding of the form *LENGTH*:*STATE*.

-w | -overwrite Overwrite existing output files. If **-w** is not provided and an output file already exists, an exception is thrown.

MODEL SPECIFICATIONS

-s *PARAM* | -states *PARAM* Definition of hidden states. *PARAM* may take the following forms [Default: **3**]:

NRSTATES The typical, simple case of an HMM: Given a single unsigned integer *NRSTATES*, the data is assumed to be one-dimensional and generated from this many hidden states.

MAPPING NRDIST NRDIM LIST State definitions in the form of a *MAPPING*. The data has *NRDIM* dimensions, and there are *NRDIST* different emission distributions. If *NRDIM* is not provided, it defaults to **1**. A state is defined as a vector S of size *NRDIM*, where S[i] denotes that the i-th dimension of an emission sampled from this state was generated from the S[i]-th emission distribution. For instance, if a 3-dimensional data point is assigned to a state with mapping S=[1 1 2], its first two dimensions are assumed to be generated by emission distribution 1, and its third dimension by emission distribution 2. The following *MAPPING* schemes are supported:

S | shared All values at a given data position are assumed to be generated by the same emission distribution across *NRDIM* data dimensions. *NRDIM* defaults to 1, and the number of states equals *NRDIST*.

C | combination States express all possible combinations of emission distributions, resulting in *NRPARAMS*^*NRDIM* different states. For instance, **-s C 2 3** generates 2^3=8 state mappings: (1 1 1), (1 1 2), (1 2 1), (1 2 2), (2 1 1), (2 1 2), (2 2 1), and (2 2 2). *NRDIM* defaults to **1**, so that **-s C 3 1** is equivalent to **-s 3**. *LIST* is ignored.

M | manual Manually specify a *LIST* of pointers for *NRDIST* emission distributions and *NRDIM* data dimensions (*NRDIM* must be set explicitely, even if it is **1**). The emission distribution for the d-th dimension of the k-th state is the (k*NRDIM* + d)-th element in *LIST* (all counts are zero-based), hence the length of *LIST* must be a multiple of *NRDIM*, its elements must be at least 0 and smaller than *NRDIST*, and the number of states is the length of *LIST* divided by *NRDIM*.

-e *DIST PARAM* | -emissions *DIST PARAM* Set the emissions to be variates of a given *DIST*ribution, and let their parameters be sampled from priors using the given hyper*PARAM*eters. The behavior of this option depends on the number of tokens: Let K be the number of hyperparameters per prior, and D the number of emission distributions (see **-s** option). If *PARAM* consists of K tokens, all priors are assumed to have those same hyperparameters. If there are N*K tokens, each prior gets its specific set of hyperparameters. In all other cases, an exception is thrown. Arguments may take the following forms [Default: **normal**. This means that **-a** has to be provided if **-p** is not.]:

normal [*PARAMs*] For Normal emissions, *PARAM* is a collection of 4-tuples *ALPHA BETA MU NU*, representing parameters to the Normal-Inverse Gamma distribution, sorted by state. If **-a** is set, *PARAM* is *VAR P* instead, where *P* is the probability to sample emission variances less or equal than *VAR*; if these parameters are not provided, they default to **0.2 0.9**.

NOTE: No other emission type than Normal is currently supported.

If neither **-a** nor *PARAM* is provided, an exception is thrown.

-a | -auto-priors Use automatic hyperparameters for emission priors, based on the wavelet transform of the data. This changes the meaning of parameters passed to **-p**.

-t *VALUES* | -transitions *VALUES* Parameters for transition probabilities. These are the parameters alpha for a Dirichlet distribution. *VALUES* can take the following forms:

ALPHA A single number means that all alpha-parameters are set to the same value.

SELF TRANS All alphas corresponding to self-transitions are set to *SELF*, the others to *TRANS*.

-S | -no-self-transitions Do not use self-transition probabilities within blocks (this has no effect for mixture sampling).

-I *ALPHA* | -initial *ALPHA* Sets the alpha parameter of the Dirichlet distribution used as a prior for the initial state distribution.

SAMPLING SCHEME

-R | -random-seed An unsigned integer value to be used to seed the random number generator. If **-R** is not set, a seed is generated from the current epoch time. A seed should be set manually using **-R** whenever reproducibility is required.

-i *SCHEME ...* | -iterations *SCHEME ...* A list of sampling *SCHEME*s, each of which consists of either a single token *FLAG*, or three tokens, *TYPE ITER THIN*.The following *FLAG*s can be used:

P Sample from priors. Since the very first action in a Gibbs sampler is a sampling from the prior, an additional **P** is always silently prepended to **-i**.

S Set compression to *static*, the block structure is determined by the current state of emission parameters and remains unchanged until **D** is provided.

D Set compression to *dynamic*, the block structure changes at every iteration based on ht current state of emission parameters and remains unchanged unto **D** is provided.

The following triples can be used:

1. The *TYPE* of sampling method to be used is one of the following:

   M *Mixture sampling* treats compression as a way to impose equality relations on otherwise exchangeable data points. It completely ignores transition probabilities passed to the model, and instead assumes transitions to be implied in the block structure alone. This is much faster than the other methods,

as it depends linearly on the number of states, but is not truly an HMM. High-variance components are prone to over-segmentation, and spurious differences in sampled values can lead to segments which come from the same true state being assigned to different states. However, if the variance is expected to be similar over all states, this variant can yield reasonably good results very fast.

F *Forward-Backward Gibbs sampling* uses a dynamic programming trellis to quickly sample state sequences unaffected by auto-correlation due to adjacent blocks. FBG is considered the state-of-the-art for Gibbs sampling in HMM. Running times depends quadratically on the number of states.

2. The number of sampling *ITER*ations.
3. The type of *THIN*ning to be used to record sampled state sequences (0=record none, 1=record all, 2=record every second sample, etc.).

[Default: **M 500 0 S P F 200 0 F 300 3**. Under this scheme, 100 unrecorded mixture iterations are performed to converge to a block structure, which is then fixed. The emission parameters are resampled from the prior so as to remove the influence of the mixture observations, and 200 FBG iterations for burn-in are performed, followed by 300 FBG iterations, every third of which is recorded, resulting in 100 recorded iterations.]

COMPRESSION

-m ***FLOAT*** **| -weight-multiplier** ***FLOAT*** Multiply weights by this factor, to avoid overcompression. [Default: **1.0**]

CAVEATS

While HaMMLET is designed to minimize memory consumption (univariate models of 100 million data points can be handled on a standard laptop), one should still be aware that the size of the marginal state records and the trellis cannot be predicted before running the inference. As a consequence, data that only allows for low compression ratios may still incur huge memory overhead, as it negates the central approach that makes FBG feasible on such scales. If memory consumption gets out of hand, you might want to try increasing the number of burn-in steps; if the sampler has not fully converged, individual iterations might have very low compression, even though the data itself would allow for better ratios. Likewise, decreasing the number of states might be an option, since superfluous state parameters will be sampled solely from the prior and yield arbitrarily low noise variances. If this does not work, using Mixture model sampling might be an option, but results should be interpreted with care, see **-i** option.

Though the model should work for any emission distribution in the exponential family (Normal, Poisson, Exponential, Laplace, Gamma, Chi-Squared etc.), only Normal emissions are implemented at the moment.

Multivariate models are supported in the sense that multiple data dimensions may share their generating parameters. True multivariate models such as Normals with non-diagonal covariance matrix are not yet supported.

Plotting the results is done using external Python libraries (NumPy, Matplotlib). As these are not optimized for large-scale applications, this can take a long time, often longer than the inference itself.

HaMMLET does not support the convention of combining single-letter options, such as replacing **-x -y -z** by **-xyz**.

HISTORY

The first version of HaMMLET was developed by Eric Brugel and John Wiedenhoeft, and published in 2016 in PLOS CompBio and RECOMB. It used a wavelet tree data structure for dynamic compression. The current version is designed for minimal memory footprint in large-scale applications. Changes include: a breakpoint array data structure for optimal wavelet compression, an in-place algorithm for its construction, run-length-encoded output, and a queue-based implementation to record run-length-encoded state sequences. It is currently developed and maintained by John Wiedenhoeft (ORCID: 0000-0002-6935-1517) at https://github.com/wiedenhoeft/HaMMLET.

REPORTING BUGS

GitHub issue tracking system: https://github.com/wiedenhoeft/HaMMLET/issues

SEE ALSO

Current hosting site: https://wiedenhoeft.github.io/HaMMLET/

Current repository: https://github.com/wiedenhoeft/HaMMLET

Stable link: https://schlieplab.org/Software/HaMMLET/

Documentation in different formats (pdf, html, txt, man) can be found in the doc/ subfolder of HaMMLET's installation directory.