Statistical Analysis and Modeling of the Behavioral Response to Humidity and Olfactory Cues by the Vinegar Fly *Drosophila melanogaster*

May, 2021

Jonathan Lind

Master's Thesis in
Biomedical Engineering
Supervisors: Anders Enjin, Ola Jakobsson

LUND
UNIVERSITY

Dept. of Experimental Medical Science,
Faculty of Medicine
Dept. of Biomedical Engineering,
Faculty of Engineering

1 Abstract

Animals are impacted by the humidity of their surrounding environment, it affects their ability of thermoregulation, water retention and their overall well-being. For small insects the effects of the surrounding humidity is even more important as they have a limited water storage which depletes faster in a drier environment. Some insects such as the vinegar fly *Drosophila melanogaster* have developed the ability to sense humidity (hygrosensation) and navigate using humidity cues. Much research has contributed to our understanding of hygrosensation but the process is still not fully understood. In this thesis a framework is developed which is able to detect and visualise results from behavioural assays. The framework consists of a statistical analysis of trajectory data and a Gaussian mixture Hidden Markov Model (HMM) which simulates fly locomotion. A behavioural assay is also conducted in this thesis, in which vinegar flies are subjected to variations in humidity and their resulting trajectories are measured. The developed framework is applied to the experiment data in order to investigate hygrosensation. However as the reaction from flies was insufficient in the experiments of this thesis, the HMM is also applied to an external data set from an assay investigating olfaction, where distinct responses to stimuli were found. Furthermore when separate HMMs were fitted to trajectory data from different responses, the resulting model fits showed notable differences. Trajectories simulated from each model also differed, where one model showed flies moving towards the stimuli source at an increased speed and another showed flies moving sporadically with a reduced speed. The developed framework could be applied for analyzing further experiments investigating hygrosensation and serve as a starting point for analyzing trajectory data in general.

2 Collaborations and acknowledgements

I worked on this project in collaboration with Master's student Kalle Andersson and we have divided the project by focusing on different areas. This thesis presents the preprocessing, statistical analysis and modelling of flies reacting to fast changes of humidity or odor stimuli. While his thesis will present the engineering behind the experimental setup along with an analysis of another data set. We both worked on the experiment setup, but it was Kalle's main focus. He also single-handedly developed certain parts of the experiment setup that will be mentioned throughout the thesis.
Therefore, a special thank you goes out to Kalle Andersson for his collaboration and support throughout the thesis. This thesis would also not have been possible without the involvement of several people. My supervisor Anders Enjin, who spent a significant amount of time helping set up experiments, explaining neurophysiological concepts and giving advice throughout the thesis. My supervisor Ola Jakobsson for assisting with Labview. Kristina Corthals, for her input on Hidden Markov models and clustering. Also Ganesh Giri, Elton Melo, Johan Wall and Alrik Schörling for their input and support. Thank you.

Contents

Abbreviations

AIC-Aikake information criterion
BIC-Bayesian information criterion
D. melanogaster-Drosophila melanogaster
GMM- Gaussian mixture model
HMM- hidden Markov model
HRN- hygrosensory receptor neuron
IR-infrared
LiCl- lithium chloride
NaCl- sodium chloride

3 Introduction

3.1 Background

For small insects the temperature and humidity of the surrounding environment impacts survival, causing them to seek out optimal conditions. Thus insects have developed the ability to sense humidity (hygrosensation) and to navigate using humidity cues.[1] Much research has contributed to our understanding of hygrosensation but the process is still not fully understood. We know that humidity is sensed in a sensory organ called sacculus, by specialized sensory hairs called hygrosensilla.[2] Each hair containing three different sensory neurons that reacts to: increasing humidity, decreasing humidity and decreasing temperature. How these neurons collectively encode humidity cues and especially the possible contribution of the neuron reacting to temperature is uncertain.[2]

3.2 Research Target

The Sensory Neurophysiology group at the Department of Experimental Medical Science of Lund University researches hygrosensation in vinegar flies, *Drosophila melanogaster*. Their recent research suggests a model for hygrosensation where humidity cues are encoded by sensory neurons swelling with relative humidity and the strength of their responding activity is modulated by the temperature. In this thesis the hypothesised model will be investigated by conducting behavioural experiments, subjecting vinegar flies to brief variations in humidity and recording their movement. The potential reactions to humidity stimuli will be investigated by applying a framework of statistical analysis and modelling to the resulting trajectory data. The initial objective was to investigate hygrosensation and the possible contribution of the temperature neuron in three steps:

- Firstly to investigate if the vinegar fly shows a significant response to our humidity stimuli.
- Secondly to measure the difference in response at two temperatures.
- Finally to silence the temperature neuron in vinegar flies and measure the difference in response.

However, when no significant reaction to humidity could be detected from the experiments, the objective of this thesis was restricted to the first step. Comparing the difference in response between two groups requires that there is a significant response overall. Furthermore a general goal is also to see how statistical analysis and modelling can be applied to behavioural assays and to provide a framework for future work involving trajectory analysis.

In order to show how the developed model could simulate flies expressing different responses, it was also applied to an external data set from another study. This data set contained trajectory data from flies that reacted significantly to odor stimuli.

3.3 Outline

In the first part of the thesis an experiment setup for the behavioural assay is developed. The setup records and measures the trajectory data of vinegar flies moving in an arena while receiving an automated humidity stimulus.
The second part consists of preprocessing and analyzing the resulting trajectory data. Several features describing the characteristics of fly movement are calculated, and potential reactions are investigated as changes in feature values at stimuli onset and offset.
Lastly, the movement of flies is statistically modelled by implementing a Gaussian mixture hidden Markov model and fitting it to trajectory data. Inference about the fly behaviour can be drawn from studying the model fit and the responses from separate periods can be differentiated. Simulated trajectories can also be generated by sampling from a fitted model.

4 Theory

4.1 Neurophysiological background

This section presents the necessary theory regarding hygrosensation in vinegar flies. Both from the perspective of the underlying mechanisms of humidity sensing and how humidity stimuli translates into fly behaviour. Providing context to how the experiments in this thesis can deduce information about fly hygrosensation and the role of the temperature neuron.
In this thesis humidity levels are measured using relative humidity, which is the ratio between the amount of water vapor in the air and the amount of water vapor at saturation for a given temperature. The relative humidity thus increases with the amount of water vapor and decreases with temperature, as more water vapor is then required for the air to be saturated.[1]

4.1.1 Hygrosensation

When insects navigate through an environment they are able to use cues from multiple senses. They process olfactory information from odor plumes to find food sources, detect objects using visual and auditory cues and they are also able to sense the relative humidity of their surroundings. [3][4] The ability of humidity sensing is called hygrosensation, and is important for the insects survival. As a consequence of their small size and in relation large surface area, insects have a limited water storage that depletes relatively quickly when the surrounding air is dry. Therefore it is beneficial to sense and seek out optimal humidity levels. Insects are adapted to different humidity levels corresponding to their natural habitat, and species with the ability of hygrosensation are drawn to their preferred humidity level. Their preference also depend on their internal state; as a dried out fly is more likely to move towards humid areas.[2]

The vinegar fly *Drosophila melanogaster* provides a good model system to study hygrosensation as it is commonly used in experiments and it is genetically accessible. Meaning that it is possible to target and silence individual neurons involved in hygrosensation. The vinegar fly senses humidity in an invagination called sacculus, located on the third antennal segment. The sacculus consists of three chambers with walls covered by sensory hairs called sensilla. There are different forms of sensilla used for olfaction, termoreception and hygroreception (hygrosensilla).[2] Each hygrosensillum are coupled to three distinct hygrosensory receptor neurons (HRN) forming a hygrosensory triad, a structure also present in other humidity sensing insects.

The hygrosensory triad consists of one moist, one dry and one temperature HRN. The discharge frequency of each of the HRN increases by the corresponding actions: an increase of water vapor for the moist neuron, a decrease of water vapor for the dry neuron and decreased temperature for the temperature neuron.[2] The sacculus, hygrosensilla and HRN are illustrated in figure 4.1. It is unclear how the information from the HRN are combined to determine the humidity level, in particular the potential contribution of the temperature neuron.[1]

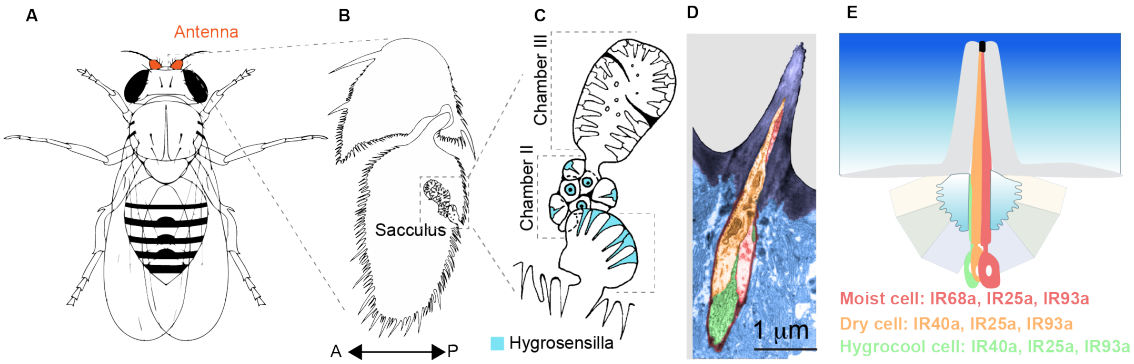


Figure 4.1: An illustration of the sacculus and hygrosensory neurons (image used with permission from Anders Enjin). (A) *D. melanogaster* with the antenna highlighted in orange. (B) A section of the antenna, displaying the location of the sacculus. (C) The internal structure of the sacculus, displaying the location of hygrosensilla. (D) A hygrosensory triad of neurons. (E) The dry, moist and temperature (hygrocool) neuron belonging to the hygrosensory triad with the ionotropic receptors (IR) expressed by each neuron.

4.1.2 Mechanisms behind hygrosensation

There are three major theories regarding humidity transduction: a mechanosensory model, a psychrometer model and an evaporation model. In the mechanosensory model the sensilla absorbs water vapor from the surrounding air, causing them to expand with rising relative humidity. The volume change of the sensillum causes mechanical stress in the neuronal membrane, sensed by ion channels known as mechanoreceptors, which encode the relative humidity into signals.[1]
In the psychrometer model humidity is measured by comparing the temperature difference between the sensillum surface and the surrounding air. The water in the sensillum evaporates into the surrounding air. Cooling the surface similar to how a body cools down from sweating, transferring thermal energy to the air. As drier air is able to absorb more humidity, the air humidity can be measured by the degree of the cooling effect.

In the evaporation model humidity is measured from changes in ion concentrations in the liquid surrounding the sensory neurons, the sensillum lymph. Water evaporates to the surrounding air, causing a concentration change in the lymph that is dependent to the saturation deficit of the air. The resulting concentration change will in turn activate ion channels on the HRN, encoding the humidity change into signals.[1]

The Sensory Neurophysiology group at Lund University has put forth an alternate hypothesis for the mechanisms behind hygrosensation, which the experiments in this thesis are designed to test. The research group proposes an extended mechanosensory model, where the moist and dry neurons react to humidity as in the previously described mechanosensory model but the strength of their response is modulated by the temperature neuron. Thus the moist and dry cells swell with increasing relative humidity, similarly to how a pine cone opens and closes with changes in air humidity. They then believe that the signal transduction is governed by ionotropic receptors which activates from mechanical stress. Additionally they believe that a decrease in temperature causes the temperature neuron to inhibit the activity of the moist and dry neuron.

Previous evidence against a mechanosensory model was that given a constant water vapor level, a rise in temperature, decreases the relative humidity. Therefore, the activity of the moist and dry neurons responding to humidity variations should be lower when increasing the temperature. But on the contrary studies have shown that the activity of the neurons increases with temperature.[1] However these models do not include the temperature neuron. The research group believes that the temperature neuron has an inhibitory effect on the two other neurons via ephaptic coupling. Meaning that since the neurons are bundled together and share the same surrounding extracellular fluid, their activities are dependent. The activity of one neuron, alters the ionic concentration in the surrounding fluid, which in turn alters the membrane potential and firing rate of the other neurons. A similar inhibitory effect has already been observed in olfactory receptor neurons of the vinegar fly.[5] The hypothesis is therefore that decreasing the temperature, rises the activity of the temperature neuron and thus inhibits the activity of the moist and dry neuron.
Studies suggest that the temperature neuron is silent for temperatures above the threshold of 25 degrees.[6] Therefore the experiment in this thesis will be performed at 24 and 26 degrees, above and below the suspected threshold. A reaction to humidity is suspected at both temperatures, with a stronger response at 26 degrees. When using genetically modified flies with a silenced temperature neuron the expectation is that the response will be equally strong at both temperatures.

4.1.3 Navigation using sensory cues

The behaviour of insect navigation induced by stimuli has been studied extensively.[2][7][3] Studies have shown that the humidity level preference of vinegar flies is relatively fine tuned. In one two-choice assay *D. melanogaster* was able to choose between two areas with different humidity levels during four hour experiments. A prominent trend was detected where flies actively moved towards areas with 70 % humidity and avoided areas with 20 % or 85 % humidity.[2] Furthermore, there are indications that flies placed in a circular arena react to a humidity increase in the center within tens of seconds.[7] This project will subject vinegar flies to a gradient of humidity where the humidity level is either increased or decreased for about twenty seconds, in order to elicit a reaction. The humidity gradient provides a finer scale of contrast in humidity compared to a two-choice assay. The flies should be able to sense the altered humidity, the question is how strong of a reaction the stimulus can evoke. Studies of vinegar fly navigation with a similar experiment setup have been conducted. These show that second long pulses of vinegar attracts starved flies [8] and that flies are able to navigate through complex plumes using olfactory cues.[3] Although hygrosenation and olfaction are two separated processes using different receptors, this shows that vinegar flies are able to find the source of an attractant by sensing cues from rapidly varying stimuli gradients.

One particular study of olfactory navigation by Alvarez et al.[8], served as an inspiration for this study. This thesis uses their arena design and inspiration has been drawn from their experimental setup and parts of their analysis. Their experiments consisted of minute long trials with constant wind flow, where odor stimuli in the form of a ten second long vinegar pulse was activated in each trial. The measured trajectory data is analyzed by investigating how features such as velocity, upwind speed (towards the source) and angular velocity vary during the course of trials. When calculating the mean feature values of all flies two prominent responses were found:

- ON-response during the stimuli onset, where flies tend to direct towards and move upwind against the gradient with an increased speed.
- OFF-response during the stimuli offset, where flies show local search behaviour, trying to find the attractant that disappeared. During the OFF-response the speed decreases while the angular velocity increases and the fly performs more sporadic turns. [8]

Flies subjected to humidity stimuli will possibly show similar responses if the stimuli acts as a powerful attractant.

4.2 Analytical background

This section presents the necessary theory for the analysis and modelling in this thesis. It first explains the Wilcoxon signed rank test which are used to determine if reactions are statistically significant. This section also explains the Hidden Markov model (HMM) that is implemented in this thesis. The model is deployed to simulate fly locomotion and further investigate differences in fly behaviour caused by stimuli, illustrating different characteristics of fly locomotion. The concepts behind the model are explained along with how the Baum-Welch algorithm is utilized to statistically determine a model that fits a given data set.

4.2.1 Wilcoxon signed rank test

This thesis deploys the Wilcoxon signed rank test in order to determine if a response is statistically significant. The Wilcoxon signed rank test compares if two samples $\mathbf{X} = [X_i; i = 1,...N]$, $\mathbf{Y} = [Y_i; i = 1,...N]$ share the same underlying distribution. Whereas other comparative tests assume the data to be normally distributed, this test is nonparametric and does not assume the type of the underlying distribution. The test is performed by first pairing randomly selected values from the two samples. The differences of all pairs are then calculated $\mathbf{D} = [D_i = X_i - Y_i; 1,...N]$ and the differences are ranked based on their magnitude. Given that the samples share the same underlying distribution, the median of the differences should be zero. The test thus calculates the probability that the median is not zero.[9]
One should note that performing several Wilcoxon signed rank test on the same data set introduces the problem of multiple comparisons. Meaning that as the number of comparisons increases it becomes more likely that some of them will be statistically significant. It is possible to compensate for this effect by adjusting the significance level α. A common method for adjusting α is the Bonferroni method where α is divided by the number of comparisons. [10]

4.2.2 Hidden Markov model

Modelling the movement of vinegar flies comprised the process of fitting a HMM to the fly trajectories of the experiment data. Previous work by Tao et al.[11], that fitted a hierarchical hidden Markov model to fly locomotion has served as an inspiration. Their model was able to differentiate several features from fly locomotion, such as when the fly was meandering or charging forward. The model could also find significant differences between the movement of individual flies and found that the presence of odor stimuli impacted the model fit.[11]

The HMM depends on two stochastic processes, here denoted $\{S(t)\}$ and $\{O(t)\}$. Where $\{S(t)\}$ is a Markov chain and $\{O(t)\}$ an observable process that depends on $\{S(t)\}$. These two concepts will be briefly explained to provide context about a HMM.
The resulting trajectory data from the experiments will be thought of as realizations of a discrete time stochastic process $\{O(t)\}$. A process with non deterministic properties of which a single sample is a time series of data. In comparison, realizations of a random variable $O(t)$ gives a single output, decided by its probability distribution.[12] Whereas the stochastic process is a family of random variables

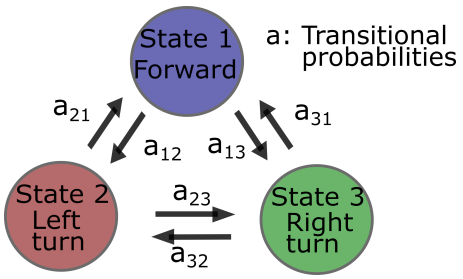
$$\{O(t), t \in T\} \tag{1}$$

Where the value of the process at t on the parameter space T (most often time) is given by a random variable $O(t)$. Thus an entire realization of a discrete time stochastic process is given by all values at points $t = 1 : T$. All possible realizations are contained in the sample space Ω.[12] In HMM applications the output of of the process is commonly denoted $\boldsymbol{o}$ (observations), with the observation at time t given as o_t.

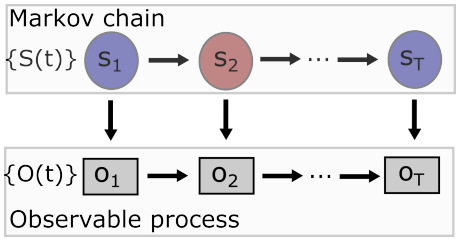
A Markov chain is a special case of a stochastic process that follows the Markov property. Here the output of the process are referred to as states $\boldsymbol{s}$. With the Markov property stating that the probability of the current state output of the process should only be determined by the state before, excluding the history of all other previous states.

$$P\left(S(n) = s_n \mid S(n-1) = s_{n-1}, \dots S(1) = s_1\right) = P\left(S(n) = s_n \mid S(n-1) = s_{n-1}\right) \tag{2}$$

If it is not intuitive that the Markov property holds, it will be used as an approximation.[13] This will be the case for the model in this thesis, where the impact of the history of the process is deemed insignificant. In reality the fly is complex enough to process information gained from various time points into its decision.[4]
Figure 4.2 (a) shows a simplified example of what the underlying discrete time Markov chain in this thesis model could be. For each time point the fly alternates between three states, either moving forward or turning left or right. Where the probabilities of transitioning between states is given by the transition matrix A.



(a)



(b)

Figure 4.2: (a) A simplified example of the underlying Markov chain $\{S(t)\}$, describing the movement of the fly. The fly switches between three states, either going forward, turning left or turning right. With transitional probabilities given by $A = (a_{ij})$.
(b) The corresponding HMM. The stochastic process $\{O(t)\}$ depends on the Markov chain $\{S(t)\}$, where the output o_t is different depending on the current state s_t. In this example, the fly starts in state one, thus o_1 will be trajectory data corresponding to the fly moving forward.

Together, the Markov chain $\{S(t)\}$ and the process behind the measured trajectory data $\{O(t)\}$, describe a HMM. Denoting all variables describing the model as λ. Here the process $O(t)$ generates our observations $\boldsymbol{o}$, following a distribution that depends on the states $\boldsymbol{s}$ of the unknown underlying process $\{S(t)\}$. Thus the likelihood of the observations depends on the conditional likelihood $f_{\boldsymbol{o}|\boldsymbol{s},\lambda}$ and the state probabilities $P_{s|\lambda}$.

$$f_{\boldsymbol{o}|\lambda}(\boldsymbol{o} \mid \lambda) = \sum_{\boldsymbol{s}} f_{\boldsymbol{o}|\boldsymbol{s},\lambda}(\boldsymbol{o} \mid \boldsymbol{s}, \lambda) P_{s|\lambda}(\boldsymbol{s} \mid \lambda) \tag{3}$$

As $\{S(t)\}$ is unknown so is also the states of the chain, along with the transition rates. However the observations, $\boldsymbol{o}$, provides information about the Markov chain and by analyzing them one can find a model for $\{S(t)\}$. Along with a relationship between the Markov chain states and the observations.[14]

Figure 4.2 (b) shows how the resulting HMM could look, continuing with the simplified example of an underlying three state Markov chain. The fly switches between a set number of states, altering its movement. However the fly does not show which state it occupies, and conclusions has to be drawn by studying $\boldsymbol{o}$. In reality, the number of states and the movement they describe is unknown.

4.2.3 Two-dimensional Gaussian Mixture Hidden Markov Model

This thesis will deploy an extension of the previously described model denoted as a Gaussian mixture Hidden Markov model. All parameters of the model and the data set are described in table 4.1. The model will be fit to a data set $\boldsymbol{o}$, consisting of sequences of two-dimensional observations. Each sequence, $\boldsymbol{o}^{(f)}$ contains two-dimensional data of measured speed, v and angular velocity, θ'. Each sequence of T data points is measured from a single fly during a specified time interval.

$$\boldsymbol{o} = \left[\boldsymbol{o}^{(1)}, \boldsymbol{o}^{(2)}, \ldots, \boldsymbol{o}^{(F)}\right]$$

$$\boldsymbol{o}^{(f)} = [o_t^{(f)} = [v_t, \theta_t']; \; t = 0, \ldots T]$$

The exact processing of the data set is described in sections 5.2, 5.4.

The probability of observations depends on the Markov states as shown in equation 3. For each state there is a different probability distribution from which observations are generated when the Markov chain occupies that state. Each probability distribution is modeled as a Gaussian mixture model, a weighted sum of Gaussian probability distributions. Where the likelihood of the arbitrary observation $o_t^{(f)}$ being generated from state i is given by

$$f_{\boldsymbol{o}|s,\lambda}(o_t^{(f)} \mid s_t = i) = \sum_{k=1}^{M} P_{ik} \mathcal{N}\left(o_t^{(f)}, \boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik}\right) \tag{4}$$

where

$$\mathcal{N}\left(o_t^{(f)}, \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk}\right) = \frac{1}{(2\pi)^{D/2}\left|\Sigma_{jk}\right|^{1/2}} e^{-\frac{1}{2}\left(o_t^{(f)} - \boldsymbol{\mu}_{jk}\right)^T \Sigma_{jk}^{-1}\left(o_t^{(f)} - \boldsymbol{\mu}_{jk}\right)}$$

Gaussian mixture models are commonly used for Markov models with continuous distributions. Whereas fitting a single Gaussian distribution also assumes that the observations belonging to a state are normally distributed, a mixture model with a sufficient number of components can approximate all sorts of distributions.[14] After fitting a HMM, the Gaussian mixture models will cover areas of our two-dimensional space where there are observations, as shown in figure 4.3 Fitting the model to a data set boils down to finding parameters for each Gaussian distribution along with stationary and transitional probabilities that fit the data set. The model fit should maximize the likelihood of the data being generated from the model.[14]

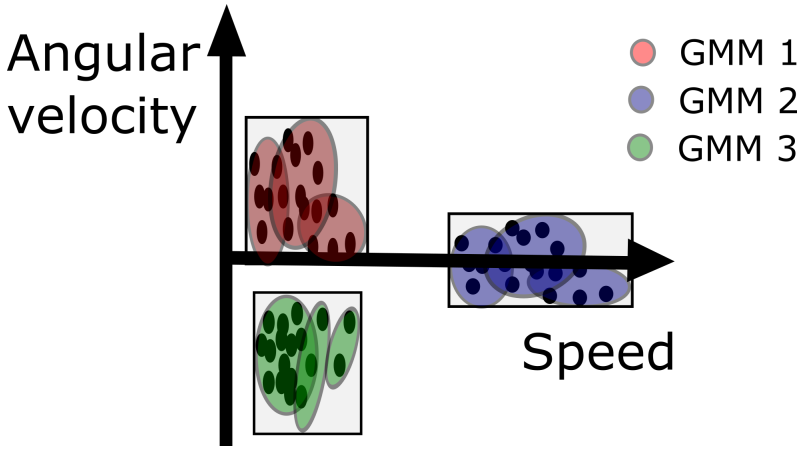


Figure 4.3: An illustration of a possible fit to the exemplary HMM from figure 4.2. The probability distributions of three Gaussian mixture models (red, blue, green), each corresponding to a state have been fitted to the data set (black dots). Here a positive angular velocity corresponds to turning left. Thus, depending on the occupied state, the output of the Gaussian mixture model results in a left turn, a right turn or going forward.

Parameter	**Description**
$s_t^{(f)} = i, i = 1:N$	The occupied state at time t, sequence f.
$\boldsymbol{s} = \left[\boldsymbol{s}^{(1)}, \boldsymbol{s}^{(2)}, \ldots, \boldsymbol{s}^{(F)}\right]$, $\boldsymbol{s}^{(n)} = [s_t^{(n)}; t = 0, \ldots T]$	The set of state sequences. Denoting the occupied state for each sequence and time point.
$\boldsymbol{o} = \left[\boldsymbol{o}^{(1)}, \boldsymbol{o}^{(2)}, \ldots, \boldsymbol{o}^{(F)}\right]$ $\boldsymbol{o}^{(f)} = [o_t^{(f)} = [v_t, \theta_t']; t = 0, \ldots T]$	The set of two-dimensional observations, each consisting of a measure of speed and angular velocity.
$\lambda = [A, \pi, P, \mu, \Sigma]$	The parameters describing a model fit.
$A = (a_{ij})$	The transition matrix. A_{ij}: probability of transitioning from state i to j.
$\pi = (\pi_i)$	The initial state distribution. Probabilities of the Markov chain starting in each state.
$f_{\boldsymbol{o}\mid\boldsymbol{s}}(o \mid s = i)$	The observation probability distribution in state i.
$\boldsymbol{\mu}_{s,m}, \boldsymbol{\Sigma}_{s,m}, \boldsymbol{P}_{s,m}$	The mean, covariance and mixture weight of the Gaussian belonging to state $\boldsymbol{s}$ and mixture m.
$\boldsymbol{F}, \boldsymbol{T}, \boldsymbol{N}, \boldsymbol{M}$	Number of: sequences, data points of a sequence, states, Gaussian mixture components.

Table 4.1: Notations describing the data set and all parameters of the model deployed in this thesis.

4.2.4 Fitting a Hidden Markov Model

When fitting a HMM the objective is to maximize the probability of the model given the sequences of observational data. This optimization is performed by varying $\lambda = [A,\pi,P,\mu,\Sigma]$, the parameters describing the model.[14] Algorithm 1 describes the fitting process, with the corresponding Matlab code in Appendix B (Fitting process). Using Bayes rule the model probability can be expressed as

$$P_{\lambda|\boldsymbol{o}}(\lambda \mid \boldsymbol{o}) = \frac{1}{f_{\boldsymbol{o}}(\boldsymbol{o})} f_{\boldsymbol{o}|\lambda}(\boldsymbol{o} \mid \lambda) P_{\lambda}(\lambda) \tag{5}$$

[15] The prior probability of the model $P_{\lambda}(\lambda)$ is chosen from an assumption and $f_{\boldsymbol{o}}(\boldsymbol{o})$ denotes the likelihood of the data without model assumptions. The prior probability is not dependent on $\boldsymbol{o}$ and the likelihood of the data serves merely as a normalizing effect. Thus maximizing the likelihood of the data set given the model parameters $f_{\boldsymbol{o}|\lambda}(\boldsymbol{o} \mid \lambda)$ will also result in maximizing the model probability.[15] The likelihood of the entire data set can in turn be written as a product of the likelihood of the individual sequences

$$f_{\boldsymbol{o}|\lambda}(\boldsymbol{o} \mid \lambda) = \prod_{f=1}^{F} f_{\boldsymbol{o}^{(f)}|\lambda}(\boldsymbol{o}^{(f)} \mid \lambda) \tag{6}$$

And the problem of finding the variables that maximizes this likelihood can be written as

$$\lambda^* = \arg\max_{\lambda}[f_{\boldsymbol{o}|\lambda}(\boldsymbol{o} \mid \lambda)] \tag{7}$$

[14]
When fitting the model, the individual likelihoods for all sequences will be calculated before using equation 6 for the total likelihood. Maximizing the likelihood $f_{\boldsymbol{o}|\lambda}(\boldsymbol{o} \mid \lambda)$ analytically is problematic, thus it is usually solved using a special case of a recursive Expectation maximization method, known as the Baum-Welch algorithm.
The Baum-Welch algorithm utilizes a set of variables to solve equation 7 which are briefly introduced in table 4.2. Calculating these correspond to the expectation step.[14]

Variable	Description
$\alpha_t^{(f)}(i) = f_{\boldsymbol{o},\boldsymbol{s}\mid\lambda}[o_1^{(f)},\ldots,o_t^{(f)}, s_t^{(f)} = i \mid \lambda]$	The forward probability of sequence f.
$\beta_t^{(f)}(i) = f_{\boldsymbol{o},\boldsymbol{s}\mid\lambda}[s_t^f = i, \boldsymbol{o}_{t+1}^{(f)},\ldots,\boldsymbol{o}_T^{(f)} \mid \lambda]$	The backward probability of sequence f.
$\gamma_t^{(f)}(i) = P_{\boldsymbol{s}\mid\boldsymbol{o},\lambda}[s_t^{(f)} = i \mid \boldsymbol{o}, \lambda]$	Probability of the current observation belonging to state i.
$\Gamma_t^{(f)}(i,j) = P_{\boldsymbol{s}\mid\boldsymbol{o},\lambda}[s_t^{(f)} = i, s_{t+1}^{(f)} = j \mid \boldsymbol{o}, \lambda]$	Probability of the current observation transitioning from state i to j.
$\zeta_t^{(f)}(i,k) = P_{\boldsymbol{s},\boldsymbol{K}\mid\boldsymbol{o},\lambda}[s_t^{(f)} = i, m_t^{(f)} = k \mid \boldsymbol{o}, \lambda]$	Probability of the current observation belonging to state i and mixture component k. Where m is a mixture component of the set K.

Table 4.2: Variables calculated using the Baum-Welch algorithm, and their statistical meaning. These variables are calculated individually for each sequence f. The formulas are lifted from [15].

Figure 4.4 provides an insight into how the Baum-Welch algorithm uses forward and backward probabilities to optimize the model. At each time point the underlying Markov chain can be in one of N possible states, resulting in sequences of states for $t = 1 : T$. Where the likelihood of our observations are impacted by the possible sequences of states. By exploiting the repetitive structure in figure 4.4, given that the likelihood of all observations up till time point t is known, the likelihood for the neighboring time points can easily be computed.
The forward probability $\alpha_t^{(f)}(i)$ is shown in table 4.2 as the likelihood of all observations up till time point t given that the state at t is i. The likelihood of the first observation becomes

$$\alpha_1^{(f)}(i) = \pi_i f_{\boldsymbol{o}\mid\boldsymbol{s},\lambda}(o_1^{(f)} \mid s_1^{(f)} = i)$$

[14] Figure 4.4 then shows that the likelihood of all observation up till the next time point is calculated as.

$$\alpha_{t+1}^{(f)}(i) = f_{\boldsymbol{o}\mid\boldsymbol{s},\lambda}(o_{t+1}^{(f)} \mid s_{t+1}^{(f)} = i) \sum_{i=1}^{N} \alpha_t^{(f)}(j) a_{ji}$$

[15]

Transversely, it is also possible to iterate backwards using the backward probability $\beta_t^{(f)}(i)$. The backward probability is shown in table 4.2 as the likelihood of all observations from the end T back to the time point $t+1$ given that the state at t is i. The likelihood of the end observations is set to one:

$$\beta_T^{(f)}(i) = 1$$

[14] Figure 4.4 then shows that the likelihood of the observations also including the previous time point becomes.

$$\beta_t^{(f)}(i) = \sum_{j=1}^{N} \beta_{t+1}^{(f)}(j) a_{ij} f_{\boldsymbol{o}|\boldsymbol{s},\lambda}(o_{t+1}^{(f)} \mid s_{t+1}^{(f)} = j)$$

[15]
α and β are used to calculate the rest of the variables in table 4.2, which in turn are used to update the model parameters λ. α also provides a way to calculate the likelihood of all observations, which is maximized by training the model. The likelihood of all observations in sequence f is given by

$$f_{\boldsymbol{o}^{(f)}|\lambda}(\boldsymbol{o}^{(f)} \mid \lambda) = \sum_{i=1}^{N} \alpha_T^{(f)}(i) \tag{8}$$

When training a model using the entire data set, the forward and backward probabilities along with the other variables in table 4.2 are calculated individually for all sequences $1:F$. However the model parameters λ are updated by weighing information from all sequences, as shown in algorithm 1. [15]

Furthermore, since γ denotes the probability of an observation belonging to a state, it gives a value of the confidence with which the model assigns a state to each observations. The amount of states assigned with a high confidence serves as an indication of how well a model fits the data set. Given the definition of γ in table 4.2, the overall state probability can also be calculated as an average of γ over all sequences and time points

$$P(S = i) = \frac{1}{TF} \sum_{f=1}^{F} \sum_{t=1}^{T} \gamma_t^{(f)}(i) \tag{9}$$

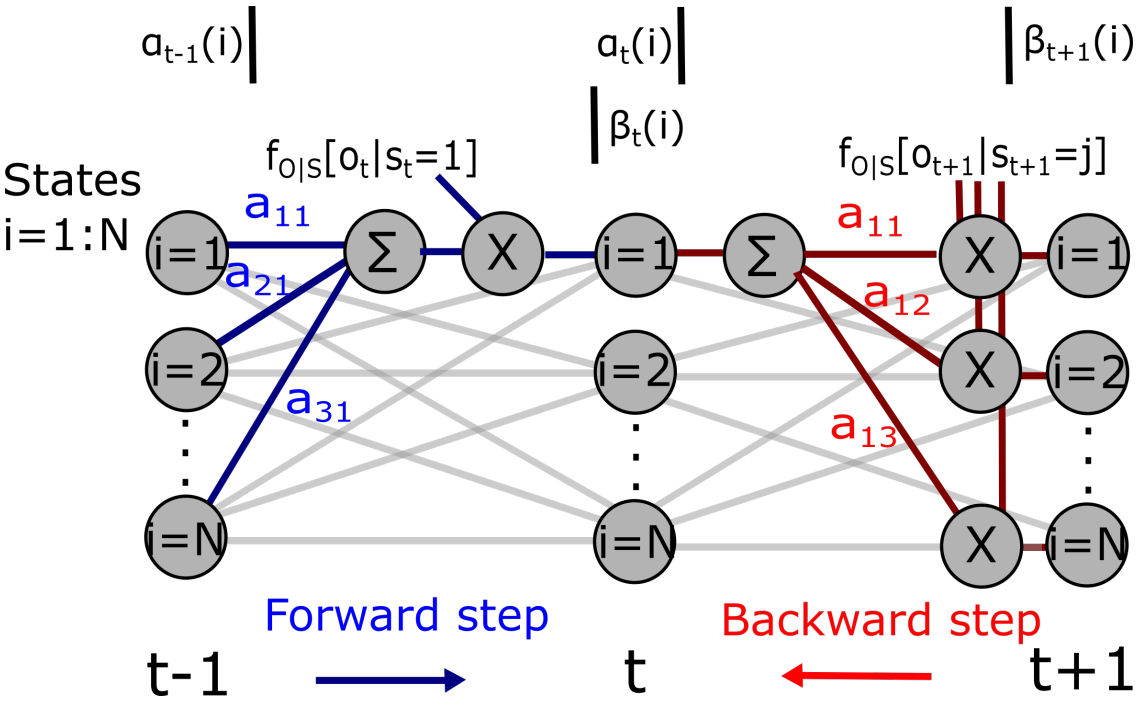


Figure 4.4: An illustration of calculating the forward probabilities (α) and backward probabilities (β), for one sequence of data (f). The gray connections show all possible sequences of states. In the forward step (blue) $\alpha_t(1)$ is updated from the forward probabilities of the previous step $\alpha_{t-1}(i)$. Calculated as a sum weighted by the transitional probabilities a_{ij} and multiplied by the likelihood of the added observation o_t. Thus computing the likelihood of all possible ways to reach state 1 at time t. In the backward step (red), the procedure is performed in reverse and $\beta_t(t)$ is computed from the backward probabilities of the next step $\beta_{t+1}(i)$.

Algorithm 1: The Baum-Welch algorithm deployed in this thesis for fitting a HMM to a given data set. The algorithm performs an expectation maximization procedure until the likelihood of the data set converges. In the expectation step the algorithm calculates a set of variables described in table 4.2. In the maximization step these variables are then utilized to update λ, the components that describe the HMM. The corresponding Matlab code is found in Appendix B (Fitting process).

Initialize $\boldsymbol{A}$,$\boldsymbol{\pi}$,$\boldsymbol{\mu}$,$\boldsymbol{\sigma}$,$\boldsymbol{P}$;
while $(P(\mathbf{o} \mid \lambda_{new}) - P(\mathbf{o} \mid \lambda_{old}))/P(\mathbf{o} \mid \lambda_{new}) > 0.0001$ **do**
  E-step
  Initialize α, β;
  for *all f,i:* **do**
    $\alpha_1^{(f)}(i) = \pi_i f_{\boldsymbol{o}|\boldsymbol{s},\lambda}(o_1^{(f)} \mid s_1^{(f)} = i)$
    $\beta_T^{(f)}(i) = 1$
  end
  for *all t,f,i:* **do**
    $\alpha_{t+1}^{(f)}(i) = f_{\boldsymbol{o}|\boldsymbol{s},\lambda}(o_{t+1}^{(f)} \mid s_{t+1}^{(f)} = i) \sum_{i=1}^{N} \alpha_t^{(f)}(j) a_{ji}$
    $\beta_t^{(f)}(i) = \sum_{j=1}^{N} \beta_{t+1}^{(f)}(j) a_{ij} f_{\boldsymbol{o}|\boldsymbol{s},\lambda}(o_{t+1}^{(f)} \mid s_{t+1}^{(f)} = j)$
  end
  for *all t,f,i,m* **do**
    $\gamma_t^{(f)}(i) = \frac{\alpha_t^{(f)}(i)\beta_t^{(f)}(i)}{\sum_{j=1}^{N} \alpha_T^{(f)}(j)}$
    $\Gamma_t^{(f)}(i,j) = \frac{\alpha_t^{(f)}(i) a_{ij} f_{\boldsymbol{o}|\boldsymbol{s},\lambda}(o_{t+1}^{(f)} | s=j) \beta_{t+1}^{(f)}(j)}{\sum_{i=1}^{N} \alpha_T^{(f)}(i)}$
    $\zeta_t^{(f)}(i,m) = \frac{\sum_{j=1}^{N} \alpha_{t-1}^{(f)}(j) a_{ji} P_{im} \mathcal{N}\left[o_t^{(f)}, \mu_{im}, \Sigma_{im}\right] \beta_t^{(f)}(i)}{\sum_{j=1}^{N} \alpha_T^{(f)}(j)}$
  end
  M-step
  Re-estimating λ:
  for *all i,m* **do**
    $\pi_i = \frac{1}{F} \sum_{f=0}^{F} \gamma_0^f(i)$
    $a_{ij} = \frac{\sum_{f=1}^{F} \sum_{t=1}^{T-1} \Gamma_t^f(i,j)}{\sum_{f=1}^{F} \sum_{t=1}^{T-1} \gamma_t^f(i)}$
    $P_{im} = \frac{\sum_{f=1}^{F} \sum_{t=1}^{T} \zeta_t^f(i,m)}{\sum_{f=1}^{F} \gamma_{t=0}^f(i)}$
    $\mu_{im} = \frac{\sum_{f=1}^{F} \sum_{t=1}^{T} \zeta_t^f(i,m) o_t^{(f)}}{\sum_{f=1}^{F} \sum_{t=1}^{T} \zeta_t^f(i,m)}$
    $\Sigma_{im} = \frac{\sum_{f=1}^{F} \sum_{t=1}^{T} \zeta_t^f(i,m) \left[\overline{o}_t^{(f)} - \overline{\mu}_{im}\right] \left[\overline{o}_t^{(f)} - \overline{\mu}_{im}\right]^{\mathrm{T}}}{\sum_{f=1}^{F} \sum_{t=1}^{T} \zeta_t^f(i)}$
  end
end
$\rightarrow$Final values of $(P(\mathbf{o} \mid \lambda^*), \lambda^*$

4.2.5 Initialization

Prior to fitting a HMM using the Baum-Welch algorithm, the model is initialized by finding a first estimate of $\lambda = [A,\pi,P,\mu,\Sigma]$. The initialization process is illustrated in figure 4.5. A and π are set to uniform probabilities, not favoring any state. However finding an initial estimate of the Gaussian mixture models requires more care. This initialization is performed by first segmenting the data into N states using a K-means clustering algorithm. Each observation is set to belong to one cluster, given by the index $k_t^{(f)}$. The cluster mean is given as the mean of all observations belonging to the cluster

$$\mu_n = \frac{1}{|k = n|} \sum_{k=n} o_t^{(f)} \tag{10}$$

The algorithm updates which observations belong to each cluster as

$$k_t^{(f)*} = \arg\min_n \left[d\left(o_t^{(f)}, \mu_n \right) \right] \tag{11}$$

with the distance d given by

$$d\left(o_t^{(f)}, \mu_n \right) = \sqrt{\left(o_t^{(f)} - \mu_n \right)^T \left(o_t^{(f)} - \mu_n \right)} \tag{12}$$

The algorithm performs an iterative optimization until conversion, where each step first updates k using equation 11 and then updates the cluster centers using equation 10. Thus minimizing the distance between all observations to the cluster means (the cluster variance).[14]

The initial parameter values for μ, σ, P are then estimated by fitting a Gaussian mixture model with M mixture components to the observations belonging to each cluster. The Gaussian mixture model is fitted using an iterative expectation maximization algorithm that updates the Gaussian mixture model variables (μ, σ, P) to maximize the likelihood of all observations belonging to the cluster, where the likelihood of one observation is given by equation 4.
Initialization using clustering before applying a HMM is commonly deployed when there is no prior information about the Markov chain states. It is of importance as the Baum-Welch algorithm converges to a local minimum and the resulting fit is dependent of the initialization.[14] When the model deployed in this thesis was initialized with clustering the resulting fit reached a higher likelihood compared to when using random initialization.

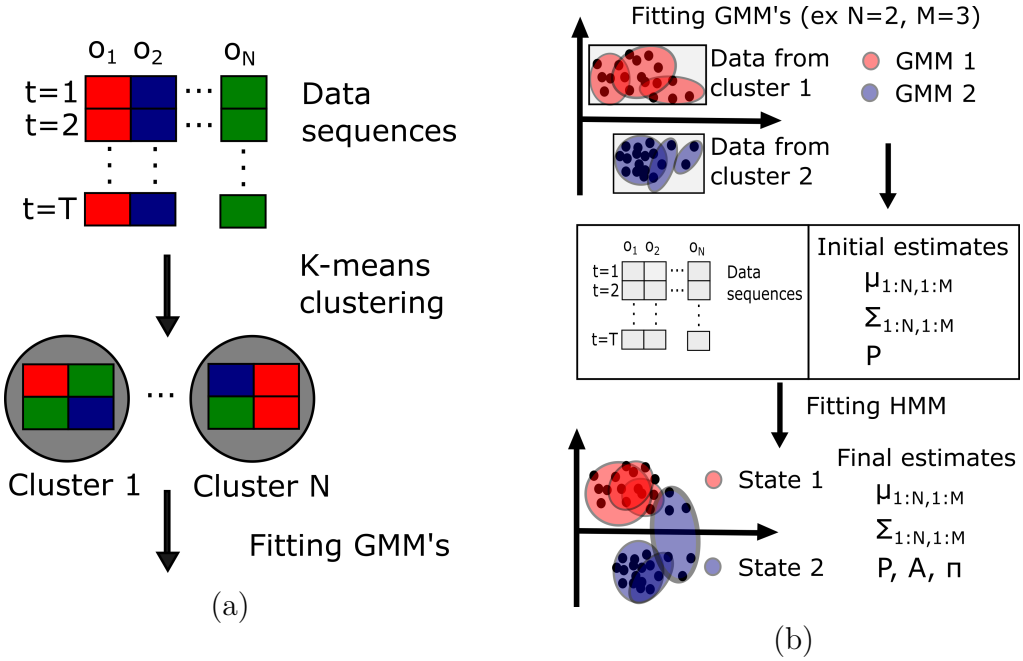


Figure 4.5: The initialization process deployed in this thesis. (a) The observations of the data set are divided into states using a K-means clustering algorithm. (b) A Gaussian mixture model is fitted to the observations in each cluster/state, giving an initial estimate of all Gaussian distributions. The Baum-Welch algorithm then fits a HMM using the data set and the initial Gaussian mixture estimates.

5 Materials and methods

5.1 Experimental protocol

This section explains the behavioural experiments conducted in this thesis. At the time of writing there is no ethical standards regarding the use of insects in scientific experiments and vinegar flies are commonly used in behavioural assays. This section describes the implemented experiment setup, the experiment procedure and the resulting data set. The experiment setup and arena design was inspired by a study from Alvarez et al.[8] The model names of the equipment and used programs is provided in table 8.1, Appendix A.

5.1.1 The experiment setup

The behavioural experiments in this thesis required an experiment setup capable of subjecting flies to controlled variations of humidity and measuring their responding movement. Thus the experiment setup illustrated in figure 5.1 was implemented.

Flies are placed in an arena, where humidity stimuli is supplied from an externally controlled pump which delivers a constant air flow, alternating between high and low humidity levels. The temperature inside the box is regulated and visual cues are removed as the box is lit using only infrared light, which is outside of the flies visible spectrum. The flies are filmed by a camera which captures the contrast of the black flies walking above the light from the IR-lamp.

The arena consists of four chambers measuring 14x4x0.17 cm (length, width, height), where one fly can walk freely in each chamber, allowing for tracking of four flies simultaneously. The arena is made out of several plastic layers that create the chambers when screwed together (the arena design is available in [8]). Each individual chamber has an inlet connected to the pump at the base of the arena and an outlet at the top end of the arena. This causes a humidity gradient when the pump is activated.

The humidity levels are altered by an airflow passing two conical flasks, which are filled with saturated salt solutions of Sodium Chloride (relative humidity of 70 %) and Lithium Chloride (relative humidity of 20 %). However, the humidity level of the air reaching the chamber is slightly different, as shown in figure 5.2. Flies are expected to show a preference for the higher relative humidity as they are dried out prior to the experiment and since it is closer to the humidity level of their natural habitat.[2]
The pump is remotely controlled using a Daq-board (data acquisition hardware which enables communication between a computer and instruments) which sends signals following the instructions of a Labview script (developed by Kalle Andersson). The script also produces an output of the pump state along with the time measure of the internal clock, which is used to synchronize the trajectory data to the corresponding humidity levels.

The camera output is measured in real time using the image analysis program Margo, which is specially designed to measure the trajectories of multiple insects.[16] The program is tuned to produce a positional value of each fly every tenth of a second. It is also modified to output the corresponding time from the internal clock, allowing for synchronizing the trajectory data to the pump output.
The wind speed in the arenas was estimated to 11 cm/s by filling a conical flask in the experiment setup with visible smoke and measuring the flow using an image analysis script (developed by Kalle Andersson). Increasing the wind speed further can cause flies to stop moving during the experiments.[8] The smoke-experiment also showed that the flow was evenly distributed throughout the arena.

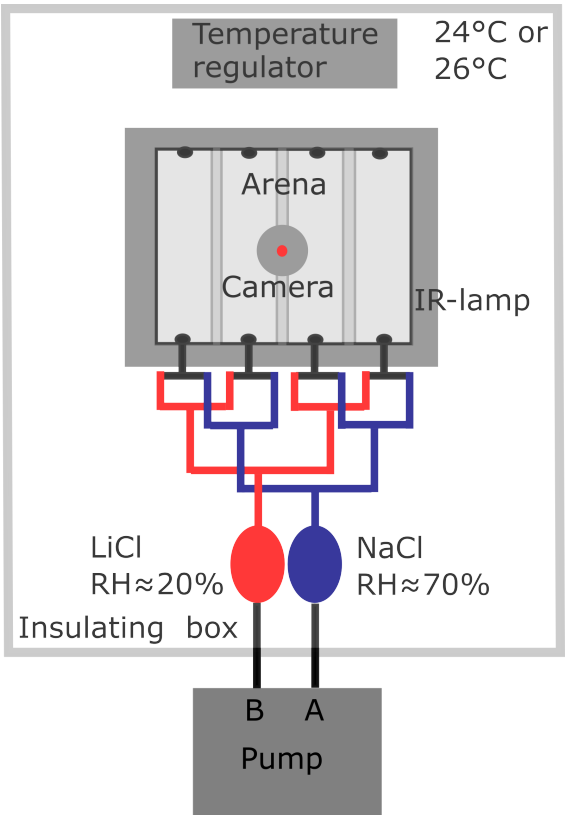


Figure 5.1: The experiment setup used for the behavioural experiments in this thesis. The wiring in red and blue shows how the pump is connected to each individual chamber inlet. Depending on the state of the pump, it provides an air stream of dry air (B, red) or humid air (A, blue).

5.1.2 Stimuli

The relative humidity and temperature in the arena was measured prior to the experiments using probes placed at the inlet and outlet of a chamber. The results are shown in figure 5.2. However, as the humidity levels were not continuously measured throughout the experiments, these expected humidity variations serve as an estimation. Experiments were conducted at both $24°C$ and $26°C$ in order to test the hypothesis regarding hygrosensation, described in section 4.1.2. It was crucial to minimize any potential temperature gradient inside the box as flies could also react to temperature variations. Figure 5.2 showed that the temperature difference caused by the pump switching states and the temperature difference between the probes was less than 0.25 $°C$, which was deemed insignificant. The temperature change when the pump switches state was deemed insignificant as it was of a magnitude of $0.2°C$, as shown in figure 5.2.

The humidity stimulus during a experiment is shown in figure 5.2. The five minute baseline of constant humidity serves as acclimatization. Whereas the following one minute trials switches between high and low humidity levels to elicit a reaction.

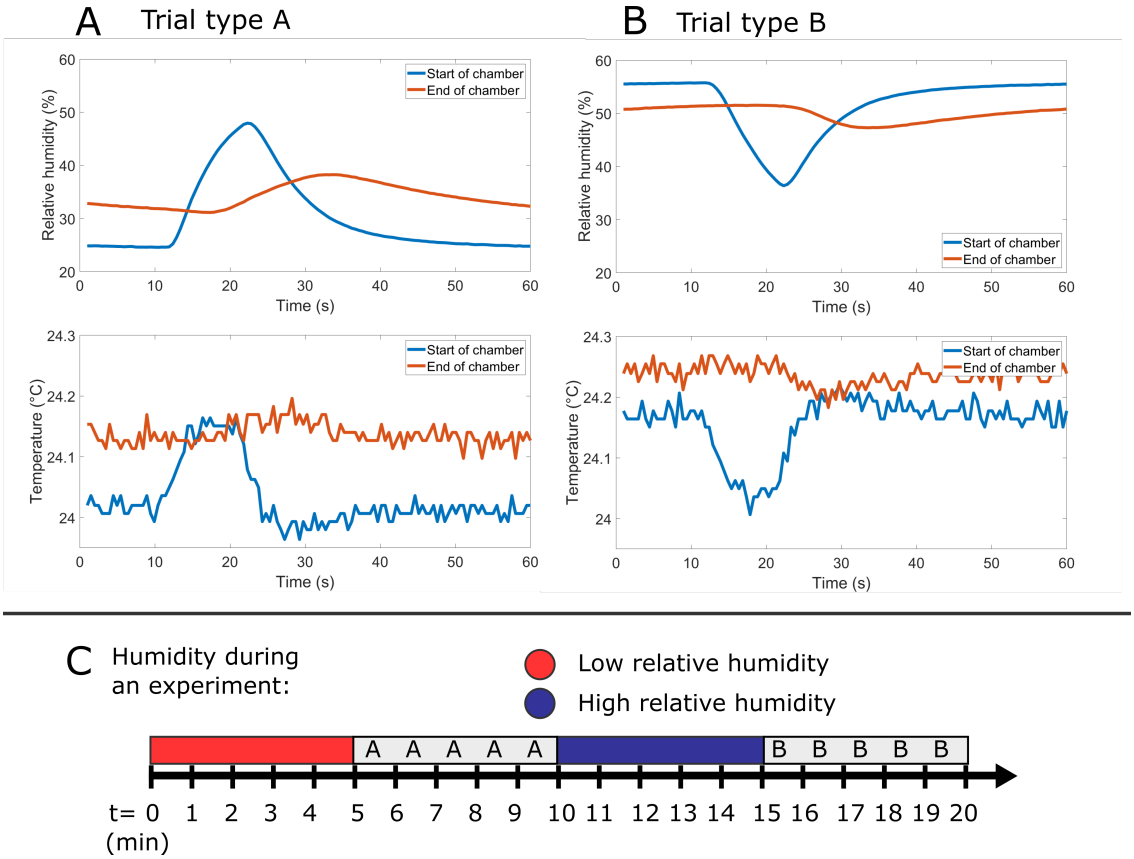


Figure 5.2: The humidity and temperature stimuli delivered to flies during an experiment. With (A) and (B) displaying the relative humidity and the temperature during trials conducted at 24 $°C$ (the relative humidity will differ slightly at 26 $°C$). With two types of minute-long trials were the humidity is either increased (type A) or decreased (type B) for a short duration. (C) The timeline showing how the relative humidity alters during the course of the experiment.

5.1.3 Fly preparation and experiment procedure

In total, the resulting data set from the experiments that was later analyzed consisted of trajectory data from 61 flies. Flies of the strain w^{1118CS} were used in the experiments. The initial plan was to first perform experiments using w^{1118CS} flies as a baseline and then repeat the experiments using transgenic flies with a silenced temperature neuron. However since the results of the experiments showed an insufficient reaction to humidity stimuli, experiments were only performed using w^{1118CS} control flies.

Prior to the experiments, the flies were housed in an incubator with a temperature of $25°C$ that subjected the flies to a 12:12 hour dark-light cycle. The flies were reared on the diet showed in table 8.1, Appendix A.
The intention was to select flies that would show a stronger response to humidity, which is why the following actions were take to enhance the response. Firstly, only male flies with an age ranging from 2-14 days old were used. Secondly, flies were also sorted out and placed in empty vials to dehydrate 14-24 hours prior to an experiment. Furthermore, the experiments were performed during the time when they were usually subjected to daylight, while trying to select time periods when the flies according to studies are expected to be active.[17]
In order to transfer the flies from their vial into the arena they were anesthetized by placing the vial in ice. Before starting an experiment, the flies were allowed to acclimatize in the arena for a few minutes until they were sufficiently active.

The experiments were performed over four days, resulting in a data set with 32 flies at $24°C$ and 29 flies at $26°C$. Thus consisting of 305 trials (five hours of measurements) for both types of trials in figure 5.2. Many days of previous experiments were discarded due to flies showing insufficient activity or technical difficulties. Further measurements of individual flies were discarded due to inactivity or the inability to transform the coordinates of the trajectory data. The image analysis program performing the coordinate transformation was sometimes disabled by missing visual cues.

5.2 Preprocessing

The resulting trajectory data from the experiment consisted of the coordinates of four flies along with the time point of each observation. The trajectory data needed to be processed before an analysis was possible as it contained inactive flies, missing values and noise. The coordinate system of the data also differed depending on the measured fly. Therefore a set of filters and transformations described in this section were implemented to process the trajectory data.

5.2.1 Interpolating missing values

Test runs were performed prior to the experiments to assure that the program Margo could sufficiently track flies in the experiment setup. Studying the retrieved data showed that Margo occasionally lost an object during tracking, resulting in missing values for data outputs at the corresponding time interval. Firstly, when starting the tracking program the initial output during the first seconds were always missing.

Thus the first seconds of an experiment had to be ignored and the starting time was delayed until the first measurement.
Secondly, there were occasionally sequential missing values throughout the experiment, lasting from a split second up to several minutes. Studying a video of the flies along with tagged time intervals where the tracker failed, revealed that all cases of detection failure was due to the fly being inactive. After the period of missing values there could be a slight offset in position, corresponding to the time before the tracker recognises the fly after it resumes moving. But since the object mostly remained in place, the missing values were replaced using linear interpolation and a more sophisticated method was deemed unnecessary. For most of the sufficiently active flies the missing values made up a few percent of the total data.

5.2.2 Transforming data coordinates

Since the trajectories of four flies were measured simultaneously, each in a distinct chamber, their coordinates would differ. The coordinate systems from different batches could also vary as the camera or arena would be in a slightly different position. Therefore an image analysis program was deployed to align the data from all flies to the same coordinate system (developed by Kalle Andersson).

5.2.3 Filters

Before the trajectory data could be analyzed, outliers and noise needed to be filtered. In a first step data sequences from inactive flies (flies with a mean speed less than 0.1 mm/s) and data sequences where the coordinate transformation malfunctioned were removed.

Secondly all shot noise in the form of a positional change larger than the threshold 20 mm/s was removed. The threshold was deemed reasonable as a visual investigation of the data revealed that positional changes of a larger magnitude did not correspond to natural movement. The shot noise value was replaced by the value of the previous time point.
Even still, the output trajectory data was visibly noisy. Filtering using a Butterworth filter was suggested by a study analyzing fly trajectories.[8] After experimenting with differently tuned filters a two-pole Butterworth filter was chosen that low-pass filtered data at 0.1 Hz, using Matlab functions "butter" and "filtfilt". Trading in the loss of information for more smoothed data that was deemed more suitable for the applications in this thesis.
Lastly the positional data for the individual one minute trials were extracted. This was possible by synchronizing the real time output of Margo with the pump output. The extraction allowed for analysis of both entire experiment runs and individual trajectories.

5.3 Analysis

Potential reactions to humidity stimuli were investigated by calculating features from the experiment data. The features shown in table 5.1 are partly inspired by a study from Alvarez et al.[8], also using their definition of curvature. Their study showed that fly reactions translated into significant changes in feature values.

5.3.1 Features

Using the processed position data (x, y) several features (table 5.1) were calculated that would show possible reactions to stimuli and illustrate characteristics of fly locomotion. The average of feature values was calculated individually for 305 trials of type A or B, shown in figure 5.2 (with 5 trials for each of the 61 flies in the data set). The feature averages of 61 flies during the entire experiment duration was also calculated. Potential responses were investigated by analyzing how the average of feature values changed with time, with a change during a time period of altered humidity being a possible reaction. The results are displayed in figures 6.4, 6.5.

The significance of a feature change was determined using Wilcoxon signed rank tests (described in section 4.2.1). The tests displayed in table 6.1, were performed using the Matlab function "signrank". With the input being the feature value average of individual trials during periods before, during and after the periods of altered humidity in trials. Thus testing the significance of possible reactions at stimuli onset and offset.
No adjustments were made to compensate for the increase in familywise error rates with multiple comparisons. Compensating using the Bonferroni method would divide the significance level α by the number of comparisons. But this compensation could be overly strict as the features are correlated and individual comparisons are not performed. However one should note that P-values close to α=0.05 would indicate an insignificant result if the significance level were to be adjusted.

Feature	Description
Activity	Flies are considered active when moving faster than 1 mm/s.
y	Distance to the humidity source (mm).
v, v_y	Speed, speed component in the y direction (mm/s).
θ	Angle towards the humidity source (rad).
$\theta^{'}$	Angular velocity (rad/s).
$C = \frac{\lvert\theta'\rvert}{v}$	Curvature (rad/mm). Excluding v<1 mm/s and $\theta^{'}$ <π/18 rad/s.

Table 5.1: Features calculated from the experiment data.

5.3.2 Further filtering of low-speed data

Previous studies investigating fly behaviour have shown that if low-speed data is included in the analysis, the prominence of a mean reaction could be dampened. They also show a clear correlation between flies moving at a slow speed and not reacting to stimuli.[17] Therefore an additional analysis of the experiment data was performed after removing slow-speed data. The data set was further processed by removing the ten flies with the slowest average speed and ignoring all data measures where the speed was slower than 1 mm/s (these limits are illustrated in figure 6.3). Removing individual trials instead of all trials belonging to a fly is also a possibility, however this could produce a more biased result. The average feature values of the data set after removing slow-speed data is illustrated in figures 8.1 and 8.2, Appendix A.

5.4 Modelling

The conclusion from the analysis (displayed in section 6.1) was that flies did not show a significant reaction to humidity stimuli in the experiments. A HMM was still fitted to the data set from this thesis to show that it could find a viable model for fly locomotion adapted to the data set. However, the intention of modelling different responses was not possible.
In order to show that the developed HMM could still be used as a framework to find differences in behaviour, the model was also fitted to an external data set from Alvarez et al.[18] They performed similar experiments using odor stimuli (previously described in section 4.1.3) and their research found significant responses at odor onset and offset. As their data set contains trajectory data from flies walking in an arena with an identical design and being subjected to a gradient of stimuli, it was preferred over other possible external data sets. Fitting models to their data set allows for showing how the model was intended to work on the experiment data from this thesis.

5.4.1 Data input

HMMs are fitted to sequences of two-dimensional data with speed and angular velocity v, θ'. With the angular velocity being transformed to be positive when a fly is turning towards the stimuli source, and negative otherwise. These features provide complete information about the positional change of a fly and also seem to captures the change in behaviour caused by stimuli. Each sequence consisted of 100 observations from experiment trials, as shown in table 4.1. With the external data set providing 1155 trials of data.

As discussed in section 5.3.2 it was deemed beneficial to filter out sequences containing low-speed data. The fit of the HMM is also affected by low-speed data and will give a higher likelihood to sequences with inactive flies. However since the sequential order of the data is important, it was problematic to remove individual data measures. Therefore entire sequences with an average speed less than 1 mm/s were removed.
The K-means clustering and Baum-Welch algorithm are also susceptible to outliers, being prone to overly adapt the fit to comply with outliers. Therefore sequences with an excessive amount of outliers were removed. Sequences containing outliers were found by looking at the histogram of the data set and comparing if they contained unusual values. The histogram of a data set before and after removing outliers is shown in figure 5.3.

Three HMMs were fitted to individual subsets of the external data set from Alvarez et al.[18] These subsets contain data from trials, with two seconds of measures from when flies are either Neutral (not receiving a stimuli), experiencing an ON-response (stimuli onset) and experiencing an OFF-response (stimuli offset). Each subset is processed as described above before fitting a HMM, resulting in three model fits. The different model fits are investigated by comparing model parameters, along with simulated trajectories from each model fit.

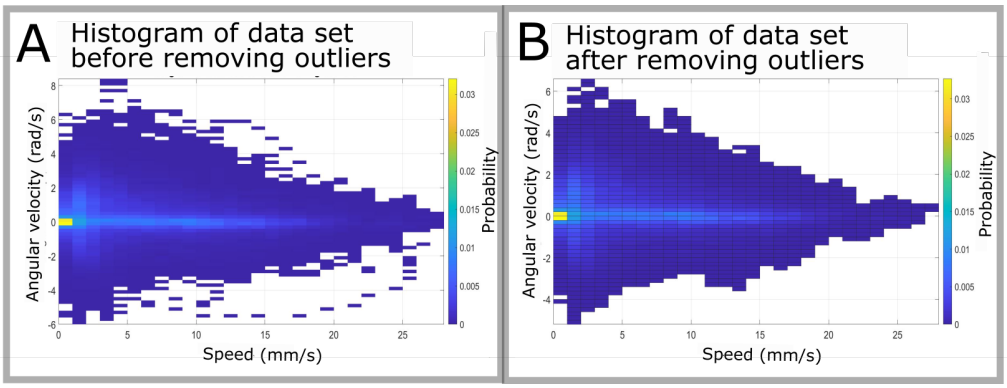


Figure 5.3: Histogram of a data set before and after removing outlier trajectories.

5.4.2 Fitting process

A HMM with six states and four mixture components was fitted to each data subset. The models were initialized as described in section 4.2.5, giving initial parameter values for all Gaussian mixture models. With the Matlab K-means clustering algorithm "kmeans", segmenting the data into states and the Matlab function "fitgmdist" in turn segmenting the states into mixture components. Both algorithms were run 100 times, choosing the best resulting fit in terms of likelihood. The transitional probabilities A and the initial state distribution π were initialized with uniform probabilities.
The final model fits were found using the Baum Welch algorithm (algorithm 1), used for 500 iterations or until convergence. The convergence criterion was chosen as when the relative change in likelihood was less than 0.01 %.

Numerical underflow is a common problem when using the Baum-Welch algorithm and is a consequence of the low likelihoods that can occur. As an example the forward probability α is a product of probabilities and likelihoods that are all less than one, which can result in very small numbers depending on the model limits. A number of adaptations suggested by a study from Mikael Nilsson[14] were made to deal with numerical underflow, the corresponding Matlab code is found in Appendix B (Fitting process). Firstly, equations were transformed to logarithmic versions when necessary. Secondly, whenever a numerical underflow occurred the value was substituted for the lowest available number in Matlab.
Furthermore, parameter values of Gaussian mixture models were modified when needed. The probability of a mixture component was set to never subceed 1e-10. The diagonal elements of a covariance matrix belonging to a Gaussian distribution were also hindered from subceeding 0.25, preventing the distribution from covering a too small area in the data space. The limit of 0.25 is larger than suggested by [14], but was deemed necessary from preventing the likelihood output of a Gaussian distribution becoming larger than one.
The covariance matrix of a Gaussian distribution was also prohibited from becoming negative definite. When this problem occurred the largest number out of the matrix determinant and 1e-4 was added to the matrix diagonal values.

5.4.3 Model selection

There is no general consensus on choosing the number of states and mixture components for a Gaussian mixture HMM. Increasing the amount of parameters improves the likelihood with the caveat of also increasing model complexity. Studies have shown that conducting likelihood ratio tests between different model fits can be problematic as the tests have been shown to be unbounded for a HMM.[19] Several studies deploy a penalized likelihood that subtracts the likelihood estimate with a penalty which increases with the number of model parameters.[19] This thesis initially deployed penalized likelihoods in the form of Aikake information criterion (AIC) and Bayesian information criterion (BIC). Selecting the model that for AIC minimizes

$$-2l(\lambda) + 2K$$

and for BIC minimizes

$$-2l(\lambda) + Kln(T * F)$$

with $l(\lambda)$ being the log-likelihood of the model with parameters λ, K the number of model parameters and $T * F$ the number of observations. [20] However the penalty in these criterions was not sufficient and the more complex model was always preferred. This is a recognized problem and studies develop other penalized likelihood specialized for certain models. At the time of writing no study with a criterion specialized for Gaussian mixture HMMs was found.
In a study by Tao et al. the number of model states and mixture components were decided by studying two factors. When the model output resulted in trajectories that were similar to each other and to which degree states and mixture components were used.[11]

This thesis deploys a similar strategy. Firstly the initial cluster segmentation gives some indication about the number of Markov states. When increasing the amount of clusters/states to more than five, the improvement in explained variation dropped of. Choosing the amount of Markov states is a separated problem, but five states was used as an starting point in the model selection. Different HMMs were then fitted, varying the amount of states and mixture components. In the end a model fit with six states and four mixture components was chosen due to generating similar trajectories while also relatively frequently using all states and mixture components.

6 Results and Analysis

6.1 Analysis of the experiment data

The resulting data set from the thesis experiments was analyzed by comparing the change in the average of feature values, as described in 5.3.1. Figure 6.1 shows individual examples of fly trajectory data that comprises the results. Data from one minute periods corresponding to a trial (either type A or B, as shown in figure 5.2) are analyzed separately and compared to data from the entire 20 minute experiment. Overall there was variation between the trajectories of different trials, with no clear trend of moving in a certain direction.

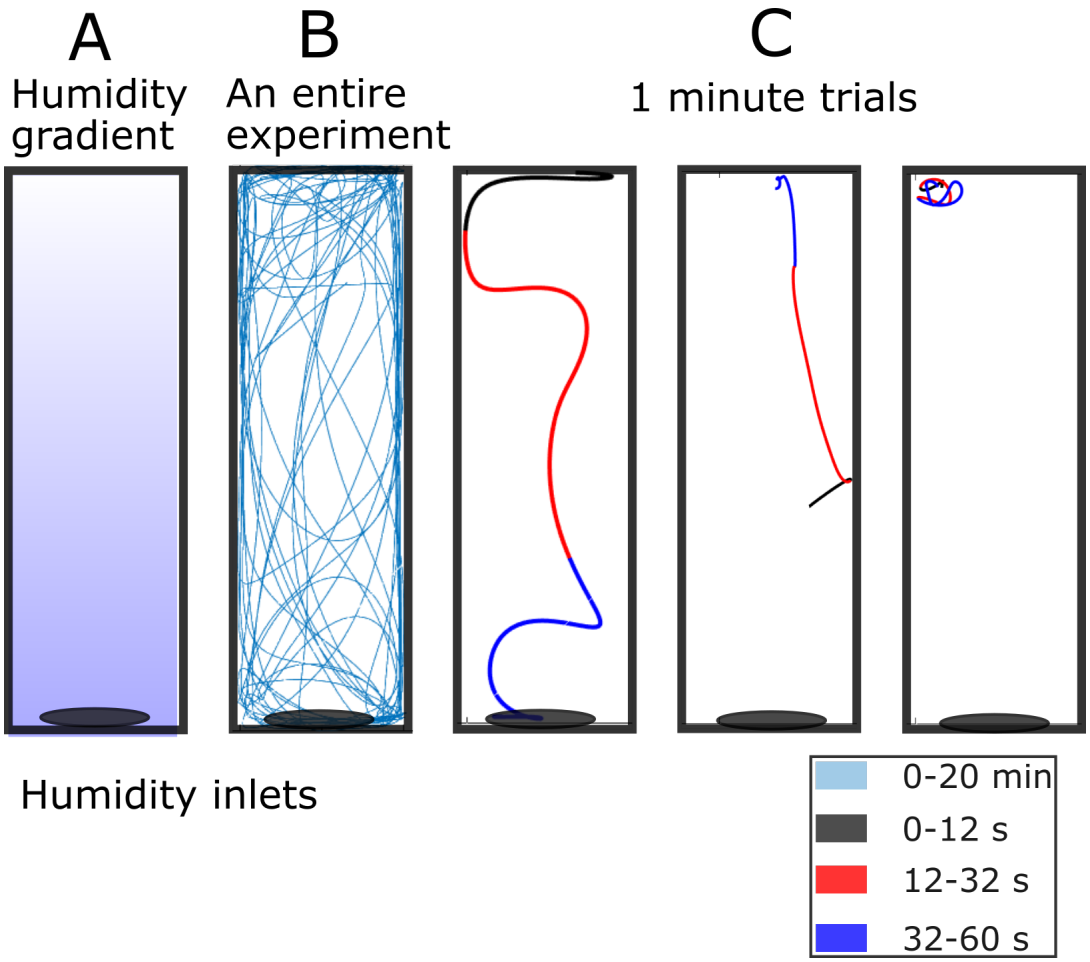


Figure 6.1: (A) The expected humidity gradient inside a chamber. With either higher or lower relative humidity at the inlet, depending on the relative humidity of the air supplied to the arena. (B) Trajectory data from the entire 20 minute experiment of a single fly. (C) Trajectory data from individual one minute trials. Segments of the trials are either colored gray (before a humidity level change), red (during the humidity level change), blue (after the humidity level change).

Figure 6.2 show the overall positional trends of flies, using positional data from the entire data set. While there is movement in the middle of the chamber, flies show a clear preference for the edges of the chamber, which is a common phenomenon in behavioural assays using flies.[21] The bottom part of the chamber is also preferred over the top part. However there is still movement and a trend of the average y-position decreasing by a magnitude of five mm over the course of the experiment.
Furthermore, the occurrences of unnaturally straight trajectories in figure 6.2 A indicates that the object tracker sometimes looses track of a fly.

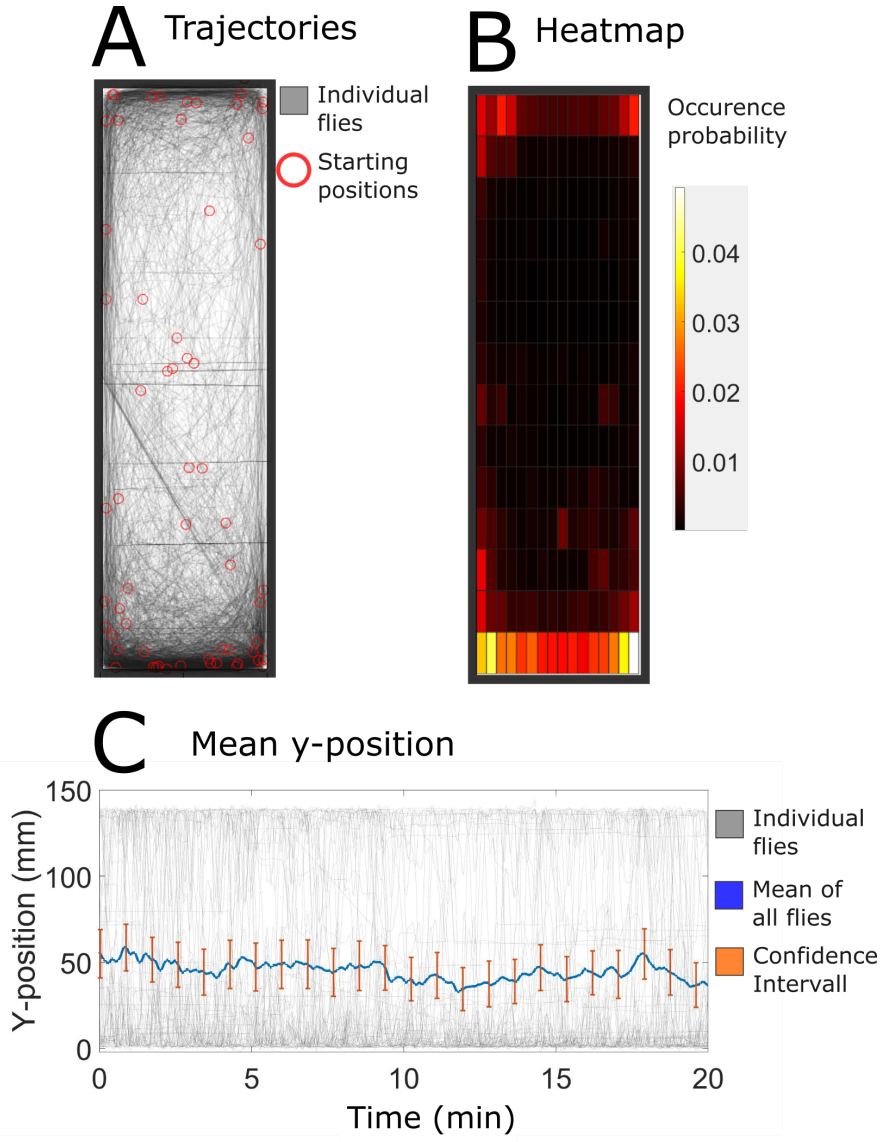


Figure 6.2: (A) the resulting trajectories from the entire data set. (B) A heat-map of the positional data in A, normalized so that the value of each bin is the probability of a fly occurring in that position. (C) Y-position data from the entire data set. The measures for individual flies (grey), the average of all flies (blue) and confidence intervals (orange).

Figure 6.3 shows the overall walking speed of flies, using the entire data set. Figure 6.3 A shows that there is a no clear difference in average speed with regards to the temperature but relatively large differences between individual flies. This indicates what previous studies have shown in that the movement characteristics can differ greatly between individual flies.[11] It can also be a sign that more data would be beneficial.
Figure 6.3 B shows how flies spend a majority of the time standing still or moving at a slow speed. There is also a correlation between the overall mean speed and the y-speed during trials, as illustrated in figure 6.3 C. Indicating that active flies are more likely to show a reaction of moving towards or away from the source.

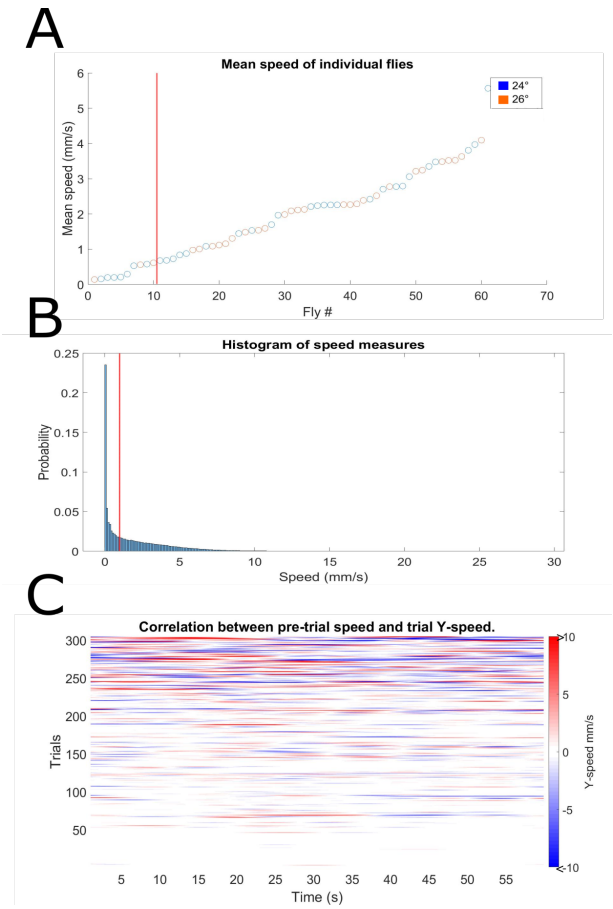


Figure 6.3: (A) The mean speed of the 60 flies belonging to the data set, sorted in ascending order. The background temperature during the corresponding experiment run of either $24°C$ (blue) or $26°C$ (orange) is highlighted. The red line denotes the limit for eliminating the ten slowest flies. (B) The histogram containing all individual speed measures of the entire data set. The red line denotes the limit for ignoring all speed measure less than 1 mm/s. (C) The y-speed (v_y) during all type B trials sorted in ascending order based on the mean speed during the five minute baseline period before the trials (see figure 5.2 for references).

The trajectory data is analyzed before (figure 6.4 and 6.5) and after (figure 8.1 and 8.2, Appendix A) further removing low-speed data. As no major differences were observed, the results from the further filtered data are shown in Appendix A. The limits for removing low-speed data is illustrated by the red lines in figure 6.3 A, B and the process is described in section 5.3.2.

The average of feature values, explained in table 5.1, during either type A or B trials are shown in figure 6.4. The results indicate a trend in activity, y-position, speed and y-speed caused by humidity stimuli. The trend also seems to be antagonistic depending on the stimulus type. From table 6.1 one can also see that there are statistically significant changes in y-position, speed, angular velocity and curvature. However the confidence intervals are relatively large compared to the trends.

Figure 6.5 B shows the average feature values during the entire experiment. When comparing peaks and dips during trials to fluctuations during periods with constant humidity, one can see that they are of the same magnitude. Excluding a few peaks during the trails, most notably in activity and velocity which are slightly larger than the ones during the baseline humidity. Figure 6.5 A also shows that averaging data from randomly selected one minute periods produces trends of the same magnitude as in figure 6.4. Indicating that possible responses during trials are not more synchronized with the stimuli than random fluctuations.

Furthermore there are indications in figure 6.5 B of long term reactions in activity, y-position and speed over the course of the experiment. Notably, the moving average of the angular velocity during the last quarter is also elevated compared to the rest of the experiment. Table 6.1 shows that there are also statistically significant differences between the average speed and angular velocity during the first and second half of the experiment.

Conclusively, there are some indications that flies react to the humidity stimuli in terms of altered activity, speed and angular velocity. However they are not clear enough to draw any conclusions. Based on these results, analyzing the differences in reaction related to temperature and conducting further experiments with genetically modified flies was not deemed relevant.

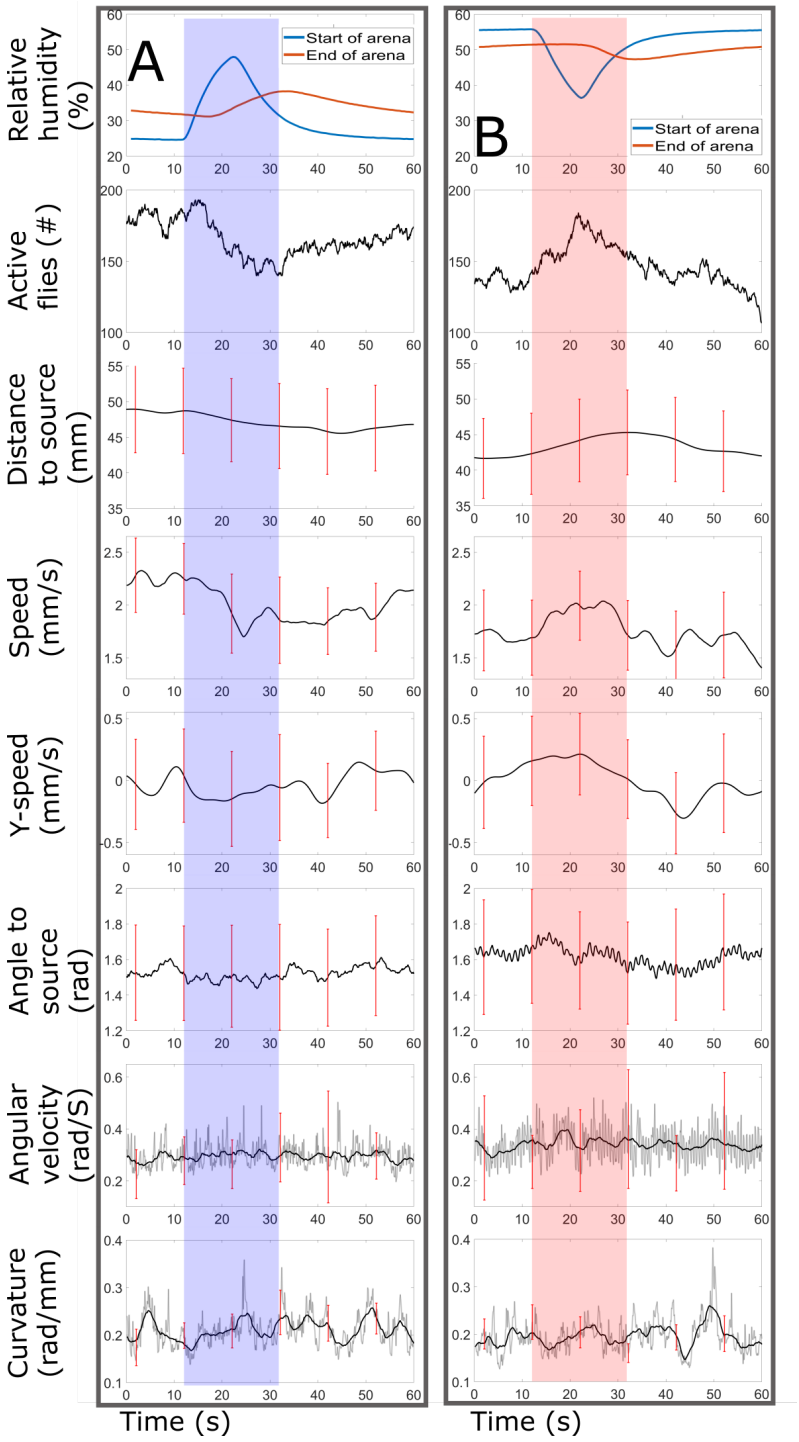


Figure 6.4: The average of feature values, described in table 5.1, during type A and type B trials. A twenty second period of altered humidity is highlighted with blue and red. Confidence intervals are showed as red bars. For angular velocity and curvature the average feature value (gray) is plotted along with a moving average (black) with a window length of three seconds; in order to more prominently show possible trends. The average features are calculated from data containing 5 trials each from 61 flies.

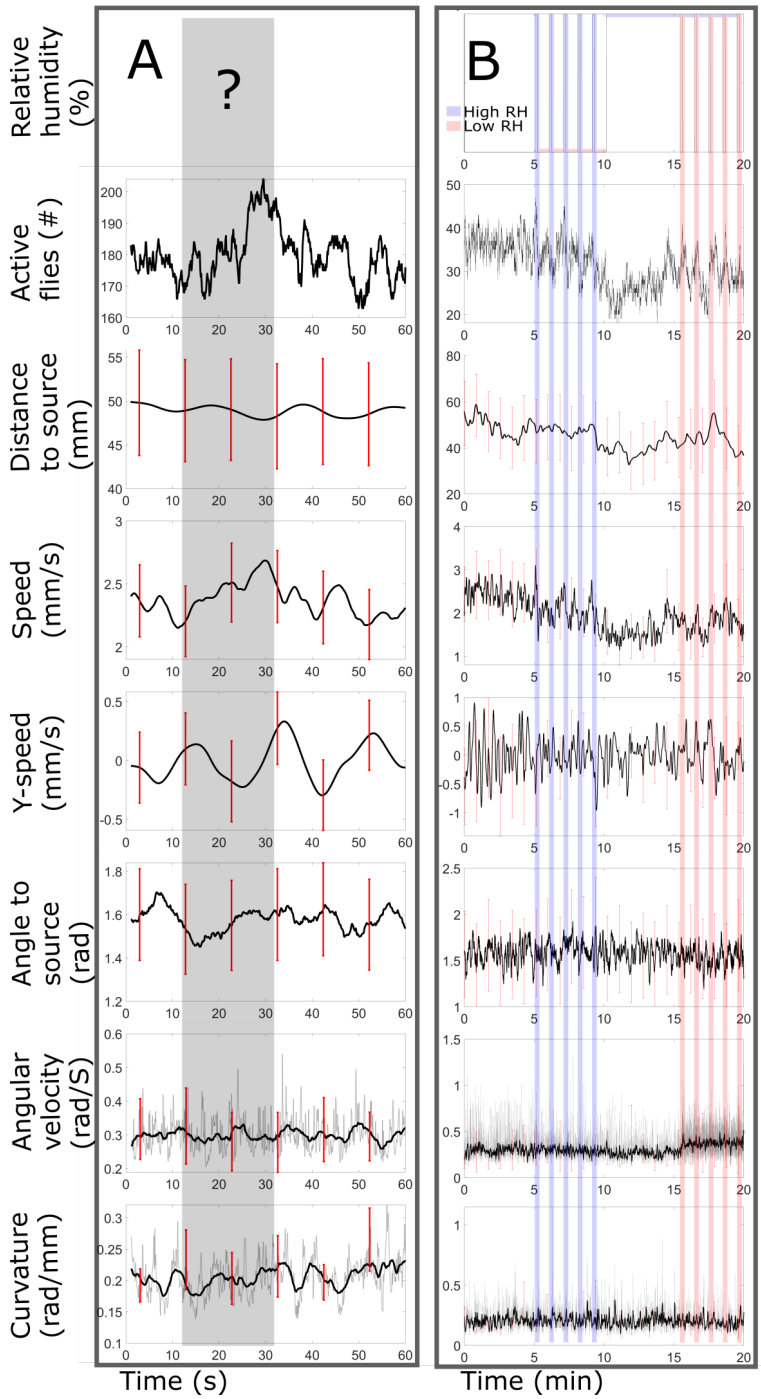


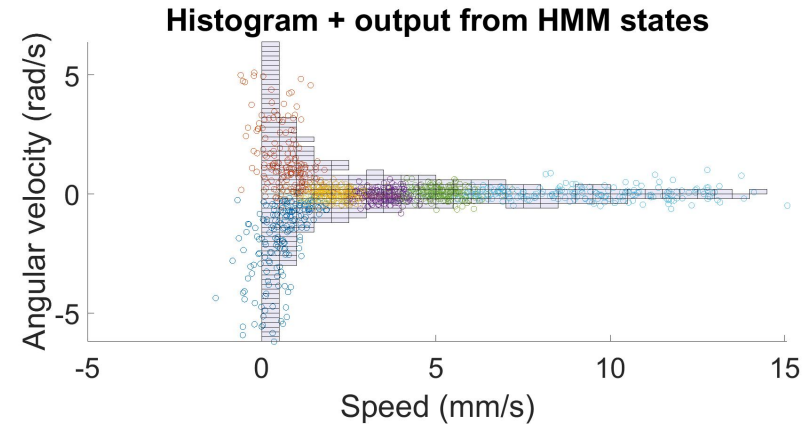
Figure 6.5: The average of feature values, described in table 5.1. Twenty second periods of altered humidity is highlighted with blue and red and a twenty second period of unknown humidity is highlighted with gray. Confidence intervals are showed as red bars. For angular velocity and curvature the average feature value (gray) is plotted along with a moving average (black) with a window length of three seconds; in order to more prominently show possible trends. (A) The average is calculated from 5 randomly selected one minute periods from each of the 61 flies. (B) The average is calculated from the 61 flies over the course of the entire experiment.

P-values Wilcoxon signed rank test	**A**			**B**			**Entire**
Time periods	0-12 s vs 12-32 s	12-32 s vs 32-52 s	0-12 s vs 32-52 s	0-12 s vs 12- 32 s	12-32 s vs 32-52 s	0-12 s vs 32-52 s	0-10 min vs 10-20 min
y	0.46	0.057	0.048*	0.026*	0.83	0.33	0.088
v	0.034*	0.27	0.0050*	2.4e-04*	2.8-04*	0.75	2.78e-06*
v_y	0.76	0.41	0.56	0.15	0.56	0.34	0.12
θ	0.33	0.36	0.74	0.77	0.62	0.89	0.062
θ'	0.12	0.85	0.052	0.016*	0.032*	0.50	9.9e-04*
C	0.68	0.053	0.52	0.28	0.075	0.012*	0.25

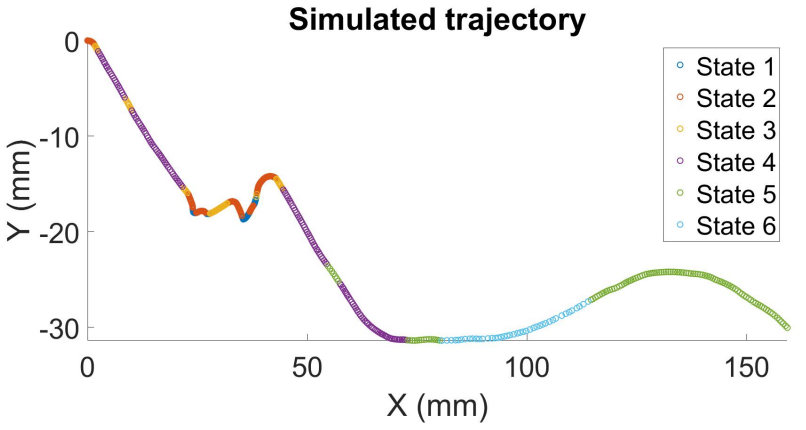
Table 6.1: The resulting p-values of comparisons between the feature averages, as described in section 5.3.1. with * denoting a significant change on a 95% confidence level. Tests are performed individually for type A trials, type B trials and the entire experiment, comparing averages from the specified time periods. The time periods in A, B correspond to before, during and after the humidity change. While the two halves of the experiment are compared in Entire.

6.2 Modelling fly movement

A HMM with six states and four mixture components was fitted to the data set from the experiments, using observations of speed and angular velocity from 11 to 21 seconds during type B trials (see figure 5.2). In this fit 81 % of the observations were assigned to belong to one of the states with a probability of 95 % or higher. The histogram of the data along with the resulting fit can be seen in figure 6.6 (a). The outputs of the fitted HMM states overlap well with the two-dimensional data set. When simulating from the HMM, it switches between the states given the fitted transition probabilities and generates observations from the probability density function of the current state. Figure 6.6 (b) shows how a simulated fly switches between states during a generated trajectory. The behaviour during a state corresponds to its Gaussian mixture model. During state one and two the simulated fly performs more sharp turns. During state six it walks straighter, with a higher speed.



(a)



(b)

Figure 6.6: (a) A fitted HMM with six states and four mixture components. Along with a histogram of the data from type B trials to which the model was fitted. The coloured dots denote 400 outputs from each of the six Gaussian mixture models. With the colour of each state shown in b. (b) A simulated fly trajectory from the fitted model in a. The fly starts in origo, with the simulation lasting one minute. The segments of the trajectory are color coded after which state the fly occupies.

However, since the results in section 6.1 showed that no inference regarding different fly behaviour caused by stimuli could be drawn, HMMs were fitted to an external data set as described in section 5.4.1. Figure 6.7 A, C, E shows histograms of subsets from the external data set to which three HMMs were fitted. Each subset uses two seconds of data per sequence, resulting in 100 data measures of speed and angular velocity. The neutral subset contains data from when flies are not subjected to stimuli, the ON-response subset contains data from stimuli onset and the OFF-response subset contains data from the stimuli offset.

Although the relative shapes of the histograms are similar, there are notable differences. The histogram of the ON-response shows a larger probability for high-speed data, while also being skewed towards positive angular velocity, corresponding to turning towards the humidity source. Contrariwise, the histogram of the OFF-response is more likely to generate low-speed data with a higher angular velocity. The histogram of the neutral fit provides somewhat of a middle ground.

Figure 6.7 B, D, F show histograms of the output of the three fitted HMMs and serves as a control that the model generates data corresponding to its data subset. Overall, the histograms of the model outputs look relatively similar to the histograms of the data subsets.

Figure 6.8 illustrates the three model fits by showing the outputs from the Gaussian mixture models of their six fitted states. One can not draw conclusions of the overall movement by individually comparing the similar states from the models, since they are not the same. I.e. although state one in all models look similar, there is not one set in stone characteristic that they all describe. The model simply fitted a state to accommodate the cluster of low speed data. Thus one needs to study all states to draw any conclusions.
Overall the three fits in figure 6.8 look relatively similar, as they all describe fly movement. There are however differences between the Gaussian mixture models, transitional probabilities and state probabilities.

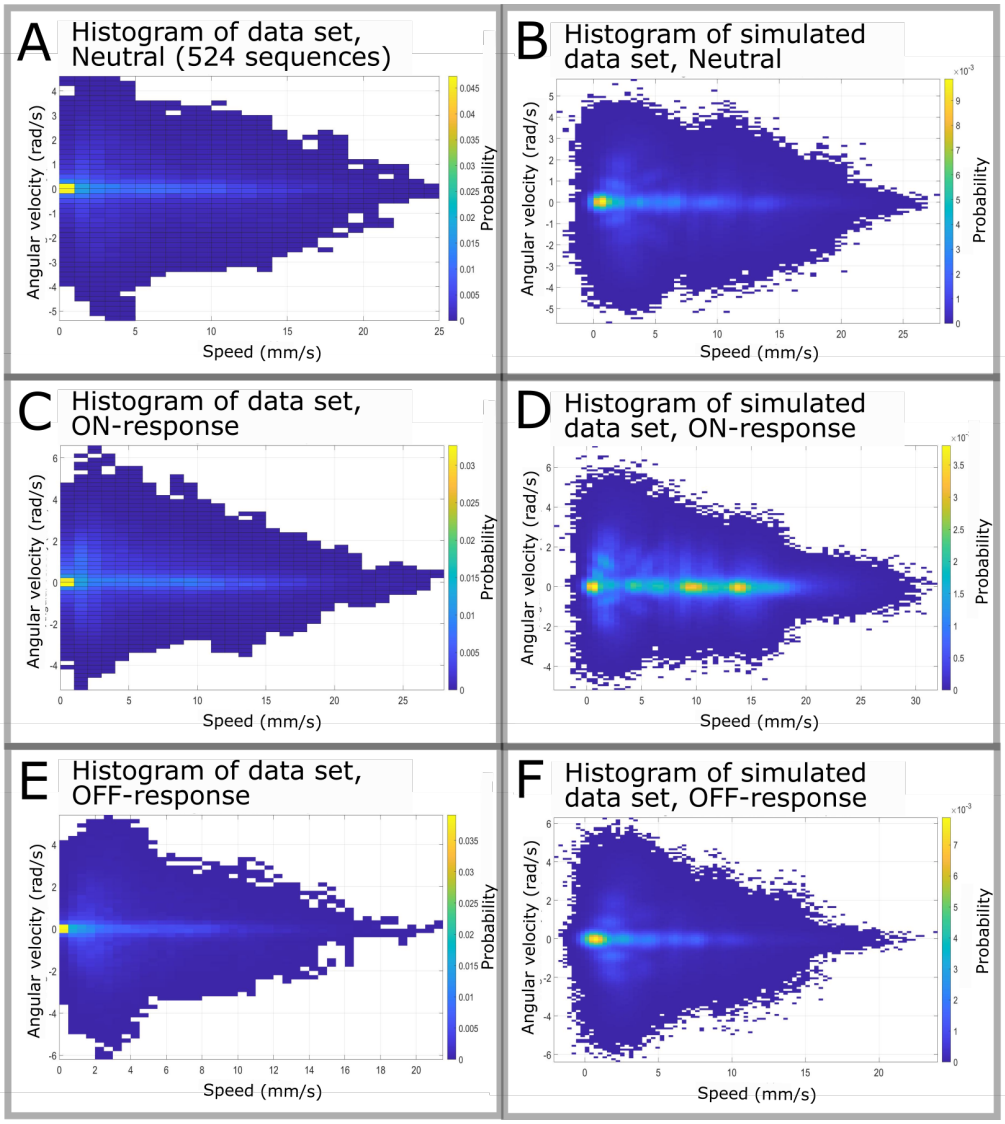
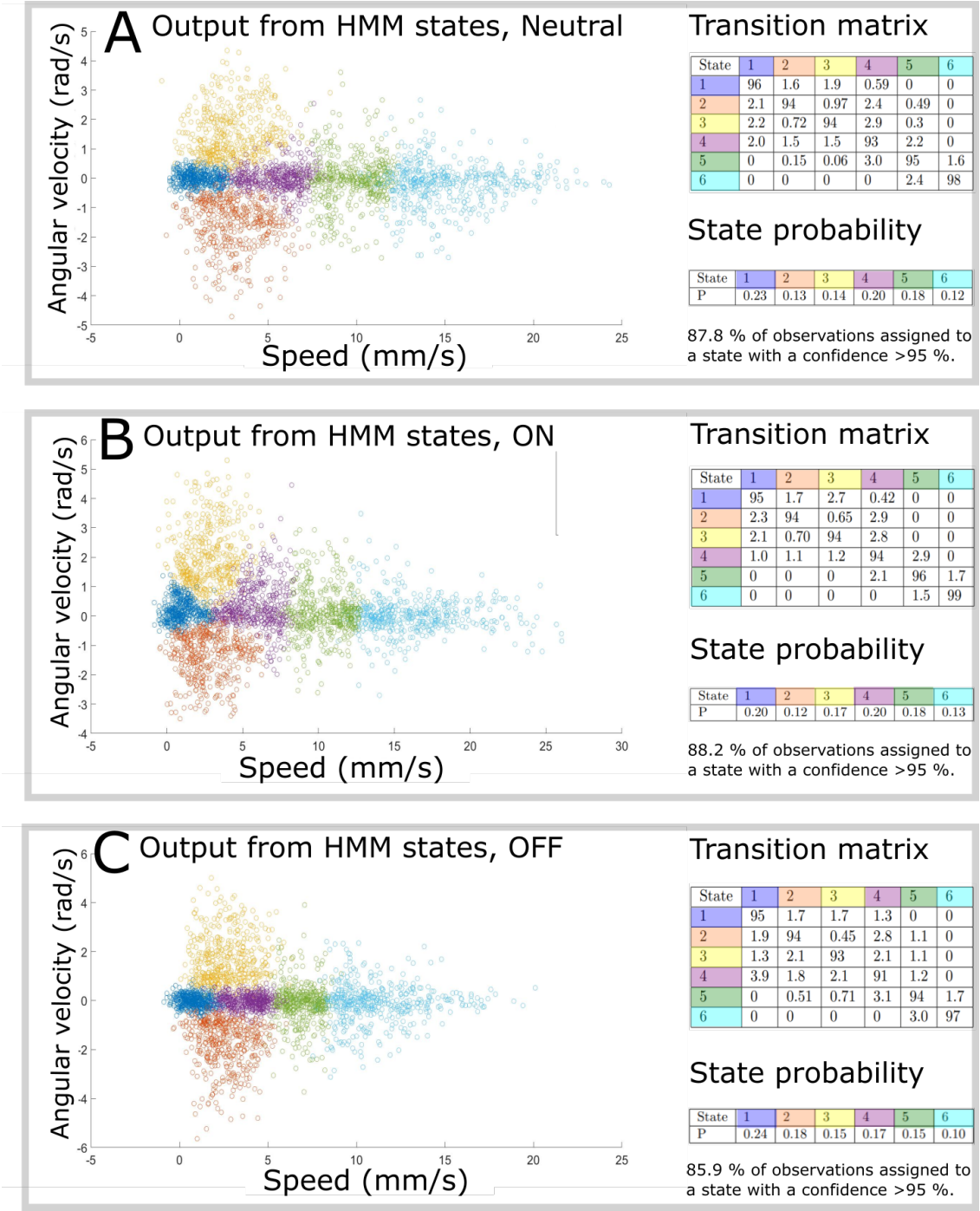


Figure 6.7: (A), (C), (E) Histograms of the data subsets from the external data set of Alvarez et al.[18] further described in section 5.4.1. Each bin covers a range of speed and angular velocity. Where positive angular velocity corresponds to turning towards the stimuli source. The color of each bin denotes the probability of a measure occurring within the bin range. (B), (D), (F) Histograms of the generated output from the three HMMs fitted to the data subsets in A,C,E.



State	1	2	3	4	5	6
1	96	1.6	1.9	0.59	0	0
2	2.1	94	0.97	2.4	0.49	0
3	2.2	0.72	94	2.9	0.3	0
4	2.0	1.5	1.5	93	2.2	0
5	0	0.15	0.06	3.0	95	1.6
6	0	0	0	0	2.4	98

State	1	2	3	4	5	6
P	0.23	0.13	0.14	0.20	0.18	0.12

State	1	2	3	4	5	6
1	95	1.7	2.7	0.42	0	0
2	2.3	94	0.65	2.9	0	0
3	2.1	0.70	94	2.8	0	0
4	1.0	1.1	1.2	94	2.9	0
5	0	0	0	2.1	96	1.7
6	0	0	0	0	1.5	99

State	1	2	3	4	5	6
P	0.20	0.12	0.17	0.20	0.18	0.13

State	1	2	3	4	5	6
1	95	1.7	1.7	1.3	0	0
2	1.9	94	0.45	2.8	1.1	0
3	1.3	2.1	93	2.1	1.1	0
4	3.9	1.8	2.1	91	1.2	0
5	0	0.51	0.71	3.1	94	1.7
6	0	0	0	0	3.0	97

State	1	2	3	4	5	6
P	0.24	0.18	0.15	0.17	0.15	0.10

Figure 6.8: The resulting fits of HMMs with six states and four mixture components. Where the models (A), (B), (C) are fitted to the data subsets in figure 6.7 A, C, E. Thus the fitted models describe fly locomotion during no stimuli (A), stimuli onset (B) and stimuli offset (C). For every subfigure, the coloured dots denote 400 outputs from each of the six Gaussian mixture models that belong to the states. The colour of each state along with the transition probabilities are shown in the transition matrix and the overall probability of a fly occurring in a state is shown in the state probability vector.

Figure 6.9 visualizes the differences between the three fits in figure 6.8 by simulating trajectories from each model fit. The trajectories from the ON-response fit clearly turn towards the stimuli source and show greater speed overall. The trajectories from the OFF-response fit show slower speeds with sharper turns. There is also a slight tendency to turn away from the stimuli source for the neutral fit and especially for the OFF-response fit. Possibly due to flies being directed towards the source before stimuli offset, in which case every turn is away from the stimuli source.

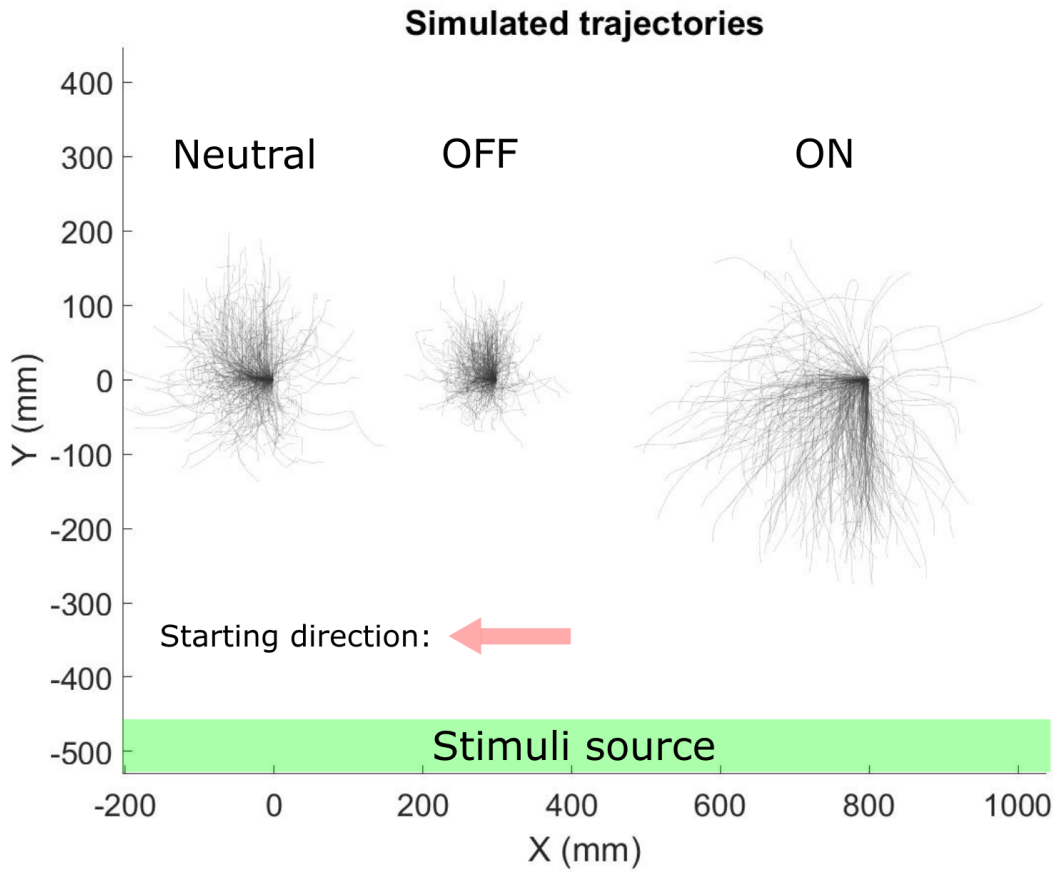


Figure 6.9: Trajectories simulated using the three model fits in figure 6.8. The artificial flies are not restricted by an arena. Each gray curve is a simulated trajectory, lasting 25 seconds with a fly starting at $y = 0$, directed to the left, with the stimuli source below the fly. 100 trajectories are generated from each model where flies are either not experiencing a stimuli (Neutral), experiencing an OFF-response (OFF) or experiencing an ON-response (ON). The trajectories from the different models are shown separately by starting at different x-values ($x = 0$, $x = 300$, $x = 800$).

Figure 6.10 shows trajectories generated from alternating between the three model fits to simulate an experiment trial with shifting stimuli. The simulated fly experiences no stimulus for ten seconds and then experiences ten seconds of ON-response followed by ten seconds of OFF-response. Generally, the fly turns towards the stimuli source with an increasing speed during the ON-response and turns more randomly, (often away from the source) with a decreasing speed during the OFF-response. However the trajectory segments of the three responses can also look relatively similar.

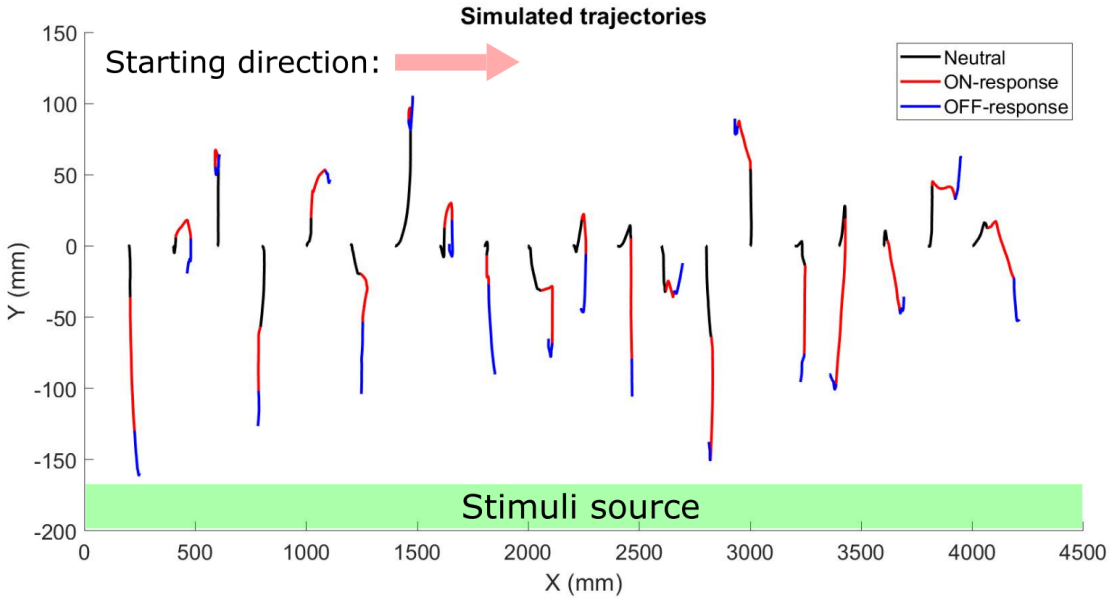


Figure 6.10: Trajectories simulated using the three model fits in figure 6.8. Each trajectory is simulated with a fly starting at $y = 0$, directed to the right, with the stimuli source below the fly. 20 trajectories are generated, each lasting 30 seconds and sampling ten seconds each from the three model fits. The simulated fly experiences no stimuli for ten seconds (grey) and then experiences ten seconds of ON-response (red) followed by ten seconds of OFF-response (blue). In order to clearly show all trajectories separately, they start at different x-values.

7 Discussion

In this thesis hygrosensation was investigated firstly by conducting behavioural experiments. Secondly by developing a framework of statistical analysis and modelling which is applied to study underlying reactions in trajectory data. This section discusses the results and potential improvements. Starting with the behavioural experiments, discussing the potential reasons for why the flies showed an insufficient response and how the experiment setup and procedure could be improved.

The methods for analysis and modelling are also discussed, in terms of their general applicability, the indications of responses that they found and potential shortcomings.

7.1 Reactions to humidity stimuli were insufficient

The results and analysis from section 6.1 concluded that although there were some indications of potential reactions, they were not significant enough to draw any conclusions. Especially with the uncertainty of large confidence intervals and naturally occurring fluctuations of feature values, as shown in figures 6.4, 6.5. Thus analyzing the differences in reaction related to temperature and conducting further experiments with genetically modified flies (with a silenced temperature neuron) was not deemed relevant. There should be a prominent response overall in order to investigate the dependence of temperature and temperature neurons. Therefore, no conclusions can be drawn regarding the hypothesized model for hygrosensation or the role of the temperature neuron. To investigate the role of the temperature neuron one would need to compare responses at two temperatures, using both flies with a silenced temperature neuron and a control line of normal flies. However, since flies are not showing significant responses to the humidity stimuli this would not be possible.

Although it was not possible to investigate this hypothesis, this thesis serves as a framework for investigating and analyzing hygrosensation. It also presents methods using statistical analysis and modelling that can quantify and simulate responses from behavioural experiments. If improving the experimental setup and procedure results in a significant response to stimuli, these methods can be used for future work.
The overall insignificant reactions during the experiments can depend on a multitude of reasons. Overall, behavioural experiments are complex in the sense that the decision of the subject is often decided by more factors than what is investigated. The experimental setup can only make sure the subject receives the desired stimulus and block out other forms of stimuli if possible. Even if the stimuli delivery is successful, the response of the receptor neurons must still be translated into action. In terms of the experiment in this thesis, the lack of a reaction could be explained by both the complexity of fly behaviour and the experiment setup.
Inactive flies were a problem throughout performing the experiments and lead to the removal of a large amount of trials. As previously discussed, activity is a necessity for reaction. Several actions were deployed to try and enhance the activity of the flies (described in section 5.1.3). Dehydrating the flies should make them more prone to seek out humid environments.

Starving the flies and selecting time periods when they are usually subjected to daylight could also increase their response.[17] However, the overall activity was still low and developing an experiment setup and experiment procedure which ensures fly activity could be an improvement. Especially when considering that the overall activity of flies was higher in the external data set of Alvarez et al. where fly responses were prominent. [8]

Another shortcoming could be that the change in relative humidity during trials was not sufficient, or that the stimuli duration was too short. Previous work showed the possibility of flies reacting to humidity in the matter of ten seconds. However, in those studies flies could decide between areas with a larger humidity difference, whereas the stimuli in this thesis is a weaker humidity gradient. Preferably the humidity levels during trials, should change between 20 % and 70 % as suggested by previous work.[2],[7] But with the limitations of the current experiment setup this was not possible. Comparatively, figure 5.2 shows a change in relative humidity of around 20 percentage points at the start of the arena and 5-10 percentage points at the end of the arena (depending on the trial). The change in humidity could be insufficient to provoke a reaction, especially for flies located closer to the end of the arena. It is also reasonable that humidity could be a less powerful attractant compared to odors, which indicate the presence of food. Therefore the stimuli duration might need to be increased for a more prominent response.

The relatively large confidence intervals for many of the calculated features such as in figures 6.4, 6.5 indicate that the amount of data could be insufficient. The limited amount of data was not a cause of planning too few experiment runs. Rather a consequence of having to sort out a large amount of experiments together with a restricted time schedule. If one wishes to repeat similar experiments, increasing the number of trials would be suggested. Especially for quantifying the difference between responses at two temperatures.

7.2 Quantifying responses using the analysis

The analysis from section 6.1 should be able to find and quantify potential reactions to stimuli, but shortcomings of the analysis can also impact the result. In terms of the magnitude of changes in feature values during trials, a more prominent difference was expected than what is shown in figures 6.4, 6.5. Especially when comparing to the responses in the external data set from Alvarez et al.[8] where changes in feature values were notably larger than random fluctuations.

As an example, if the flies were to respond to humidity stimuli by moving down the gradient. Then a majority of flies moving in the same direction should result in changes in the distance to the humidity source and y-speed that are larger than natural fluctuations.

Additionally, the feature changes in figure 6.5 B does not seem to be more synchronized with the humidity stimuli than random fluctuations, as indicated by figure 6.5 A. Because of the relatively large random fluctuations, averaging feature values from randomly selected one minute trials can produce results that look like potential trends.
The data set was also analyzed after removing slow-speed data (as described in section 5.3.2), resulting in figures 8.1, 8.2, Appendix A. A method of dividing the data set based on temperature was also tested to investigate if there were prominent responses for one of the temperatures. But no major increase in responses could be observed from either of these methods and the uncertainty increased as the average values were calculated from less flies.
Furthermore, figure 6.2 showed that flies had a preference for the bottom and top part of the arena. Flies preferring a section of the arena could be a reaction to humidity stimuli or the airflow. It is problematic for the analysis that flies are often close to an edge as it limits the possible directions in which they can move. In combination with the low activity indicated by figure 6.3, this makes it harder to quantify a response using feature values. Additionally, the flies do not traverse toward the humidity source over time to a significant degree, which previous studies would suggest.[2]

There were however some indications of changes in activity, y-position, speed, angular velocity and curvature caused by the humidity stimuli (figures 6.4, 6.5, table 6.1). Since activity is defined as moving faster than a threshold of 1 mm/s, activity and speed are correlated. The results potentially indicate that these are the most important of the list of features for finding a response. Especially speed, for which the P-values were significantly lower than the other features. One should note that the selected time intervals in table 6.1 also impact the result.
There were also signs of long term responses, with a significant difference in speed and angular velocity between the first and second halves of the experiment. Since the humidity baseline changes from low to high after the first half, the flies can show a long term reaction to humidity over the course of the experiment. It remains difficult to deduce whether this is due to the change in humidity or time. The flies could become more acclimatized to the arena and potentially learn about their environment. To properly test the cause of the reaction one could perform control experiments, reversing the stimuli and starting with a high humidity baseline.

Furthermore, drawing conclusions from directional features was also problematic as the feature values fluctuate with a high frequency (shown in figures 6.4, 6.5). Therefore the moving average of the feature values was also studied in order to find potential trends. The moving average results in a further loss of information which is not ideal, but was still deemed to be able to detect feature differences over longer periods of time.
A potential shortcoming of the analysis could be that not enough attention was given to directional features. Measuring the heading direction directly and possible even the positions of the individual antennas could be an improvement. A distribution of the direction of flies over the course of trials could also provide insight into their behaviour.
Lastly, it would be easier to analyze the potential responses if randomly initiated blank trials were incorporated in the experiments. Currently the response during trials is compared to the response during a different time period. If flies alter their behaviour with time, this will become a source of error. Randomly altering between blank trials would eliminate this problem and allow for a more unbiased comparison.

7.3 The results and general applicability of the HMM

The fitted models were able to segment the trajectory data into states that corresponded to different parts of fly locomotion. Overall the results in section 6.2 indicate that the HMMs provided a good fit. More than 80 % of observations were assigned to a state with a confidence higher than 95 %. The model outputs also overlap well with the underlying data subset as shown in figures 6.6 A, 6.7. Studying the model fits in figure 6.8 shows that the distributions of the states are logically restricted in the sense that they don't contain dissimilar data points. The transition probabilities are also logical in the sense that flies have a probability larger than 90 % of remaining in the same state and the most common transitions are to neighbouring states. Remaining in the same state implies regular movement, as frequently transitioning between states could results in an unnatural movement and also makes it harder to draw conclusions about the underlying behaviour. Both these effects could also be an implication of a high sampling frequency when measuring the trajectories.

The results also showed differences between the three model fits in figure 6.8. Overall the ON-response fit corresponded to the behaviour of flies turning towards the stimuli source and an increased walking speed. Whereas the OFF-response fit corresponded to the behaviour of flies performing sharper turns and a decreased walking speed. From studying the outputs in figure 6.8, the ON-response fit has more states with a higher mean speed.

While the OFF-response fit show states with lower speed. State two and three of the OFF-response fit also have a higher mean angular velocity.
The state probabilities and transition probabilities in figure 6.8 also provide information about the model fits. For the OFF-response fit a larger percentage of data was estimated to belong to low speed states and it is overall more likely to transition to lower speed states. The state probabilities of the ON-response fit and neutral fit look relatively similar, with the ON-response fit being slightly more prone to states with higher speeds. The ON-response fit is also more likely to transition to states with higher speeds.

The differences between the model fits are also prominent when studying the generated trajectories, shown in figures 6.9, 6.10. Where flies behave differently depending on if they experience an ON-response or OFF-response. However individual trajectories generated from the different model fits can also look relatively similar. Which is likely a consequence of that all models have states of both low and high speed. It is not improbable that the OFF-response fit generates high-speed data, just more unlikely. Differences in the distribution of speed and angular velocity can also be seen by studying the histograms of the observations in figure 6.7, but the HMMs are able to show the magnitude in which the statistical differences impact fly locomotion.

With that said, there is still a clear similarity between the model fits in figures 6.6, 6.8, especially between the three fits in 6.8, using the same data set. It is unclear to what degree this indicates that one model can be provided for fly locomotion, possibly even insect locomotion general. Where slight changes in the state Gaussian mixture models and in transition probabilities provide the differences in behaviour. Or if the similarity is more of a consequence of the shared experimental procedure and data processing. The results in this study suggests that tweaking one general model by fitting it to different trajectory data sets could be enough to differentiate behaviour. However one should still adapt each model and the amount of model components in regards to the application. Studies have also found better general models using a hierarchical structure.[11]

It is also important to note that the trajectories generated in figures 6.9, 6.10 do not look like real insect trajectories, but rather as smoothed trajectories with higher activity, which is a consequence of the data processing. The model does not incorporate flies stopping either, which is done in other studies.[3] Furthermore the analysis of this study along with others found that there is a great variation between the movement of individual flies. There is no single model for all flies, but a single model with data from many flies is useful for showing overall trends and differences in behaviour.

Although there are differences between the model fits in figure 6.8, the trajectories of the ON-response and OFF-response are slightly different than expected. Based on the study of the external data set [8], a more sporadic search pattern with random turns was expected for the OFF-response model. The simulated flies were also expected to react quicker during the ON-response. Some simulated flies turn towards the source in the matter of seconds, while some require the 25 seconds, possibly due to starting in a high-speed state. The fact that the model incorporates data from a multitude of differently behaving flies can also result in a less prominent response, compared to a model using data from a single reactive fly.

7.4 Potential improvements to the HMM

There are a number of problems that possibly impact the resulting trajectories, one of them being how data subsets are chosen. Firstly, the model processes data sequences of 100 measures. Which corresponds to ten seconds of measurements from the experiments in this thesis, but only to two seconds using the external data set. The responses generally last longer. Increasing the sequence length can cause numerical underflow when running the program but implementing a form of down sampling, or including sequences from a few seconds apart could be a possibility. Perhaps the amount of measurements could also be different depending on the response. With shorter periods for briefer responses and longer periods for the neutral movement.

Secondly, the behaviour also alters during the response periods. Especially during the ON-response, in which the experiments of Alvarez et al.[8] showed that the fly starts by turning and then increases its speed. The ON-response could therefore be segmented into two models.
Lastly the data subsets of the three periods are more similar if individual measures of low speed data are not removed. Filtering out low speed data while not disrupting the sequential order could possibly enhance the simulated response and show a greater difference in speed and angular velocity between models.

The fitting procedure also impacts the resulting models. As previously mentioned, the resulting fits look reasonable, and during the fitting process the likelihood of the observations increased with every iteration. The initialization process also seemed to cause the models to converge to a higher likelihood. However, the resulting fits can still be merely local minimums, as there is no real guarantee with a non-convex optimization.

Another potential problem is the fact that although the total likelihood of the data set increased with every iteration, the likelihoods of the individual sequences did not necessarily increase. Where the product of the sequence likelihoods gives the total likelihood, described in equation 6. The individual sequences also have likelihoods of different magnitudes, which could differ by a factor of 10^{100}. This is caused by the recursive calculation of the likelihoods, shown in section 4.2.4. Depending on if observations are close or far from the center of a Gaussian distribution, the resulting likelihood can differ greatly. Perhaps the fitting procedure should take the individual likelihoods of sequences into account. But since several studies suggested using the total likelihood and no mention of the problem was found from researching, this standard method was chosen. Perhaps the likelihood discrepancies is caused by differences between individual flies.

Choosing the amount of components for the model was also problematic. Section 5.4.3 explains how the amount of states and mixture components were selected. Overall, choosing a HMM in this thesis is more a case of finding a model that provides a good approximation rather than finding some secret underlying model that perfectly describes the behaviour. The number of states and mixture components are not set in stone. However, an insufficient amount of components may fail to describe the complexity of the behaviour, while too many components can result in over-fitting. Comparing models using statistical tests was tried, but without success. A model selection method that is viable for this specific application would be beneficial.
Since fitting the HMM is an unsupervised problem, it is not possible to control if the degree of over-fitting on a test data set. However splitting up the current data set and looking for differences in the fitted models could be a reasonable action.

The similarity between the histogram of a data subset and the histogram of the generated trajectories (shown in figure 6.7) gives some indication that the models fit their data subset. There are slight differences,such as how the histograms of the outputs cover a larger area. There is also a higher occurrence probability than expected around the center of states for the ON-response. But overall the histograms of the outputs and data sets are similar. This is no official test, but still indicates that the models are not heavily under-fitted. The fact that all states and mixture components are used relatively frequently also gives some indication that the models are not heavily over-fitted.

7.5 Conclusions

In conclusion the framework of statistical analysis and modelling from this thesis should be able to detect and characterize reactions when deployed on further experiments. The flies reactions to humidity stimuli in the experiments from this thesis were insufficient. This is primarily thought to be a consequence of a limited data set and an experimental setup and procedure which could be improved significantly. Given the right data set, fly hygrosensation and the role of the temperature neuron could be investigated using this framework. Although the analysis and modelling of an assay should be adapted to each application, this framework could also serve as a starting point to investigate trajectory data from other assays.
The HMM could be a useful tool for more deeply investigating and visualising reactions to stimuli. Admittedly, some conclusions can also be drawn from simply studying speed statistics. But the models are able to show the magnitude in which the statistical differences impact fly locomotion. Previous work had already showed how HMMs can be deployed to simulate fly locomotion and differentiate flies based on if they were subjected to stimuli.[11] But the HMMs in this thesis could simulate different responses caused by stimuli onset and offset. Improvements on model selection and processing of the input data would be beneficial.

7.6 Future work

This thesis was performed in collaboration with Kalle Andersson, who will continue to investigate hygrosensation using a similar experimental setup and framework for analysis. He will present a thesis with an improved experimental setup that hopefully elicits a reaction from flies. The major changes include increasing the duration of the stimuli along with greater differences in relative humidity.
In the long term, the research group wishes to perform similar experiments with an improved redesigned setup that allows greater control over temperature and humidity. Potentially this setup will also incorporate the ability to directly silence the temperature neuron of mutated flies using light of a specific wave length. Successfully finding a relation between temperature, humidity and the responding reaction of flies would make us one step closer to understanding the mechanisms behind hygrosensation.

8 References

[1] Harald Tichy, Maria Hellwig, and Wolfgang Kallina. "Revisiting Theories of Humidity Transduction: A Focus on Electrophysiological Data". In: *Frontiers in Physiology* 8 (2017), p. 650. ISSN: 1664-042X. DOI: 10.3389/fphys.2017.00650. URL: https://www.frontiersin.org/article/10.3389/fphys.2017.00650.

[2] Anders Enjin. "Humidity sensing in insects—from ecology to neural processing". In: *Current Opinion in Insect Science* 24 (2017). Neuroscience * Pheromones, pp. 1–6. ISSN: 2214-5745. DOI: https://doi.org/10.1016/j.cois.2017.08.004. URL: https://www.sciencedirect.com/science/article/pii/S2214574517300706.

[3] Mahmut Demir et al. "Walking *Drosophila* navigate complex plumes using stochastic decisions biased by the timing of odor encounters". In: *eLife* 9 (Nov. 2020). Ed. by Agnese Seminara et al., e57524. ISSN: 2050-084X. DOI: 10.7554/eLife.57524. URL: https://doi.org/10.7554/eLife.57524.

[4] Timothy A. Currier and Katherine I. Nagel. "Multisensory control of navigation in the fruit fly". In: *Current Opinion in Neurobiology* 64 (2020). Systems Neuroscience, pp. 10–16. ISSN: 0959-4388. DOI: https://doi.org/10.1016/j.conb.2019.11.017. URL: https://www.sciencedirect.com/science/article/pii/S095943881930128X.

[5] Ye Zhang et al. "Asymmetric ephaptic inhibition between compartmentalized olfactory receptor neurons". In: *Nat Commun* 10 (2019). URL: https://doi-org.ludwig.lub.lu.se/10.1038/s41467-019-09346-z.

[6] Michael H. Alpert et al. "A Circuit Encoding Absolute Cold Temperature in Drosophila". In: *Current Biology* 30.12 (2020), 2275–2288.e5. ISSN: 0960-9822. DOI: https://doi.org/10.1016/j.cub.2020.04.038. URL: https://www.sciencedirect.com/science/article/pii/S0960982220305522.

[7] Zhu Yan Ji Feiteng. "A Novel Assay Reveals Hygrotactic Behavior in Drosophila". In: *PLOS ONE* 10.3 (Mar. 2015), pp. 1–14. DOI: 10.1371/journal.pone.0119162. URL: https://doi.org/10.1371/journal.pone.0119162.

[8] Efrén Álvarez-Salvado et al. "Elementary sensory-motor transformations underlying olfactory navigation in walking fruit-flies". In: *eLife* 7 (Aug. 2018). Ed. by Ronald L Calabrese and Eve Marder, e37815. ISSN: 2050-084X. DOI: 10.7554/eLife.37815. URL: https://doi.org/10.7554/eLife.37815.

[9] Denise Rey and Markus Neuhäuser. "Wilcoxon-Signed-Rank Test". In: *International Encyclopedia of Statistical Science*. Ed. by Miodrag Lovric. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 1658–1659. ISBN: 978-3-642-04898-2. DOI: 10.1007/978-3-642-04898-2_616. URL: https://doi.org/10.1007/978-3-642-04898-2_616.

[10] J Martin Bland and Douglas G Altman. "Multiple significance tests: the Bonferroni method". In: *Bmj* 310.6973 (1995), p. 170.

[11] Liangyu Tao et al. "Statistical structure of locomotion and its modulation by odors". In: *eLife* 8 (Jan. 2019). Ed. by Ronald L Calabrese and K VijayRaghavan, e41235. ISSN: 2050-084X. DOI: 10.7554/eLife.41235. URL: https://doi.org/10.7554/eLife.41235.

[12] Georg Lindgren, Holger Rootzén, and Maria Sandsten. *Stationary stochastic processes for scientists and engineers*. CRC press, 2013, pp. 10–12.

[13] Wai-Ki Ching and Michael K Ng. "Markov chains, second edition". In: *Models, algorithms and applications* (2006), pp. 2–5.

[14] Mikael Nilsson. *First order hidden markov model: Theory and implementation issues (2005)*, Acessed: 1 February 2021. URL: https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A833697&dswid=-3772.

[15] Saeed V. Vaseghi. *Advanced Digital Signal Processing and Noise Reduction, fourth edition*. John Wiley Sons Ltd, 2008, pp. 147–161.

[16] Zach Werkhoven et al. "MARGO (Massively Automated Real-time GUI for Object-tracking), a platform for high-throughput ethology". In: *PloS one* 14.11 (2019), e0224243.

[17] Floris van Breugel, Ainul Huda, and Michael H Dickinson. "Distinct activity-gated pathways mediate attraction and aversion to CO 2 in Drosophila". In: *Nature* 564.7736 (2018), pp. 420–424.

[18] Efrén et al Álvarez-Salvado. *Data from: Elementary sensory-motor transformations underlying olfactory navigation in walking fruit flies, Dryad, Dataset*. Accessed 2021-02-10. 2019. URL: https://doi.org/10.5061/dryad.g27mq71.

[19] Zoé van Havre et al. "Overfitting hidden Markov models with an unknown number of states". In: *arXiv preprint arXiv:1602.02466* (2016).

[20] Noura Dridi and Melita Hadzagic. "Akaike and Bayesian Information Criteria for Hidden Markov Models". In: *IEEE Signal Processing Letters* PP (Dec. 2018), pp. 1–1. DOI: 10.1109/LSP.2018.2886933.

[21] KG Götz and R Biesinger. "Centrophobism in Drosophila melanogaster I: Behavioral modification induced by ether". In: *Journal of Comparative Physiology A* 156.3 (1985), pp. 319–327.

Appendix A

Equipment

Equipment	**Model name**
Pump	Syntech Stimulus Controller CS-55 V2
Daq-board	NI USB-6343
Temperature regulator	Ascon Technologic TLK 38
Temperature/ humidity sensor	SEK-SensorBridge Probe: SHT35-DIS-B
IR-lamp	Advanced Illumination BXXXYY
Smoke-pen	Björnvax
Arena	Designed by Alvarez et al.[8]
Sensor program	Sensirion ControlCenter
Image analysis program	MARGO
Fly diet	Water, sugar syrup, corn flour, yeast, agar, soya flour, propionic acid.

Table 8.1: Equipment and programs used for conducting the behavioural experiments in this thesis.

Supplementary figures

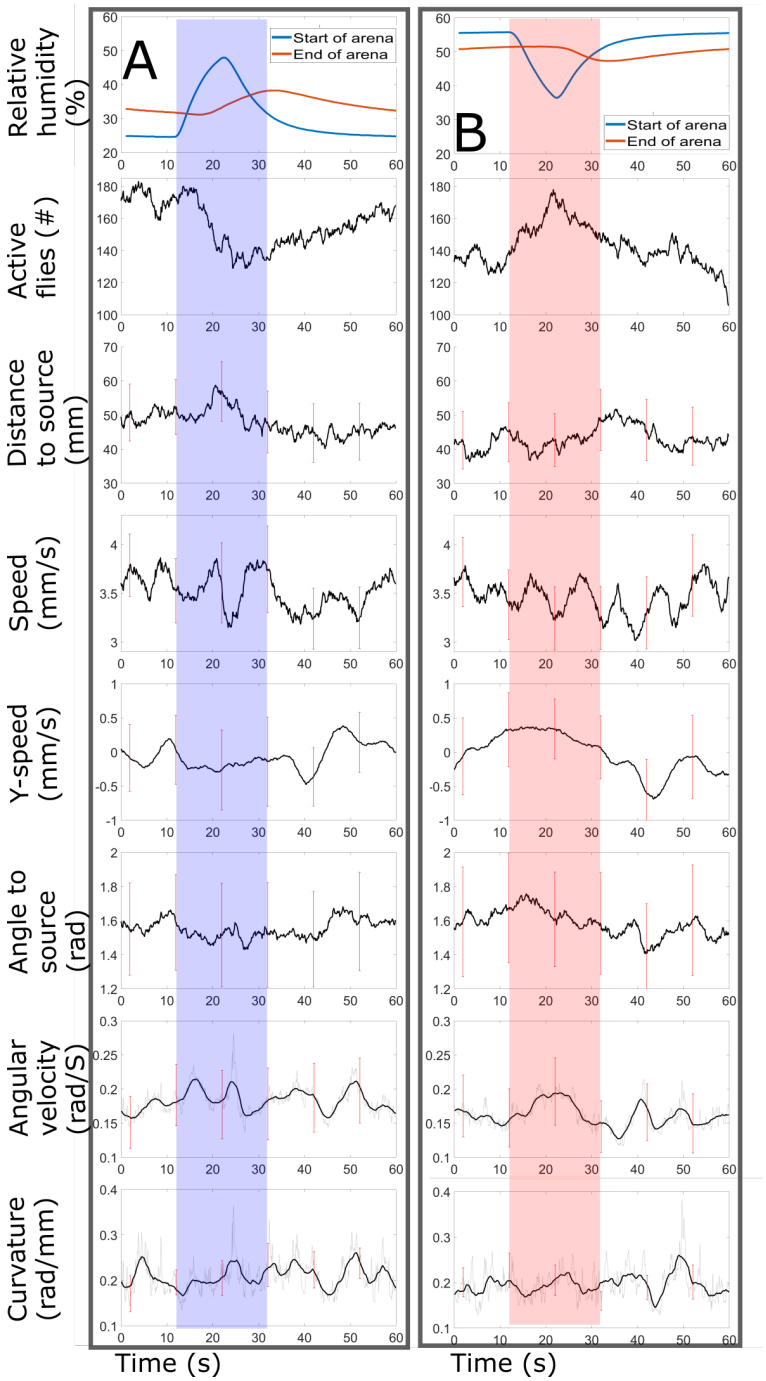


Figure 8.1: The average feature values, described in table 5.1. Calculated for type A and type B trials from the further filtered data set (with removed slow-speed data), containing 5 trials each from 51 flies. A twenty second period of altered humidity is highlighted with blue and red. Confidence intervals are showed as red bars. For angular velocity and curvature the average feature value (gray) is plotted along with a moving average (black) with a window length of three seconds; in order to more prominently show possible trends.

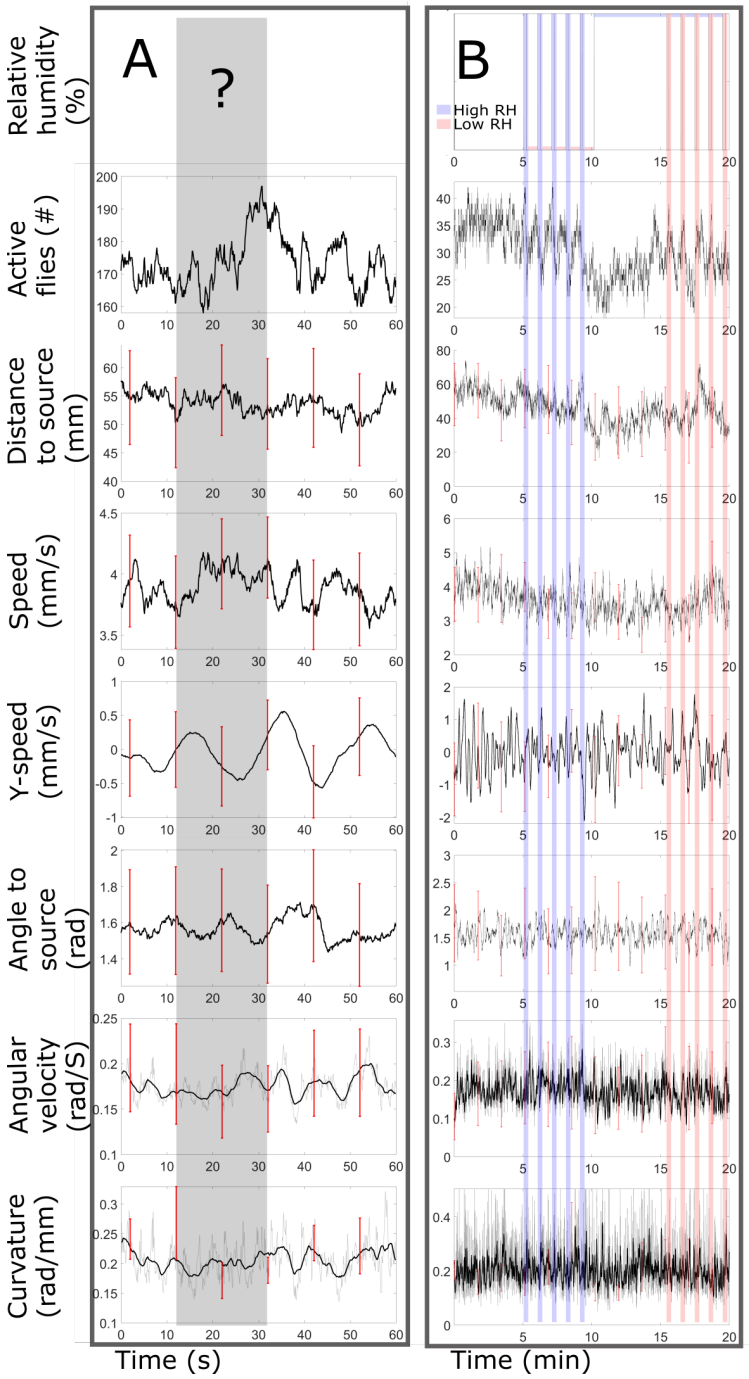


Figure 8.2: The average of feature values, described in table 5.1, using the further filtered data set (with removed slow-speed data). Twenty second periods of altered humidity is highlighted with blue and red and a twenty second period of unknown humidity is highlighted with gray. Confidence intervals are showed as red bars. For angular velocity and curvature the average feature value (gray) is plotted along with a moving average (black) with a window length of three seconds; in order to more prominently show possible trends. (A) The average is calculated from 5 randomly selected one minute periods from each of the 51 flies. (B) The average is calculated from the 51 flies over the course of the entire experiment.

Appendix B

Matlab code for fitting a Hidden Markov model

Initialization

```
function [mu,sigma,P]=initialize(N,M,O)
%INITIALIZE
%Finding the initial parameters for all Gaussian mixture models as
%described in section 4.2.4

%See table 1 for definitions
%INPUT
%N,M: number of states and mixture components
%O: observations with O(speed,angular velocity, sequence)

%OUTPUT
%mu, sigma, P: mean covariance and mixture probability of all Gaussian
%mixture models
%-------------------------------------------------------------------------


F=size(O,3); %Number of sequences
%Concatinating all observation sequences into one matrix O2, to make the
%programming easier
O2=[];
for f=1:F
    O2=[O2;squeeze(O(:,:,f))];
end

%Segmenting the data into states using a k-means clustering algorithm.
%Choosing the best clustering out of 100 runs
[IDX,c,sumd,D]= kmeans(O2,N,'MaxIter',500);
best_sumd=sum(sumd);
 for j=1:100
[idx,c,sumd,D]= kmeans(O2,N,'MaxIter',500);
     if sum(sumd)<best_sumd
         IDX=idx;
         best_sumd=sum(sumd);
     end
 end

```

```
%Fitting a Gaussian Mixture Model to each state
%Choosing the best fit out of 100 runs
mu=zeros(N,M,2);
sigma=zeros(2,2,N,M);
P=zeros(N,M);
options = statset('MaxIter',2000);
for i=1:N
    X_i=O2(IDX==i,:);
    best_loglikelihood=10^90;
    for j=1:100
        GMModel = fitgmdist(X_i,M,'Options',options,'
            Regularize',0.0001);
        if GMModel.NegativeLogLikelihood<best_loglikelihood
            best_loglikelihood=GMModel.NegativeLogLikelihood
                ;
            mu(i,:,:)=GMModel.mu;
            sigma(:,:,i,:)=GMModel.Sigma;
            P(i,:)=GMModel.ComponentProportion;
        end
    end
end
end
```

Fitting process

```
function [parameter_fit] = HMM(N,M,mu,sigma,P,O,
    parameter_fit)
%HMM: Hidden Markov Model
%Fits a hidden Markov model to a data set according to
    algorithm 1.

%INPUT
%See table 1
%N: The number of states
%M: The number of mixture components
%mu, sigma, P: initial estimates of all Gaussian mixture
    models
%O: observations. With O(speed,angular velocity, sequence)
%parameter_fit: A struct containing the resulting parameter
    values from
%several model fits

%OUTPUT
%parameter_fit: A struct containing the resulting parameter
    values from
%several model fits, with the added model fit from this run.
%
```

```
18
19 clearvars -except v angv O mu sigma P w5905all N M
       parameter_fit
20 T=size(O,1); %Number of observations
21 F=size(O,3); %Number of flies
22
23 %Initializing all variables/parameters (with logarithmic
       versions when
24 %necessary. See tables 1 and 2.
25 pi=ones(1,N);
26 pi=pi./sum(pi);
27 A=ones(N,N);
28 A=A./sum(A,2);
29 %B and B2 are matrices containing observation likelihoods
       from
30 %Gaussian mixture models. Introduced to make the programming
        easier.
31 B=zeros(T,N,M,F);
32 B2=zeros(T,N,M,F);
33 alpha=zeros(T,N,F);
34 log_A=log(A);
35 log_P=log(P);
36 log_alpha=zeros(T,N,F);
37 log_beta=zeros(T,N,F);
38 log_gamma=zeros(T,N,F);
39 log_zeta=zeros(T,N,M,F);
40 log_Gamma=zeros(T,N,N,F);
41
42 %The algorithm runs until conversion or for a maximum of 500
        iterations
43 stop_criterion=0.0001;
44 max_iterations=500;
45 for iteration=1:max_iterations
46     %Breaking if the model fit has converged
47     if iteration>3
48         %Introducing temporary variables a, b to calculate
               the improvement
49         %in likelihood.
50         a=log(sum(previous_alpha(end,:,:)));
51         b=log(sum(alpha(end,:,:)));
52         a(isinf(a))=log(realmin);
53         b(isinf(b))=log(realmin);
54         %The stopping criterion showed in algorithm 1
55         if abs(sum(a)-sum(b))/abs(sum(a))<stop_criterion
56             break;
57         end
```

```
58      end
59
60      %Reseting alpha
61      previous_alpha=alpha;
62      alpha=zeros(size(alpha));
63
64      %Calculating B and initializing alpha
65      for f=1:F
66          for n=1:N
67              for m=1:M
68                  B(:,n,m,f)=mvnpdf(O(:,:,f),squeeze(mu(n,m,:))',...
69                  squeeze(sigma(:,:,n,m)));
70                  alpha(1,n,f)=alpha(1,n,f)+pi(n)*P(n,m)*B(1,n,m,f);
71              end
72          end
73      end
74
75      %Initializing log_alpha and log_beta
76      log_alpha(1,:,:)=log(alpha(1,:,:));
77      log_beta(T,:,:)=0;
78
79      %Calculating B2
80      clear B2
81      for n=1:N
82          for m=1:M
83              B2(:,n,m,:)=P(n,m).*squeeze(B(:,n,m,:));
84          end
85      end
86      B2=squeeze(sum(B2,3));
87
88      %logarithmic versions of B,B2
89      log_B=log(B);
90      log_B2=log(B2);
91
92      %Replacing values with numerical underflow
93      log_B(isinf(log_B))=log(realmin);
94      log_B2(isinf(log_B2))=log(realmin);
95      log_P(isinf(log_P))=log(realmin);
96
97      %Calculating forward and backward probabilities (
            log_alpha, log_beta)
98      for t=1:T-1
99          for i=1:N
100             log_alpha(t+1,i,:)=squeeze(log_B2(t+1,i,:))'+...
101             logsumexp(squeeze(log_alpha(t,:,:))+repmat(log_A
```

```
                (:,i),1,F));
102             log_beta(T-t,i,:)=logsumexp(squeeze(log_B2(T+1-t
                    ,:,:))+...
103             squeeze(log_beta(T+1-t,:,:))+repmat(log_A(i,:)
                    ',1,F));
104         end
105     end
106
107     %Replacing values with numerical underflow
108     log_alpha(isinf(log_alpha))=log(realmin);
109     log_beta(isinf(log_beta))=log(realmin);
110
111     %Calculating log_gamma, log_zeta, log_Gamma
112         for t=1:T
113             for i=1:N
114                 log_gamma(t,i,:)=squeeze(log_alpha(t,i,:))
                        +...
115                 squeeze(log_beta(t,i,:))-...
116                 squeeze(logsumexp(squeeze(log_alpha(T,:,:)))
                        )';
117                 if t>1
118                     for m=1:M
119                         log_zeta(t,i,m,:)=...
120                         logsumexp(squeeze(log_alpha(t-1,:,:)
                                )')+...
121                         repmat(log_A(:,i)',F,1)+repmat(log_P
                                (i,m),F,N)+...
122                         repmat(squeeze(log_B(t,i,m,:)),1,N)
                                +...
123                         repmat(squeeze(log_beta(t,i,:)),1,N)
                                ,2)-...
124                         logsumexp(squeeze(log_alpha(T,:,:)))
                                ';
125                     end
126                 end
127                 if t<T
128                     for j=1:N
129                         log_Gamma(t,i,j,:)=squeeze(log_alpha
                                (t,i,:))+...
130                         repmat(log_A(i,j),F,1)+squeeze(
                                log_B2(t+1,j,:))+...
131                         squeeze(log_beta(t+1,j,:))-...
132                         squeeze(logsumexp(squeeze(log_alpha(
                                T,:,:))))';
133                     end
134                 end
135             end
```

```
136         end
137
138      %Replacing values with numerical underflow
139      log_zeta(isinf(log_zeta))=log(realmin);
140      log_gamma(isinf(log_gamma))=log(realmin);
141      log_Gamma(isinf(log_Gamma))=log(realmin);
142
143     %Calculating transition probabilities (log_A)
144     for i=1:N
145         for j=1:N
146             log_A(i,j)=...
147             logsumexp(logsumexp(squeeze(log_Gamma(1:T-1,i,j
                    ,:)))')-...
148             logsumexp(logsumexp(squeeze(log_gamma(1:T-1,i,:)
                    ))');
149         end
150     end
151
152     %Calculating the initial state distributions (pi)
153     pi=(1/F).*exp(logsumexp(squeeze(log_gamma(1,:,:)),2))';
154
155     %Calculating mu
156     zeta=exp(log_zeta);
157     for i=1:N
158         for m=1:M
159         mu_sum=zeros(1,2);
160         for f=1:F
161             mu_sum=mu_sum+[zeta(2:end,i,m,f)'*...
162             O(2:end,1,f),zeta(2:end,i,m,f)'*O(2:end,2,f)];
163         end
164         mu_sum=mu_sum./sum(sum(zeta(2:end,i,m,:)));
165         mu(i,m,:)=mu_sum;
166         end
167     end
168
169     %Calculating sigma
170     zeta=exp(log_zeta);
171     for i=1:N
172         for m=1:M
173             sigmasum=zeros(2,2,T-1);
174             for t=2:T
175                 sigmasum(:,:,t)=((squeeze(O(t,:,:))-...
176                 repmat(squeeze(mu(i,m,:)),1,F)).*...
177                 (repmat(squeeze(zeta(t,i,m,:)),1,2)')*  ...
178                 ((squeeze(O(t,:,:))-repmat(squeeze(mu(i,m,:)
                        ),1,F))');
179             end
```

```
180                 sigmasum=sum(sigmasum,3);
181                 sigmasum=sigmasum./sum(sum(zeta(2:end,i,m,:)));
182                 sigmasum=real(sigmasum);
183                 %Preventing diagonal values from preceeding 0.25
184                 if sigmasum(1,1)<0.25
185                     sigmasum(1,1)=0.25;
186                 end
187                 if sigmasum(2,2)<0.25
188                     sigmasum(2,2)=0.25;
189                 end
190                 %Preventing the covariance matrix from becoming
                        negative
191                 %definite
192                 if det(sigmasum)<=0.0001
193                     sigmasum(1,1)=sigmasum(1,1)+ max(abs(det(
                            sigmasum)),0.00001);
194                     sigmasum(2,2)=sigmasum(2,2)+ max(abs(det(
                            sigmasum)),0.00001);
195                 end
196                 sigma(:,:,i,m)=sigmasum;
197             end
198         end
199
200         %Calculating mixture probabilities (P)
201         for i=1:N
202             for m=1:M
203                 log_P(i,m)=logsumexp(logsumexp(squeeze(log_zeta
                        (2:end,i,m,:)))')-logsumexp(logsumexp(squeeze
                        (log_gamma(2:end,i,:)))');
204             end
205         end
206         %Preventing mixture probabilities from becoming too
                small
207         log_P(log_P<log(10^-10))=log(10^-10); %Blekinge
208
209         %Calculating the exponential of variables
210         P=exp(log_P);
211         P=P./sum(P,2);
212         A=exp(log_A);
213         alpha=exp(log_alpha);
214         beta=exp(log_beta);
215         gamma=exp(log_gamma);
216         big_gamma=exp(log_Gamma);
217         zeta=exp(log_zeta);
218     end
219
220     %Conversion of the model fit
```

```
221
222 %Adding the fitted parameters to the struct parameter_fit
223 %bic=2*(-sum(log(sum(alpha(end,:,:)))))+(N^2+N*M*6+N)*log(T*
    F);
224 %aic=2*(-sum(log(sum(alpha(end,:,:)))))+2*(N^2+N*M*6+N);
225 parameter_fit(N,M).mu=mu;
226 parameter_fit(N,M).sigma=sigma;
227 parameter_fit(N,M).P=P;
228 parameter_fit(N,M).A=A;
229 parameter_fit(N,M).pi=pi;
230 parameter_fit(N,M).log_gamma=log_gamma;
231 %The observation likelihood of the data set
232 parameter_fit(N,M).loglike=sum(log(sum(alpha(end,:,:))));
233 %The observation likelihood of the data sequences
234 parameter_fit(N,M).loglikeseries=log(sum(alpha(end,:,:)));
235 %parameter_fit(N,M).aic=aic;
236 %parameter_fit(N,M).bic=bic;
237 %Overall probability of the fly being in each state
238 parameter_fit(N,M).state_probability=sum(squeeze(sum(gamma
    ,1)),2)./...
239 sum(sum(squeeze(sum(gamma,1)),2));
240 'Converged'
241 end
```