CHMM : an R package for coupled Hidden Markov Models

Julie Aubert, Xiaoqiang Wang, Emilie Lebarbier, Stephane Robin

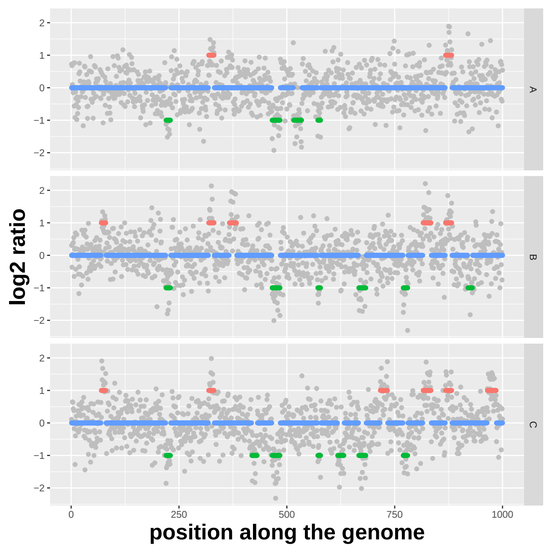
CHMM: an R package for coupled Hidden Markov Models

Julie Aubert (1), Xiaoqiang Wang (1,2), Emilie Lebarbier (1) & Stéphane Robin (1)

(1) UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005, Paris, France.
(2) School of Mathematics and Statistics, Shandong University, Weihai, China.

Detection of CNV taking into account dependency between individuals

Copy number variations (CNVs) are genomic alterations that result in an abnormal number of copies of one or more genes: duplication (green), normal (blue), deletion (red).



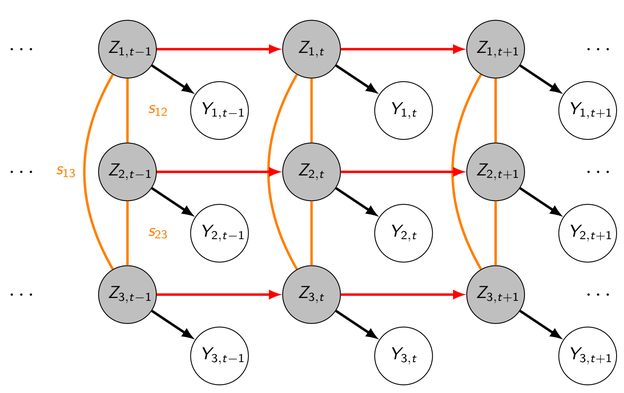
CNV detection of a simulated sample.

	A	B	C
A	1	0.61	0.56
B	0.61	1	0.75
C	0.56	0.75	1

Kinship data

Kinship measures pairwise genetic relatedness between individuals.

Coupled Hidden Markov Models: graphical representation [2]



Notations

- $Y_{i,t}$: observation
- $Z_{i,t}$: hidden status
- s_{ij}: similarity between i and j

within series dependence: $(Z_{i,t-1}, Z_{i,t})$ are Markov-dependent

between series dependence: $\forall (i \neq j), (Z_{i,t}, Z_{j,t})$ are not independent

Coupled Hidden Markov Models (CHMM): model

- **Observed** process:

$$(Y_{i,t}|Z_{i,t} = q) \sim \mathcal{N}(\mu_q, \sigma^2)$$

where μ_q is the mean value in the state q ($q = 1, \cdots, Q$).

- Joint **hidden** process: $(\mathbf{Z}_t)_t$, with $\mathbf{Z}_t = (Z_{1,t}, \cdots, Z_{I,t})$: Q^I states.

$$P(\mathbf{Z}_t = \ell | \mathbf{Z}_{t-1} = k) \propto W_\ell \prod_i \pi_{k_i,\ell_i}$$

where

- π is a $Q \times Q$ transition matrix
- dependency relationships among individuals is encoded in $W_\ell = \prod_{i,j \neq i} \omega^{s_{ij} \mathbb{1}_{\{q_j^\ell \neq q_i^\ell\}}}$ with $\omega < 1$.
- $\omega = 1$: independent case. Equivalent to independent HMM (*iHMM*).

Variational inference [1,2]

When I (the number of individuals) is large, $P(\mathbf{Z}|\mathbf{Y})$ is not computable.

Mean-field approximation

$$\tilde{P}(Z) = \arg\min_{\tilde{P} \in \mathcal{P}} \mathcal{KL}\left[\tilde{P}(Z); P(Z|Y)\right]$$

where $\mathcal{P} = \left\{\tilde{P}(Z) \,|\, \tilde{P}(Z) \propto \prod_i \prod_t \tilde{P}(Z_{i,t}|Z_{i,t-1})\right\}$ (independent Markov chains)

Forward part of the VE-step

Let denote $p_{itqr} = \tilde{P}(Z_{i,t} = r | Z_{i,t-1} = q)$, then we obtain a set of fixed point equations for p_{itqr}:

$$p_{itqr} \propto \pi_{qr}\ f(Y_{i,t}, \mu_r, \sigma^2)\ \times \omega^{\sum_{j \neq i} s_{ij}\left(1 - \mathbb{E}_{\tilde{P}} Z_{j,t}^r\right)}$$

References and acknowledgements

[1] Ghahramani, Z. and Jordan, M. (1997). Machine learning, 29(2-3):245-273.
[2] Wang, X. et al. (2017). *Submitted*.
[3] Daudin, J.-J., Picard, F. and Robin, S. (2008). Stat. Comput. 18, 173-83.

This work was supported by the CNV-Maize program funded by the french National Research Agency (ANR-10-GENM-104) and France Agrimer (11000415).

Selection criterion [2,3]

$$\widehat{Q} = \arg\max_Q \mathcal{J}_Q(Y, \widehat{\theta}, \tilde{P}) - [1 + Q(Q-1)] \log(IT)/2,$$

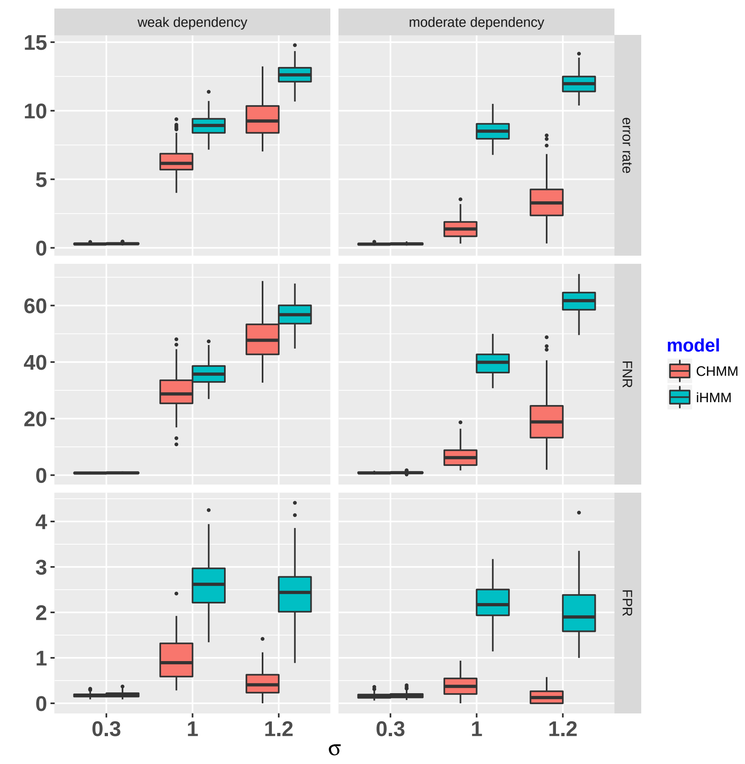
where $\mathcal{J}_Q(Y, \widehat{\theta}, \tilde{P})$ is the maximized lower bound of the Q-state model.

Simulation study

Runtime (in second), Weak dependency, $\sigma = 1$, I: number of lines

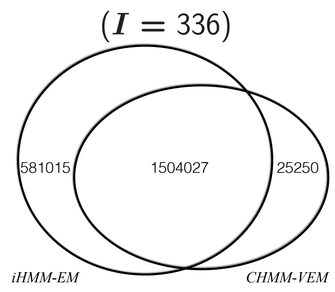
I	*iHMM-EM*	*CHMM-VEM*	*CHMM-EM*
2	0.8	0.4	2.0
3	1.1	0.5	11.2
4	1.2	0.6	79.4
5	1.6	0.8	920.2

Classification accuracy (%) for $I = 3$



Coupled HMM applied to the detection of CNV in the maize

Loci detected as deleted ($I = 336$)



Classification accuracy (validated 58 Fv2 alterations)

I	1	6	49	80	153	336
$\overline{s}_I$	1.0	0.7	0.7	0.7	0.6	0.6
FPR(%)	12.6	10.4	10.0	9.3	8.9	8.9
FNR(%)	24.1	24.1	24.1	25.9	25.9	25.9

$\overline{s}_I$: mean kinship within the panel.

The joint analysis with correlated lines reduces the proportion of falsely detected alterations.

CHMM package

```
library(CHMM)
data(toyexample)
# Variational inference of a coupled hidden Markov Chains
resCHMM <- coupledHMM(X = toydata, nb.states = 3, S =
    cor(toystatus), omega.list = c(0.3, 0.5, 0.7, 0.9))
# Breakpoints positions and status of segments
info <- clusterseg(resCHMM$status)
```

	sample	*posbegin*	*posend*	*status*
1	*Sample_5*	*1*	*17*	*2*
2	*Sample_5*	*18*	*30*	*1*
3	*Sample_5*	*31*	*66*	*2*

Conclusions

A model and associated inference for the detection of CNV taken into account dependency.

- Selection criterion
- Heuristic for choosing the value of the parameter ω.
- CHMM R package available from the CRAN.