oHMMed

Ordered Hidden Markov Models for Compact Genomic Segmentations

Wassim Alkouri Maximilien Godin
Université Libre de Bruxelles

Introduction

- Genomic tracks (GC content, SNV burden, histone marks) vary smoothly along chromosomes.
- Classical HMMs detect domains but suffer from K^2 transitions and label-switching.
- **oHMMed** fixes this by ordering states, restricting transitions to neighbours, and sharing one dispersion parameter.

Data sets

Arabidopsis thaliana GC content
5 chromosomes, 100 kb windows, GC % from TAIR10 FASTA.

TCGA-BRCA tumour SNVs
Patient TCGA-BH-A201, 100 kb windows across chr1–22; SNVs counted with `maftools`.

Do the tracks justify an oHMMed?

Variance of first-differences vs. a shuffled control shows strong autocorrelation:

Dataset	$\hat{\sigma}^2_{\text{orig}}$	$\hat{\sigma}^2_{\text{shuff}}$
GC %	2.8×10^{-4}	6.1×10^{-4}
SNV	36.2	50.2

F-tests: $p < 2.2 \times 10^{-16}$ in both cases.

The oHMMed model

Hidden chain. Reversible, tridiagonal T with only $2K - 2$ free transition parameters.

Emissions (convex families with shared scale).

$$y_l \mid \theta_l = i \sim \begin{cases} \mathcal{N}(\mu_i, \sigma) & \text{GC \% (continuous)} \\ \text{Gamma–Poisson}(\alpha, \beta_i) & \text{SNV counts} \end{cases}$$

A common σ or α makes log-density ratios convex, so states can be ordered.

Neighbour rule. $T_{ij} = 0$ if $|i - j| > 1$, preventing unrealistic long-range jumps.

Graphical Representations of the model

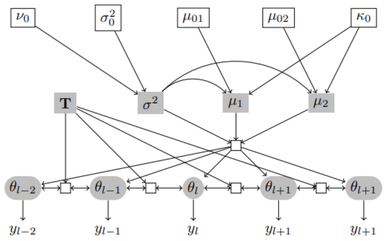


Figure 1: Graphical representation of oHMMed with normal emission densities: Starting at the top of the graph, the hyperparameters from which the emitted densities are drawn must be set: The variance σ^2 is assumed to be drawn from a prior inverse chi-square distribution with prior degrees of freedom ν_0 and prior variance σ_0^2; the means are assumed to be drawn from a normal distribution with variance σ^2 (since we set the coupling constant $\kappa_0 = 1$) and prior means $\boldsymbol{\mu}_0$. The priors for the transition matrix $\mathbf{T}$ are omitted here. The sequence of hidden states θ_l near the bottom of the graph each 'select' a μ_i, and conditional on the chosen μ_i and σ^2, the emitted/observed sequence of data points are drawn from the respective normal distribution.

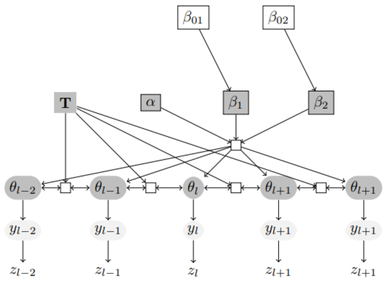


Figure 2: Graphical representation of oHMMed with gamma-poisson emission densities: Starting at the top of the graph, the hyperparameters from which the emitted densities are drawn must be set: The rate parameters β_i are assumed to be drawn from exponential distributions with priors β_{0i}. The improper prior for the shape parameter α cannot be properly shown and the priors for the transition matrix $\mathbf{T}$ are omitted.The sequence of hidden states θ_l near the bottom of the graph each 'select' a β_i, and conditional on the chosen β_i and α, the emitted sequence of data points y_l are drawn from the gamma normal distribution. These are rate parameters, which are used to draw the sequence of poisson distributed observed data points z_l.

Model Parameters selection

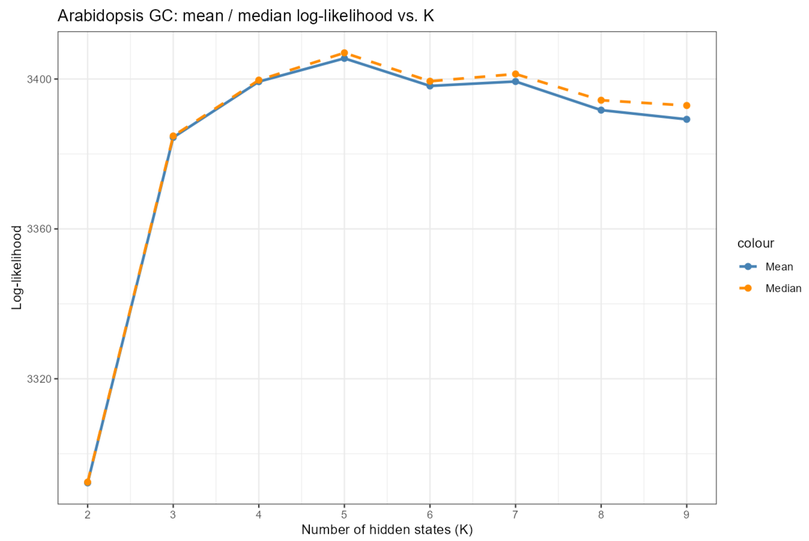
Gibbs sampler (8 k iters, 25 % burn-in)

1. Forward–Backward sample the full path $\theta_{1:L}$.
2. Update $\{\mu_i\}$ or $\{\beta_i\}$; sort to enforce order.
3. Update the tridiagonal T from its Dirichlet posterior.

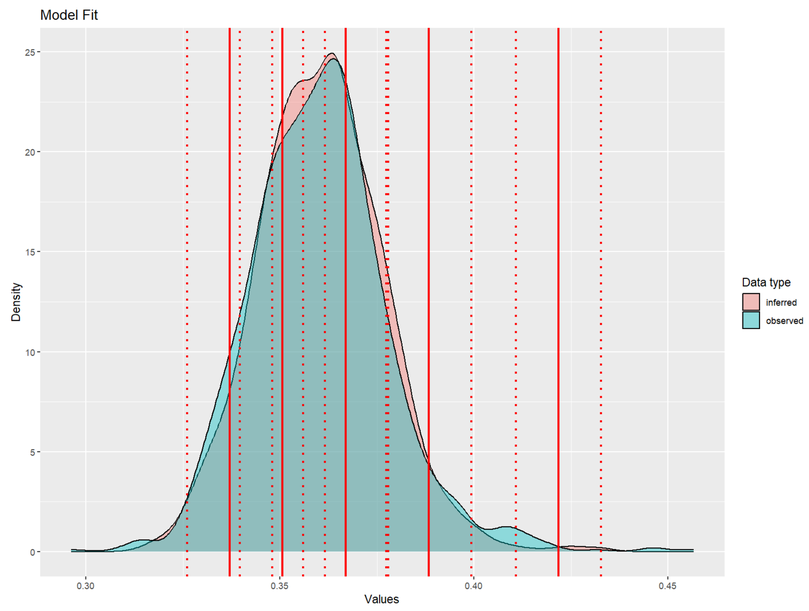
Ordering eliminates label-switching and speeds convergence.

Selecting the number of states

Fit $K = 2, \ldots, 9$ and pick the elbow of the median log-likelihood curve. Both data sets favour $K = 5$.



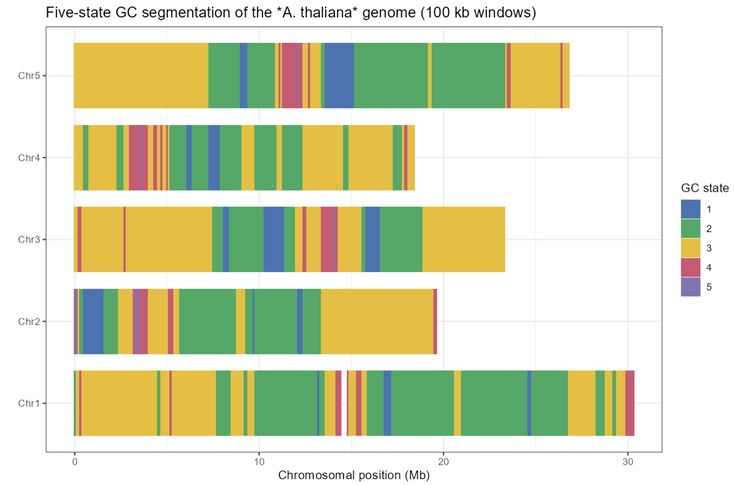
Log-likelihood vs. K



The observed overall density (blue) of the GC proportion superimposed on the posterior (inferred) density, with the inferred means per chosen number of states plus the 68% confidence intervals drawn in vertical lines.

A. thaliana GC segmentation

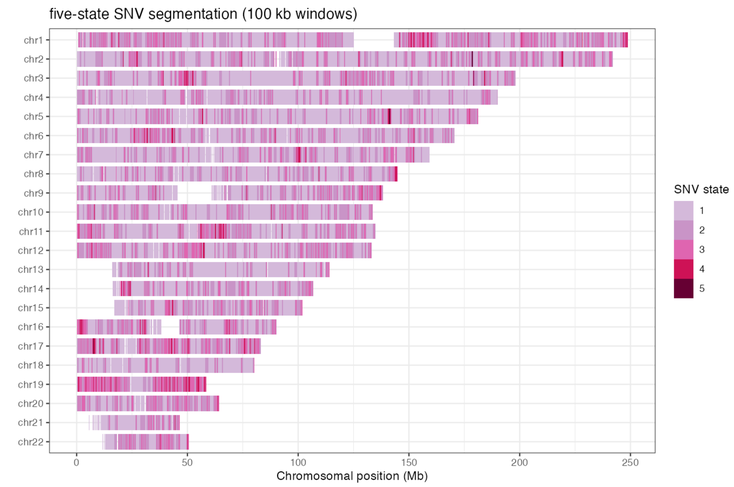
- Five domains with mean GC 0.337 $\rightarrow$ 0.415 and shared $\sigma = 0.011$.
- Recovers AT- and GC-isochores with finer-scale structure compared to window-less Z-curveZhang&Zhang; centromeres are GC-rich.



State track across chromosomes 1–5

Breast-cancer mutation landscape

- Gamma–Poisson, $K = 5$; 15 % of windows fall in high-burden states.
- Hyper-mutation on chr17 & chr19.



SNV states (1 = light … 5 = dark)

Conclusion & Outlook

- **Compact**: $2K - 2$ transitions.
- **Stable**: no label-switching.
- **Interpretable**: ordered emissions $\rightarrow$ natural biological ranking.
- **Versatile**: same core handles continuous or count data; useful wherever tracks are autocorrelated.
- **Future**: relax shared-dispersion.

Key references

Vogl et al., BMC Bioinf 25:151 (2024)
Alkouri & Godin, Tech. Report (2025)
Zhang & Zhang, J Mol Evol 59:227 (2004)