Quantifying the behavioral dynamics of *C. elegans* with autoregressive hidden Markov models

E. Kelly Buchanan
Columbia University
ekellbuchanan@gmail.com

Akiva Lipshitz
Columbia University
aclscientist@gmail.com

Scott Linderman
Columbia University
scott.linderman@columbia.edu

Liam Paninski
Columbia University
liam@stat.columbia.edu

Abstract

In order to fully understand the neural activity of *Caenorhabditis elegans*, we need a rich, quantitative description of the behavioral outputs it gives rise to. To this end, we quantify the behavioral dynamics of the worm with autoregressive hidden Markov models (AR-HMMs), a class of models that has recently yielded some insight into mouse behavior [1]. These models explicitly encode three hypotheses: (i) while the instantaneous posture of the worm is represented as a high-dimensional vector of points along the body, the first four principal components, or *eigenworms*, capture a significant fraction of the postural variance; (ii) within this four dimensional space, the postural dynamics are well-approximated with linear autoregressive models; and (iii) the linear autoregressive model switches over time as the worm transitions between different discrete behaviors, like forward crawling, reverse crawling, pausing, and turning. We show how AR-HMMs segment recordings of freely crawling *C. elegans* into meaningful discrete behaviors, providing a quantitative description of postural dynamics and a rigorous framework for assessing, comparing, and simulating worm behavior.

1 Introduction

Caenorhabditis elegans is a model organism for neuroscience. We know its complete connectome [2] and, with recent developments [3, 4], we can optically image nearly all of its neurons as it freely explores its environment. However, in order to fully understand this neural activity, we need a quantitative description of its behavioral outputs. We provide such a description with autoregressive hidden Markov models (AR-HMMs), which decompose behavior into a sequence of discrete latent states, or *syllables*: basic units of behavior that are strung together in time to create smooth behavioral trajectories. These are popular methods for unsupervised behavioral analysis [1] of freely behaving mice; we find they are also well-suited to worm behavior.

We begin with measurements of the worm's body posture as it crawls around its environment. At each point in time, the posture is represented as a sequence of tangent angles at evenly spaced points along the worm's body. Following [5], we project this posture onto the first four principal components to obtain vectors $x_t \in \mathbb{R}^4$ for each time frame t. As these vectors evolve, the posture is represented as a four-dimensional time series. Since the worm is roughly sinusoidal, the first two principal components are approximately a sine and a cosine wave. The projections onto the first two principal components trace out a ring of fixed radius, and the angular position along the ring encodes the "phase" of the worm's posture. This preprocessing pipeline is illustrated in Figure 1A. Our goal is to model the temporal dynamics of these points in PCA space.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

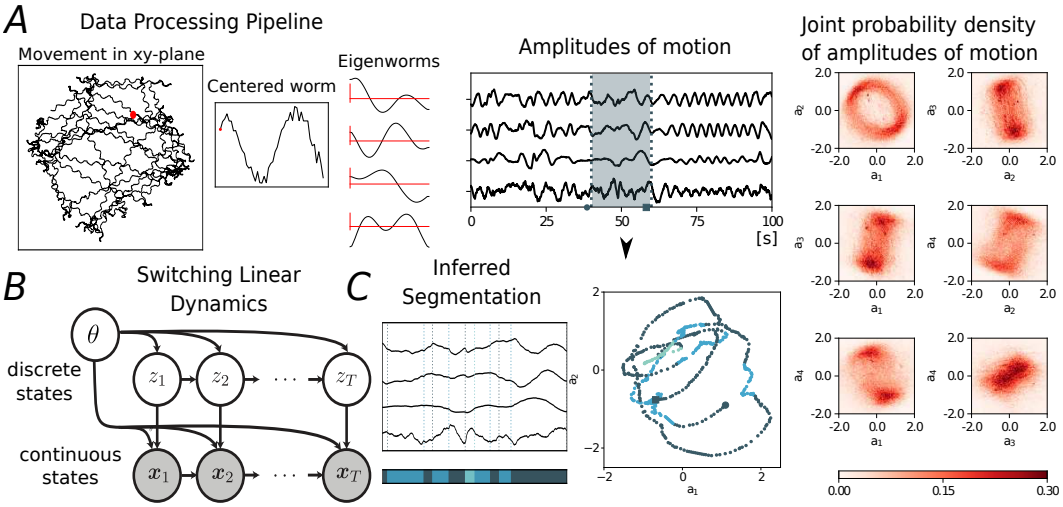

Figure 1: (A) Summary of data preprocessing pipeline. (B) Graphical model of an AR-HMM. (C) Model output: inferred segmentation, with each color representing a different state.

2 Autoregressive hidden Markov models (AR-HMMs)

We model behavior as a sequence of discrete states, or *syllables*, like forward crawling, reverse crawling, turning, and resting. Each syllable is modeled as an autoregression linking x_t to x_{t+1}. When the worm switches from one syllable to another, it employs a different set of autoregressive dynamics. This is formalized with an AR-HMM.

At each of the T time steps we have a discrete latent state $z_t \in 1, 2, ...K$ that follows Markovian dynamics,

$$z_{t+1} \,|\, z_t, \{\pi_k\}_{k=1}^K \sim \pi_{z_t} \tag{1}$$

where $\{\pi_k\}_{k=1}^K$ is the Markov transition matrix and $\pi_k \in [0,1]^K$ corresponds to the k-th row. Given discrete state z_t, the postural dynamics are given by,

$$x_t \,|\, x_{t-1}, z_t \sim \mathcal{N}\left(A_{z_t} x_{t-1} + b_{z_t},\, Q_{z_t}\right) \tag{2}$$

where $A_k, Q_k \in \mathbb{R}^{4\times 4}$ are linear dynamics and covariance matrices, respectively, and $b_k \in \mathbb{R}^4$ is a bias. The complete parameter set is given by $\theta = \{(\pi_k, A_k, Q_k, b_k)\}_{k=1}^K$.

Fig. 1B shows the graphical model of the AR-HMM. The discrete states follow Markovian dynamics and the instantaneous posture x_t (continuous state) depends only on the preceding posture and the current discrete state. The global parameters θ include the discrete state transition matrix and the autoregressive dynamics parameters. Fig. 1C shows a segment $x_{1:T}$ with the inferred discrete states $z_{1:T}$ below.

We fit the AR-HMMs with 1000 iterations of Gibbs sampling, alternating between sampling the discrete latent states $z_{1:T}$ and the global parameters θ from their conditional distributions. To learn the number of states, we fit AR-HMMs for different values of K and evaluate the likelihood of held-out data $p(x_{\text{test}} \,|\, \theta)$. While the held-out likelihood often continues to increase as we add more discrete states, the benefits tend to diminish as similar states are subdivided. We will show this in our results.

Based on our understanding of *C. elegans*, we believe that syllables tend to persist for more than just a few frames. We encode this belief in the form of a *sticky* prior on the rows of the transition matrix,

$$\pi_k \sim \text{Dir}(\alpha 1_K + \kappa e_k), \tag{3}$$

where 1_K is a length-K vector of all ones and e_k is a length K unit vector with a one in the k-th position. The hyperparameters $\alpha > 0$ and $\kappa > 0$ control the "concentration" and "stickiness," respectively. Larger values of κ bias the prior toward transition matrices closer to the identity matrix. This in turn prefers discrete state sequences with longer duration states. To set the stickiness parameter, we use manually-labeled state sequences to obtain a baseline for how long certain highly-identifiable states, like reverse crawling, should last.

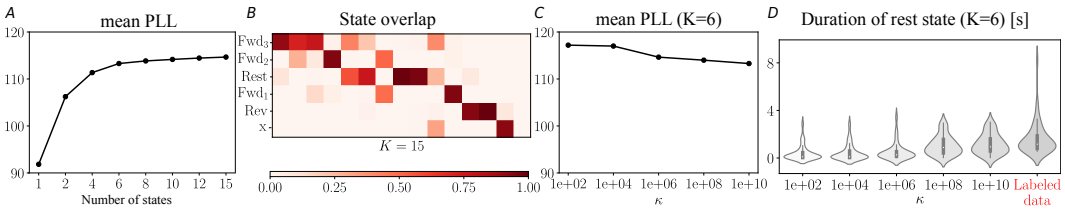

Figure 2: (A) Predictive log likelihood with respect to multivariate normal distribution of AR-HMMs with increasing numbers of states and hyperparameters $\alpha = 10$ and $\kappa = 10^{10}$. (B) Overlap between state assignments in the $K = 6$ and $K = 15$ models for a single worm. The $K = 6$ states are labeled as in Fig. 3, with 'X' denoting the one state that was used less than 1% of the time. (C) Mean predictive log likelihood for $K = 6$ for different values of κ. (D) Histogram of state durations as a function of κ, with the manually labeled data on the right for reference.

3 Results

We studied the dataset of worm behavior from Yemini et al. [6]. From this dataset, we chose nine worms with wild-type N2 alleles recorded at the same temporal scale. The postural coordinates lasted 15 minutes sampled at 30 frames per second and were preprocessed as described in [5] and in Fig. 1. We then split the data into contiguous train (70%) and test (30%) sets. As a baseline, we evaluated the held-out likelihood of the test data under a simple multivariate Gaussian model $x_t \sim \mathcal{N}(\mu, \Sigma)$ without any discrete states or autoregressive dynamics. We used maximum likelihood estimates of the mean and covariance.

We swept through different hyperparameter settings and found that the held-out likelihood was relatively insensitive to the concentration parameter α (data not shown). We focused our analysis on the effect of the number of states K and the stickiness κ.

Fig. 2A shows the held-out log likelihood for increasing numbers of discrete states K at a fixed intermediate value of κ. Here, held-out log likelihood is measured in terms of nats per time-frame improvement over the baseline Gaussian model. We see diminishing returns as we go beyond $K = 6$ states, which leads us to choose the six state model.

This choice is further substantiated by Fig. 2B, which shows the overlap in discrete states between the $K = 6$ and $K = 15$ models. To produce this matrix, we first counted how many times the 6-state model was in state k when the 15-state model was in state k' for all pairs (k, k'). Then we normalized the columns to sum to one. We we see that as K increases, states are split into substates. Upon further investigation, we found that these splits are largely based on differences in crawling velocity. This will be shown shortly.

We fit AR-HMMs with values of κ logarithmically spaced between 10^2 and 10^{10}. Large values of the stickiness parameter introduce a strong preference for self-transitions and penalize switches between different states. In terms of held-out likelihood, this stickiness hurts our performance, as shown in Fig. 2C. We see a decrease of about 3 nats per time-frame over this range of κ.

Stickiness, however, plays an important role in producing more realistic state durations. We may be willing to trade a 3 nat decrease in predictive performance for a segmentation that recapitulates the known time-scales of behavior. To formalize this, we used the empirical distribution of state durations in the manually-segmented states of Yemini et al. [6] as a baseline. This is shown as a "violin plot" in Fig. 2D (right). We choose $\kappa = 10^{10}$ as this yields a distribution of discrete state durations that best matches the manual data.

Having set the model hyperparameters (K, κ, α), we now interrogate the inferred state sequences and behavioral dynamics. We visualize the dynamics parameters (A_k, b_k) for discrete state k with the vector plots in Fig. 3A. Each arrow in this vector field shows where a point x_t would be mapped to under the linear map $A_k x_t + b_k$. While these dynamics are applied in 4-dimensional space, we show their projection onto only the first two principal components. We show the five of the $K = 6$ states that were used in at least 1% of time bins. Furthermore, we draw arrows only at the points where state k was actually used. Thus, the arrows lie around a ring in the first two principal components in accordance with the postural constraints of the worm (refer back to Fig. 1A).

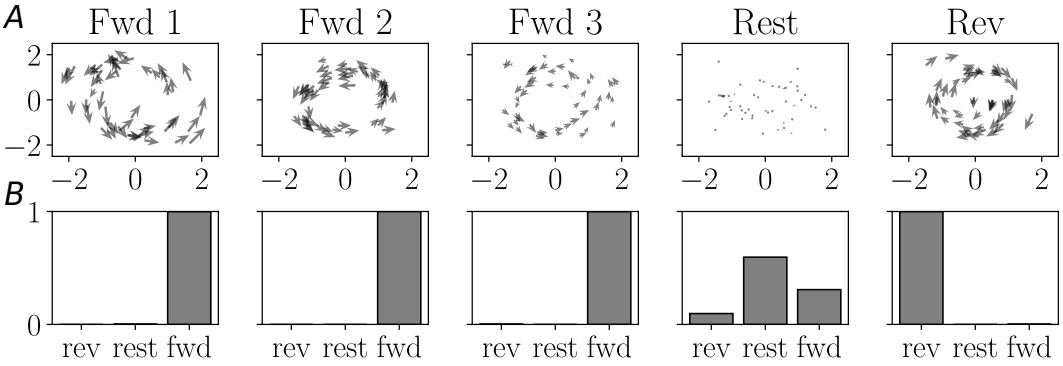

Figure 3: (A) Vector plot dynamics for a given worm for x_{t1} and x_{t2}. (B) Distributions of the manually label states, classified as forward, reverse and rest.

The first thing we notice is that three of the states ($k \in \{1, 2, 3\}$) correspond to counter-clockwise rotations in the space of the first two principal components. Moreover, they correspond to rotations with different angular velocity (length of arrows). When we compare these to the manually labeled states of Yemini et al. [6], we find that these three states are examples of forward crawling at different speeds, as shown via the probability distributions in Fig. 3B.

Likewise we find that our inferred state $k = 5$ corresponds to reverse crawling in the manual labels. In PCA space, this is a clockwise rotation. Referring back to Fig. 2B, we see that the $K = 15$ model eventually splits state 5 into more states, but forward crawling states are subdivided much more dramatically. This could be because reverse crawling has a smaller range of velocities, or because it is a less common behavior. Furthermore, we see that state 4 is most often manually labeled as a rest state. State 6 was excluded since it was rarely employed ($< 1\%$). We are continuing to investigate these states and their relation to higher order components for evidence of turning related behavior.

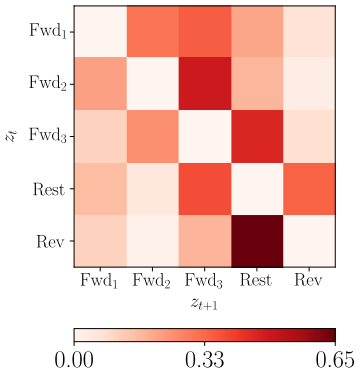

Figure 4: Transition matrix.

The transition matrix illustrated in Figure 4 represents the probability of going from state z_t to state z_{t+1}. We have subtracted the diagonal and re-normalized the rows to better visualize the relative differences between probabilities. The forward-crawling states transition to other forward states with different velocities, or to the rest state. The resting state goes to forward and reverse states with low velocities. Reverse goes back to rest with high probability. Comparing this to the neural state transitions inferred by Kato et al. [3], we see a similar pattern of transitioning between forward and reverse via an intermediate slow/pause state. We do not find strong dorsal or ventral turning states in our models, but we expect that these will become apparent as we increase the amount of data and include better preprocessing to resolve coiled worm postures [7].

4 Conclusion

Quantitative generative models of behavior provide many key features for systems neuroscientists: the ability to simulate and thereby evaluate models of behavior; an interpretable and unsupervised decomposition of complex, high-dimensional behavioral recordings; and a set of parameters that may have measurable neural bases. We are exploring autoregressive hidden Markov models as a general-purpose framework for quantifying the behavior of *C. elegans*, and the initial results presented here suggest that this is indeed a promising route toward understanding the behavior of this important model organism.

References

[1] Alexander B Wiltschko, Matthew J Johnson, Giuliano Iurilli, Ralph E Peterson, Jesse M Katon, Stan L Pashkovski, Victoria E Abraira, Ryan P Adams, and Sandeep Robert Datta. Mapping Sub-Second structure in mouse behavior. *Neuron*, 88(6):1121–1135, December 2015.

[2] Lav R Varshney, Beth L Chen, Eric Paniagua, David H Hall, and Dmitri B Chklovskii. Structural properties of the Caenorhabditis elegans neuronal network. *PLoS Comput. Biol.*, 7(2):e1001066, February 2011.

[3] Saul Kato, Harris S Kaplan, Tina Schrödel, Susanne Skora, Theodore H Lindsay, Eviatar Yemini, Shawn Lockery, and Manuel Zimmer. Global brain dynamics embed the motor command sequence of Caenorhabditis elegans. *Cell*, 163(3):656–669, October 2015.

[4] Jeffrey P Nguyen, Frederick B Shipley, Ashley N Linder, George S Plummer, Mochi Liu, Sagar U Setru, Joshua W Shaevitz, and Andrew M Leifer. Whole-brain calcium imaging with cellular resolution in freely behaving Caenorhabditis elegans. *Proceedings of the National Academy of Sciences*, 113(8):E1074–E1081, 2016.

[5] Greg J Stephens, Bethany Johnson-Kerner, William Bialek, and William S Ryu. Dimensionality and dynamics in the behavior of C. elegans. *PLoS Comput. Biol.*, 4(4):e1000028, April 2008.

[6] Eviatar Yemini, Tadas Jucikas, Laura J Grundy, André E X Brown, and William R Schafer. A database of Caenorhabditis elegans behavioral phenotypes. *Nat. Methods*, 10(9):877–879, September 2013.

[7] Onno D Broekmans, Jarlath B Rodgers, William S Ryu, and Greg J Stephens. Resolving coiled shapes reveals new reorientation behaviors in C. elegans. *eLife*, 5:e17227, 2016.