

COLLECTIVE BEHAVIOUR



JACK WALTON

March 2020

*Thesis submitted for the degree of
Doctor of Philosophy*

to the

*School of Maths, Statistics & Physics
Newcastle University
Newcastle upon Tyne
United Kingdom*

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Overview of thesis	2
2	Literature review	5
2.1	Biological function	5
2.2	Mathematical approaches	6
2.2.1	Lagrangian models	6
2.2.2	Eulerian models	9
2.3	Empirical studies	10
2.4	Numerical studies	13
3	Bayesian statistics	15
3.1	Bayesian inference	15
3.2	Markov chain Monte Carlo (MCMC)	16
3.2.1	Gibbs sampling	16
3.2.2	Metropolis-Hastings	16
3.2.3	Convergence Diagnostics	19
4	Directional statistics	21
4.1	Conventions	21
4.2	Visualisation	22
4.3	Summary statistics	23
5	Model fitting	27
5.1	Surf scoter dataset	27
5.1.1	Visualisation	28
5.1.2	Missingness	28
5.2	Model development	31
5.2.1	Vicsek Model	32

5.2.2	Hierarchical Vicsek	33
5.3	Inferring parameters	35
5.4	Inferring missingness	38
5.5	Output	38
A	Data plots	49

1

Introduction

1.1 MOTIVATION

Many of us have been struck by the inherent beauty of animals moving collectively; starlings gathering at dusk in huge numbers to perform the most mesmerising of ballets, the entire flock moving as if some fluid object; fish forming tight milling structures in defence against predation, changing direction in the blink of an eye and with a flash of silver. Collective motion has been observed over many different length scales, and is not constrained only to vertebrates. At the mesoscale, studies have been made of turbulent microbial suspensions (Dunkel *et al.*, 2013). Swarms of locusts, which are capable of occupying up to one fifth of the Earth's land surface during a plague, are also known to exhibit collective motion (Bazazi *et al.*, 2008).

Over the years collective behaviour has become a thriving topic of multidisciplinary research, capturing the imaginations of physicists, biologists, mathematicians and statisticians. Our understanding has evolved significantly from early suggestions that collective behaviour results from thought-transference and telepathy between individuals (Selous, 1931). Though we can often explain why animal aggregations are evolutionary advantageous (Giardina, 2008), much less is known about how these structures are formed and maintained.

Much work has been invested in developing theoretical models which seek to explain emergent behaviour by interactions at an individual level. Such models have shown that individual interactions are sufficient to produce group-level structures. Many different

2 INTRODUCTION



Figure 1.1: A particularly startling example of a starling murmuration, captured near Gretna in the Scottish Borders. Photograph: Owen Humphreys/PA.

simulations, implementing disparate interaction rules, are able to produce behaviour reminiscent of real flocking systems. However, these models have largely only been verified with comparison to empirical observation at a qualitative level, and a thorough quantitative comparison between field data and theory has been lacking.

In recent years technological and methodological advances have made it possible to capture the movements of large groups of animal aggregates (Ballerini *et al.*, 2008). With this data, it is only now that we are in a position to make a robust comparison between model predictions and real-world observations.

1.2 OVERVIEW OF THESIS

We begin this thesis by giving the reader a review of the literature surrounding collective behaviour. Important results and ideas of the field are introduced and discussed. After relaying the main results from the literature we discuss open problems and the future of research in the field.

[Chapter 3](#) introduces the reader to the field of Bayesian statistics. Important results, techniques and algorithms from the field will be outlined as well as problems that the practitioner may encounter, and how they may address these problems.

Now with a grasp of Bayesian statistics, the reader is given a short introduction to

directional statistics in [Chapter 4](#). In this chapter we discuss why standard methods and techniques are inappropriate for handling circular data. With the shortfalls of standard techniques considered, we introduce methods from the study of circular statistics which are appropriate to use with directional data.

2

Literature review

There is a large body of literature relating to the phenomenon of collective behaviour. Particularly unique to this literature is the variety of backgrounds in which the authors are trained. Biologists, physicists, applied mathematicians and statisticians have all made significant contributions to the field.

In this chapter we shall discuss some of the most important ideas and results from the literature surrounding collective behaviour; first by providing a quick overview of the advantages which collective behaviour offers individuals, before discussing the main two approaches to modelling: Eulerian and Lagrangian. After this we shall review previous work which focused on recording and utilising empirical data to inform model selection.

2.1 BIOLOGICAL FUNCTION

Behaving as a group can bring many advantages to the individuals involved. One classically considered benefit of aggregation is an improved defence against predation. Shoaling groups of fish have the ability to confuse predators, as predators have difficulty selecting an individual target (Landau & Terborgh, 1986). As well as a confusion effect groups of individuals can be more vigilant than a single individual, allowing for the earlier detection of predators (Pitcher & Parrish, 1993). Despite these advantages, groups may in fact attract predators (Wittenberger & Hunt, 1985).

As well as providing defence against predation grouping can aid in foraging for food; collections of individuals are able to gather more information about an environment than

6 LITERATURE REVIEW

lone individuals. In addition to foraging, collective motion aids group navigation and migration (Simons, 2004). For birds group navigation often brings an energetic advantage as individuals can work to form aerodynamically efficient shapes (Weimerskirch *et al.*, 2001). As well as these advantages, group living can aid in facilitating reproduction and the rearing of young.

As we have seen, collective living can bring many advantages to the individuals involved. However, we have yet to discuss the underlying mechanisms which generate and sustain the collective behaviours which are seen in nature.

2.2 MATHEMATICAL APPROACHES

Models of collective behaviour can largely be divided into two classes: Lagrangian and Eulerian. These descriptions are analogous to the models of fluid dynamics, where Lagrangian models consider flow in terms of interactions of fluid parcels and Eulerian models consider the changing fluid properties at a given point in space and time. In the context of collective behaviour, Lagrangian models simulate the movements and interactions of individuals and Eulerian models consider the changing properties of a group through space and time.

2.2.1 *Lagrangian models*

So called agent-based models (ABMs), also referred to as Lagrangian models, have proven a useful tool in modelling collective behaviours. In these models the behaviour of an agent is simulated at the individual level. An agent's behaviour is determined by social interactions with neighbouring individuals. Examples of typical interactions include the desire to move in the same direction as neighbours (alignment, or orientation), the desire to avoid collisions (repulsion) and a desire to remain close to neighbours (attraction). As well as simulating social behaviours, ABMs also specify how an individual identifies neighbours. An agent may, for example, identify neighbours as those; within a certain distance (metric interaction); positioned inside a field of vision or as one of a fixed number of closest individuals (topological interaction).

In a pioneering paper, Aoki (1982) developed an ABM to simulate the movements of schooling fish in two-dimensions. Here it was shown that collective behaviour can arise from simple interactions at an individual level and without the need of a leader. The model simulated zonal interactions in which the area around an individual is partitioned into zones of repulsion, alignment and attraction. The partitioning of space in this way is illustrated

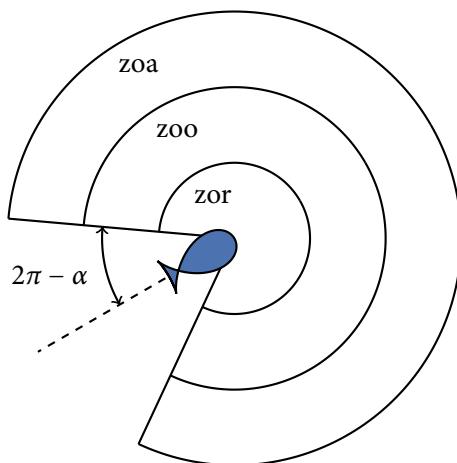


Figure 2.1: An illustration of the area around an agent partitioned into multiple zones. zor: zone of repulsion, zoo: zone of orientation (or alignment), zoa: zone of attraction. The missing segment behind the agent represents a blind zone into which it cannot see.

in [Figure 2.1](#), and has remained a popular idea in following literature. As well as zonal interactions this model accounted for fish having incomplete fields of vision, that is: a blind spot into which they cannot see. The simulation of a blind spot was utilised in further studies. Later, other models were also devised to simulate fish schools (Okubo, 1986; Huth & Wissel, 1992).

Following this, Reynolds (1987) formulated a theoretical model, motivated by the production of computer animations, which described the movement of flocking birds in three-dimensional space. To produce more aesthetically pleasing animations, the software, “Boids”¹ implemented additional sophistications such as banking during turns. This focus on developing simulations which produce elegant behaviour made rigorous scientific analysis difficult. Interestingly, Tim Burton’s 1992 Batman Returns used a modified version of the Boids software to simulate animations of bat swarms and penguin flocks. Substantially more complex than Boids was the software package MASSIVE (Multiple Agent Simulation System in Virtual Environment), developed by Stephen Regelous for Peter Jackson’s Lord of the Rings trilogy. This software was used in the striking battle sequences of the trilogy, where each individual orc, elf and miscellaneous creature was simulated as an agent with behaviour governed by a set of interaction rules (Robbins, 2017).

Not motivated by Lord of the Rings, but instead motivated by research within statistical physics, Vicsek *et al.* (1995) introduced a simple two-dimensional model in which self-propelled particles move with a fixed absolute velocity and align with neighbours within an interaction radius. This model is commonly referred to as “Vicsek Model” (VM). Despite

8 LITERATURE REVIEW

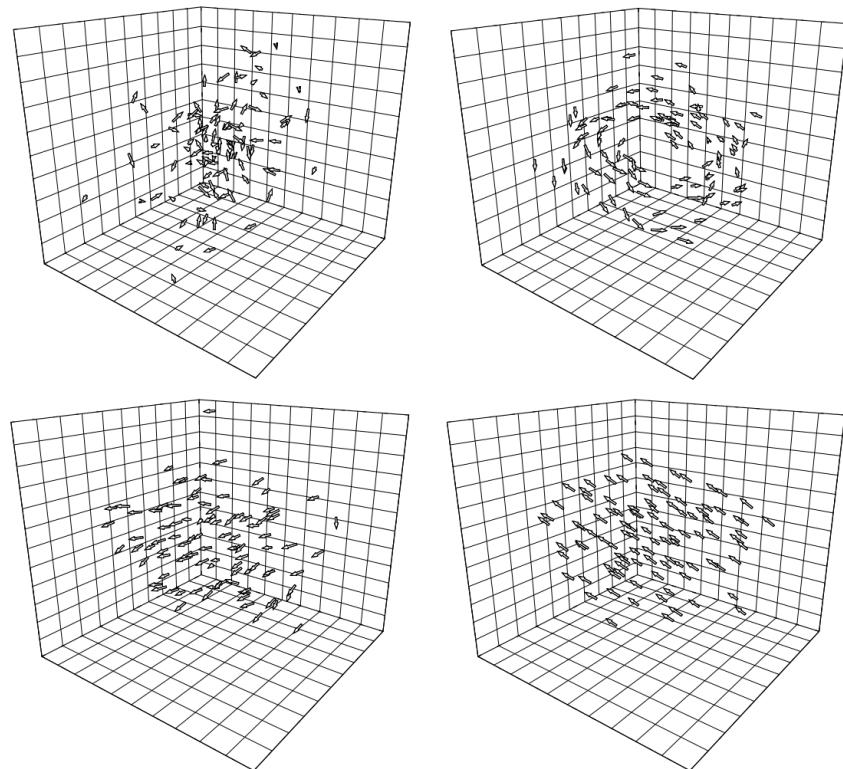


Figure 2.2: Taken from Couzin *et al.* (2002), the different steady-state solutions (swarm, torus, dynamically parallel and highly parallel) obtained by making small changes to model parameters of a three-dimensional flocking model.

its simplicity this model produces complex behaviour resembling that of a real biological system. Vicsek *et al.* (1995) investigated the phase transition between ordered and disordered motion as the density of particles and noise in the system varied. This transition from order to disorder is an example of a spontaneously breaking (rotational) symmetry, as the group has no preferred direction of motion *a priori*, but under simulation each group chooses some arbitrary direction to travel in. Because of this, the Vicsek model stands as an apparent violation of the Mermin–Wagner Theorem, which states that continuous symmetries cannot be spontaneously broken by systems that are able to achieve long range order in dimensions $d \leq 2$ (Mermin & Wagner, 1966). However, VM is out of equilibrium and Mermin–Wagner only applies to systems in equilibrium — so there is no violation after all.

Later, models were developed to explore the movements of mammals and other vertebrate groups. Using a 3-dimensional model that follows the zonal approach of Aoki (1982), Couzin *et al.* (2002) showed major group-level behavioural changes as minor changes in individual interaction rules were made. With small changes in the model parameters, groups

transitioned from disordered, swarm-like behaviour, to toroidal milling structures, to forming dynamic and highly parallel groups, as illustrated in [Figure 2.2](#). In addition to this the author's simulations demonstrated evidence of the collective memory of a group, such that previous group structure influences future behaviour as interactions change.

Further research was made by Couzin *et al.* (2005) which investigated how leaders influence the motion of travelling groups. A zonal repulsion-attraction-alignment model was used as the basis for this work. Here, though, a proportion of the flock were given information about a preferred direction of motion, and so balanced their social interactions with the desire to move in this direction. Individuals in the flock did not know which members of the group, if any, had information. Simulations showed that only a small proportion of leaders are necessary to guide groups with a high degree of accuracy. Further results investigated how groups of individuals make collective decisions in the face of conflicting desires.

As a method for exploring collective behaviour, Lagrangian models are very appealing in their intuitiveness and in the ease of implementing explicit behavioural rules. Though for many years the simulation and exploration of these models was limited by computing power; modern computation allows for the simulations of large groups over many time steps. With these advances in computing, and a growing interest in the field, a significant proportion of the literature focuses on the analysis and exploration of agent-based models.

2.2.2 Eulerian models

Sometimes known as continuum models, Eulerian models are complementary to the Lagrangian method and work at a coarse-grained level (Giardina, 2008). Eulerian models are typically constructed of a set of partial differential equations which describe how density and other properties of a group develops over time. This approach to modelling is often used to investigate the long-time spatial and density properties of groups.

One such Eulerian approach by Guernon & Levin (1993) modelled the movements of large groups of wildebeests. The predictions of the model were compared with aerial observations of migrating wildebeest in the Serengeti. The large-scale front patterns seen in the aerial photography were reproduced by the model.

Later, Toner & Tu (1998) introduced a quantitative continuum theory of flocking. There are similarities between the hydrodynamic equations introduced by the authors and the Navier–Stokes equation for simple incompressible fluids. This model is capable of predicting the existence of an ordered phase of motion, as is often observed in the field, and

propagating density waves. Detailed analysis of the model is made using techniques (e.g. dynamical renormalization group) from nonequilibrium condensed matter physics and can be used to make quantitative predictions of the properties of the long-distance, long-time behaviour of the ordered state.

Eulerian models have also been used to analyse vortex solutions (Topaz & Bertozzi, 2004) and stationary clump solutions (Topaz *et al.*, 2006).

However, the Eulerian approach is limited. Most analyses are restricted to a single dimension and the approach has not proven appropriate for modelling groups of low densities. With this in mind, and with the advantages of the Lagrangian approach, in this thesis we will concentrate entirely on modelling in the Lagrangian framework.

2.3 EMPIRICAL STUDIES

Models of collective motion rely on aprioristic assumptions about the properties and behaviours of individuals. We also understand that the emergence of a biologically realistic pattern from simulating a theoretical model is not sufficient evidence of model correctness. That is, the emergence of a desired pattern is not evidence that a model is correctly capturing the interactions of individuals. This observation is further compounded by the understanding that models employing different local interactions can produce similar looking behaviour at the group level. As such, real data describing the dynamics of animal aggregations is essential to assess the validity and appropriateness of theoretical models and their assumptions.

Thorough comparison between real data and model has proven difficult largely because of the scarcity of appropriate data. The collection of suitable data can be a complicated and convoluted process. Taking observations in the field is technically demanding, requiring the precise calibration of sensitive measurement equipment, not to mention the additional difficulty of the typically three-dimensional nature of animal aggregations. Collecting data in a laboratory setting seems an obvious workaround, however this imposes restrictions on the types of behaviour which can be captured. A laboratory may be an appropriate environment to capture the movements of fish in a tank, but it certainly isn't appropriate to capture the movements of flocking birds. Despite the difficulties associated with collecting data, significant effort has been made to track the movements and dynamics of groups of individuals.

Initial work was limited to tracking small numbers of individuals in groups. In these studies individuals were not linked through frames and hence the collected data had no

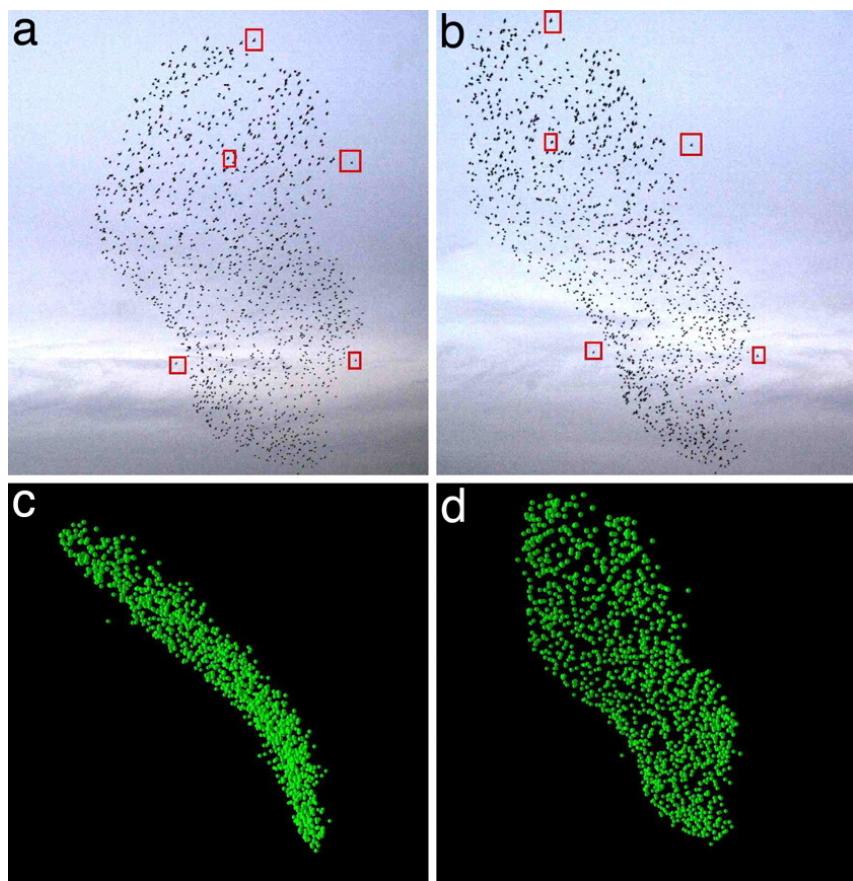


Figure 2.3: A flock of 1246 starlings reconstructed in three dimensions. Photographs taken at the same instant but 25m apart (a–b) are used to reconstruct their three dimensional positions (c–d). To perform reconstruction Ballerini *et al.* (2008) needed to match each bird in (a) to its corresponding image in (b). The red squares show five matched pairs of birds.

dynamic component. The first breakthrough came from Cullen *et al.* (1965) who used stereo photography to record the positions of fish in three dimensions.

Fish are an appealing subject to study as experiments are easily conducted in a laboratory setting. Furthermore, the movements of fish can effectively be restricted to two dimensions by conducting the experiments in shallow water. Because of these benefits, further research also concentrated on fish (Partridge *et al.*, 1980; Van Long *et al.*, 1985). Having collected empirical data, these studies investigate properties such as the distance of individuals to their nearest neighbour, or the direction toward their nearest neighbour. Empirical studies were also made of small groups of flocking birds, with similar statistics and properties realised (Major & Dill, 1978; Budgey, 1998).

More recently, a breakthrough study by Ballerini *et al.* (2008) reconstructed the three dimensional positions of flocks of starlings consisting of up to 2600 individual members

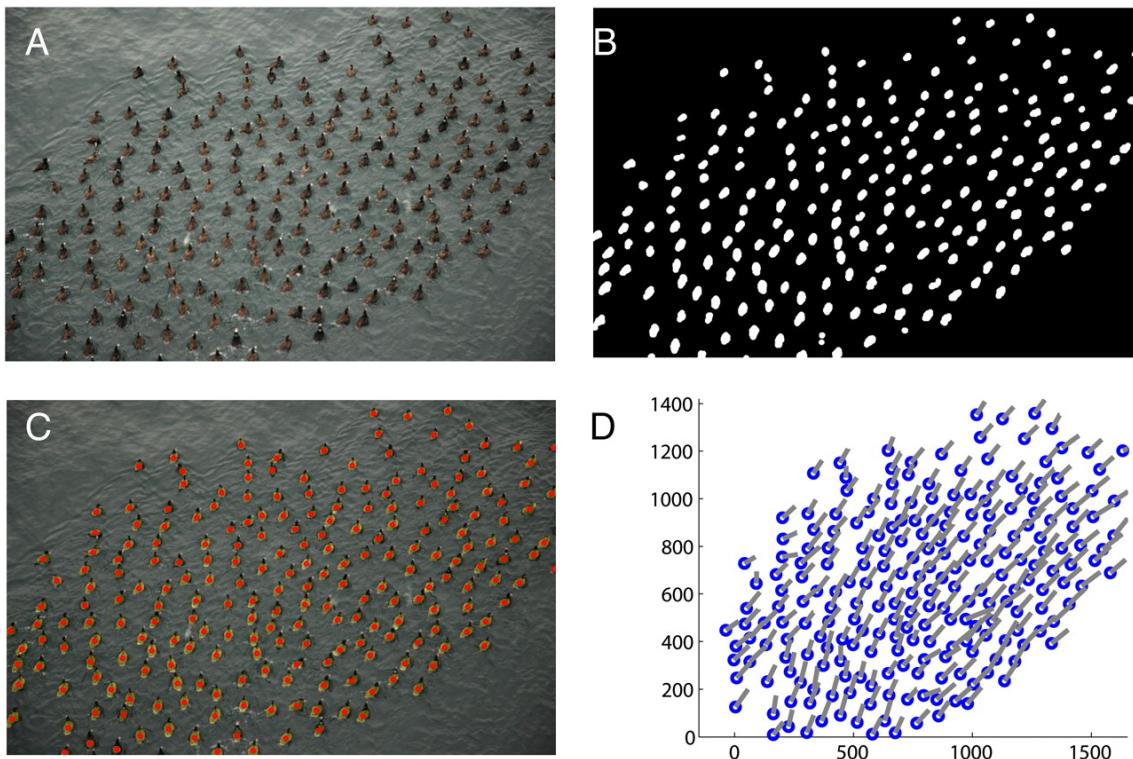


Figure 2.4: Image field data and the process of transformations to extract positions of a flock of surf scoters (Lukeman *et al.*, 2010).

(Figure 2.3). To collect the data the authors used a combination of stereometric and computer vision techniques. Having collected and extracted the dataset, the authors began by constructing angular density plots of nearest neighbours. These plots revealed a strong anisotropy in the flock, with a lack of nearest neighbours positioned along the direction of motion. Having investigated how this anisotropy decays as a function of nearest neighbour, the authors concluded that interactions are not dependent on metric distance (interactions with agents within a fixed distance), as most models in the literature assume, but on a topological distance (interaction with a fixed number of closest agents, irrespective of distance). This analysis suggested that on average a starling interacts with between six and seven of its closest neighbours.

A significant contribution to the field was made by Lukeman *et al.* (2010), whom collected and analysed data of large numbers of diving ducks interacting on the surface of a lake. Crucially, this dataset tracked individuals between frames and therefore allowed the reconstruction of a bird's trajectory through space and time. This data showed an increase by factor of ten the number of individuals which could be reliably tracked though time (Lukeman, 2009). The extracted dataset was investigated and plots of nearest neighbour densities

were realised. It was observed from these plots that the highest density of neighbours occurs at some preferred distance, in front of and behind the focal bird. Further analysis fitted varying zonal models to the data. Optimal parameters were fitted to best reproduce the spatial neighbour densities and density as a function of circumferential distance. It was concluded that a zonal repulsion-alignment-attraction model with an additional frontal interaction was best able to reproduce the desired spatial and angular neighbour distributions.

Following this, Katz *et al.* (2011) investigated two and three fish shoals of golden shiners. Data was recorded by placing fish in shallow tanks of water and using custom tracking software to convert video footage into data describing the centre of mass of fish through time. Working in a classical mechanics framework the authors map the effective forces acting on a focal fish as a function of position and velocity. It was found that the dominant interaction between fish was the regulation of their speed. No evidence was found of explicit alignment of direction between individuals; instead, alignment occurred as a product of attraction and repulsion between individuals. Pairwise interactions were seen to predict the spatial distributions of neighbours, and this observation was validated for shoals of 10 and 30 individuals.

Analysis of empirical data has so far focused on properties of individuals such as nearest neighbour distances or angular neighbour densities. Research has then focused on fitting models which are best able to replicate these properties. With technological advances we expect that more and more empirical data will become available in the future.

2.4 NUMERICAL STUDIES

Mann (2011) acknowledged that an important aspect of model fitting is knowing the associated uncertainty of inferred parameters. The author continued to stress the importance of quantifying uncertainty in parameter inference on collective behaviour models, as the associated empirical datasets often have high levels of noise. With the importance of capturing uncertainty in mind, Mann demonstrated a fully Bayesian approach to parameter inference on data simulated from a collective behaviour model. Here, in contrast to the empirical studies made, parameters were inferred on their ability to reproduce the trajectories of agents, as opposed to the ability to reproduce epiphenomena such as nearest neighbour densities or angular neighbour plots.

The agents in Mann's model moved under a weighted sum of alignment and attraction. After ten timesteps the simulated data transitioned from disordered motion to a steady state rotating mill. The author then compared the ability to infer the weighting parameter,

14 LITERATURE REVIEW

interaction radius and other properties of the agents in two situations: before and after the achievement of steady state. It was discovered that the interaction radius could not be reliably inferred when the agents had formed the rotating mill structure, although it could be inferred in the disordered motion before steady state. This result can be understood by considering that stable groups present a limited number of particle configurations, and are therefore less informative than out of equilibrium groups.

3

Bayesian statistics

In this thesis we utilise techniques from Bayesian inference to fit theoretical models of collective behaviour to real data. Bayesian inference allows the practitioner to capture uncertainty about fitted model parameters. In addition to this the Bayesian framework permits flexible model structures and potential inclusion of expert information via the prior distribution. With this we seek to fit empirical data to a generalisation of a popular model from the literature.

In this chapter we shall introduce and give overviews of some important concepts of Bayesian inference, outline schemes which can be used to infer model parameters, and discuss some of the problems which may arise and how we might address them.

3.1 BAYESIAN INFERENCE

Using Bayesian inference we wish to quantify beliefs and uncertainties about parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_p)^T$, using data \mathbf{x} which we observe. Given this observed data, the likelihood function for the parameters is defined

$$L(\boldsymbol{\theta} | \mathbf{x}) = f(\mathbf{x} | \boldsymbol{\theta}).$$

The likelihood quantifies the probability distribution of the data in terms of the parameters. We may then specify our prior knowledge about the parameters $\boldsymbol{\theta}$ through the prior distribution $\pi(\boldsymbol{\theta})$. Bayes Theorem can then be used to incorporate both the likelihood function

and our prior beliefs, to form the posterior distribution

$$\pi(\boldsymbol{\theta} | \mathbf{x}) = \frac{\pi(\boldsymbol{\theta})L(\boldsymbol{\theta} | \mathbf{x})}{\int_{\boldsymbol{\theta}} \pi(\boldsymbol{\theta})L(\boldsymbol{\theta} | \mathbf{x}) d\boldsymbol{\theta}}. \quad (3.1)$$

Because the integral in the denominator is not a function of $\boldsymbol{\theta}$, we may consider it a constant of proportionality and express our posterior beliefs as proportional to the product of the likelihood and prior, that is

$$\begin{aligned}\pi(\boldsymbol{\theta} | \mathbf{x}) &\propto \pi(\boldsymbol{\theta}) \times L(\boldsymbol{\theta} | \mathbf{x}) \\ \text{posterior} &\propto \text{prior} \times \text{likelihood}\end{aligned}$$

3.2 MARKOV CHAIN MONTE CARLO (MCMC)

For the most part, the normalising constant (given in the denominator of Equation (3.1)) will have multiple dimensions, not produce a density function of standard form, and be difficult to evaluate in all but the most trivial cases. Markov chain Monte Carlo algorithms provide methods to sample from the targeted density $\pi(\boldsymbol{\theta} | \mathbf{x})$, whilst avoiding evaluating the troublesome normalising constant.

3.2.1 Gibbs sampling

One may use the full conditional distributions of parameters to sample from a multivariate density. Doing so is to implement the Gibbs algorithm. So, instead of sampling from the full posterior, we sample from the conditional posteriors of the parameters one at a time. The Gibbs algorithm is useful when the conditional densities can be expressed in standard form and are easy to sample from.

Say we wish to target the density $\pi(\boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_p)^T$, where the full conditional densities are $\pi(\theta_i | \theta_1, \theta_2, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_p)$, for $i = 1, \dots, p$, then we may use the Gibbs sampler, as described in Algorithm 1.

3.2.2 Metropolis-Hastings

The Metropolis-Hastings algorithm is another MCMC scheme. The algorithm was introduced by Metropolis *et al.* (1953), and this work was later generalised by Hastings (1970). The algorithm works by constructing a Markov chain which has stationary distribution equivalent to the distribution of interest.

Algorithm 1 Gibbs Algorithm targeting the density $\pi(\boldsymbol{\theta} | \mathbf{x})$.

- o. Initialise chain with $\boldsymbol{\theta}^0$. Set $j = 1$.
- 1. Generate $\boldsymbol{\theta}^j$ from $\boldsymbol{\theta}^{j-1}$ by simulating from:

$$\begin{aligned}\boldsymbol{\theta}_1^j &\sim \pi(\boldsymbol{\theta}_1^j | \boldsymbol{\theta}_2^{j-1}, \dots, \boldsymbol{\theta}_p^{j-1}, \mathbf{x}) \\ \boldsymbol{\theta}_2^j &\sim \pi(\boldsymbol{\theta}_2^j | \boldsymbol{\theta}_1^j, \boldsymbol{\theta}_3^{j-1}, \dots, \boldsymbol{\theta}_p^{j-1}, \mathbf{x}) \\ &\vdots \\ \boldsymbol{\theta}_p^j &\sim \pi(\boldsymbol{\theta}_p^j | \boldsymbol{\theta}_1^j, \dots, \boldsymbol{\theta}_{p-1}^{j-1}, \mathbf{x})\end{aligned}$$

- 2. Increment j to $j + 1$. Repeat from step 1.
-

Algorithm 2 The Metropolis-Hastings used to target the posterior distribution $\pi(\boldsymbol{\theta} | \mathbf{x})$.

- o. Initialise chain with $\boldsymbol{\theta}^0$. Set $j = 1$.
 - 1. Propose $\boldsymbol{\theta}^*$ by sampling from $q(\cdot | \boldsymbol{\theta}^{j-1})$, where q is some proposal distribution
 - 2. Construct the acceptance probability $\alpha(\boldsymbol{\theta}^* | \boldsymbol{\theta}^{j-1})$ as
- $$\alpha(\boldsymbol{\theta}^* | \boldsymbol{\theta}) = \min \left\{ 1, \frac{\pi(\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}^{j-1})} \frac{L(\boldsymbol{\theta}^* | \mathbf{x})}{L(\boldsymbol{\theta}^{j-1} | \mathbf{x})} \frac{q(\boldsymbol{\theta}^{j-1} | \boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^* | \boldsymbol{\theta}^{j-1})} \right\}.$$
- 3. With probability $\alpha(\boldsymbol{\theta}^* | \boldsymbol{\theta}^{j-1})$ set $\boldsymbol{\theta}^j = \boldsymbol{\theta}^*$, otherwise set $\boldsymbol{\theta}^j = \boldsymbol{\theta}^{j-1}$.
 - 4. Increment j to $j + 1$. Repeat from step 1.
-

The algorithm begins by initialising the chain with parameters $\boldsymbol{\theta}^0$. Next, the algorithm proposes new values $\boldsymbol{\theta}^*$ from a proposal distribution, $q(\boldsymbol{\theta}^* | \boldsymbol{\theta}^{j-1})$, which is chosen to have the same support as the target distribution. After this, the proposed values $\boldsymbol{\theta}^*$ are either accepted or rejected, depending on the evaluation of the acceptance probability $\alpha(\boldsymbol{\theta}^* | \boldsymbol{\theta}^{j-1})$. Because the acceptance probability depends on a ratio of $\pi(\cdot | \mathbf{x})$, the normalising constants cancel and therefore the target distribution only has to be known to a constant of proportionality. Metropolis-Hastings is described more formally in [Algorithm 2](#).

Choosing a Proposal Distribution

The practitioner must choose a suitable proposal distribution $q(\boldsymbol{\theta}^* | \boldsymbol{\theta})$. Ideally the choice of proposal distribution will give rapid convergence to $\pi(\boldsymbol{\theta} | \mathbf{x})$ and efficiently explore the

support of $\pi(\boldsymbol{\theta} | \mathbf{x})$.

A special case of Metropolis-Hastings arises when the proposal distribution is symmetric, that is

$$q(\boldsymbol{\theta}^* | \boldsymbol{\theta}) = q(\boldsymbol{\theta} | \boldsymbol{\theta}^*).$$

In this case we observe cancellation in the acceptance ratio, as it simplifies to become

$$\alpha(\boldsymbol{\theta}^* | \boldsymbol{\theta}) = \min\left\{1, \frac{\pi(\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}^{j-1})} \frac{L(\boldsymbol{\theta}^* | \mathbf{x})}{L(\boldsymbol{\theta}^{j-1} | \mathbf{x})}\right\}.$$

Another special case of Metropolis-Hastings is the random walk sampler. In this case proposals are realised as

$$\boldsymbol{\theta}^* = \boldsymbol{\theta}^{j-1} + \boldsymbol{\omega}^{j-1},$$

where the $\boldsymbol{\omega}$ are drawn from

$$\boldsymbol{\omega}^{j-1} \sim \mathcal{N}_p(\mathbf{0}, \Sigma),$$

and \mathcal{N}_p denotes a p -dimensional multivariate normal distribution. The parameter Σ is called the tuning parameter and controls how the chain moves around the parameter space. Mixing describes how efficiently the chain moves around the sample space and how long it takes for the chain to converge to the target distribution.

Crucially then, the parameter Σ controls the mixing of the chain. So, naturally, we wish to select some optimum Σ to try and improve mixing. If the target distribution is Gaussian, it has been shown that 0.234 is an optimum acceptance probability (Roberts & Rosenthal, 2001). In an attempt to tune Σ to obtain the optimum acceptance probability, a common technique is to use

$$\Sigma = \frac{2.38^2}{p} \widehat{\text{Var}}(\boldsymbol{\theta} | \mathbf{x}).$$

Blocking Parameters

In the schemes considered so far the proposal, acceptance and rejection of the entire parameter space $\boldsymbol{\theta}$ happened simultaneously. This approach becomes inefficient for high-dimensional problems. Consider that as the dimension of the problem increases, the chances of proposing a value θ_i^* in the tails of the posterior distribution increases. Increasing the chance of proposing a component θ_i out in the tails of the distribution in turn decreases the acceptance rate of the chain and leads to slower convergence.

To overcome this problem the parameter space $\boldsymbol{\theta}$ can be split into blocks of parameters

$\theta_1, \dots, \theta_d$ which are proposed and accepted or rejected separately. There are no theoretical results which determine how best to block the parameter space, though typically blocks are chosen to contain related parameters.

Blocking doesn't reduce the risk of a parameter being proposed in the tails of the distribution, however when such a proposal does occur only a subset of all the chains are affected. In this way blocking can alleviate the lower acceptance rates that are associated with high dimensionality.

There is, however, an additional computational cost that comes with blocking parameters. Consider that if the parameter space θ is partitioned into d blocks, then for each iteration of the scheme the acceptance ratio, and therefore the likelihood, proposal density and prior distribution must be evaluated d times. In application the practitioner will likely seek some compromise between acceptance rate and computation time.

3.2.3 Convergence Diagnostics

Though there are theoretical methods to assess the convergence of chains, it is an attractive idea to analyse the output of our schemes in an attempt to assess whether the chains have converged. One of the simplest informal methods to assess convergence is to inspect the trace plots of the scheme and check for any irregularities. It is also good to use autocorrelation plots to assess autocorrelation between samples at different lags.

One way to lower autocorrelation between samples is to thin the output. When thinning, every k -th sample from a chain is kept and the remaining samples are discarded. Another common technique is to allow for a burn-in period. The purpose of a burn-in period is to discard any samples from before the chain has converged.

4

Directional statistics

Circular data arises naturally in the study of collective behaviour. In particular, a researcher in the field will undoubtedly encounter data describing the direction of motion of individuals. Given any dataset, the first instinct of the practitioner is to summarise and visualise the data. However, such a researcher should proceed with caution — circular data cannot be treated as if it were its linear counterpart.

In this short chapter we shall consider why standard techniques, methods and summaries are inappropriate to use with circular data. After this realisation, we proceed to introduce some useful techniques which can be used to handle and visualise directional data.

4.1 CONVENTIONS

Directions can be represented as rotations with respect to some zero-direction, or origin. The practitioner is free to chose the zero-direction as they feel appropriate. In a similar way, the practitioner may choose whether a clockwise or anti-clockwise rotation is taken as the positive direction.

Recall that angles may be represented in units of degrees or radians. To convert between degrees and radians we may multiply by a factor of $\pi/180^\circ$.

In this thesis we shall define the zero angle as the direction from $(0, 0)$ and along the positive x -axis. For the most part, we shall measure angles in units of radians, and take anti-clockwise rotations as the positive direction. The schematics of this setup are illustrated in

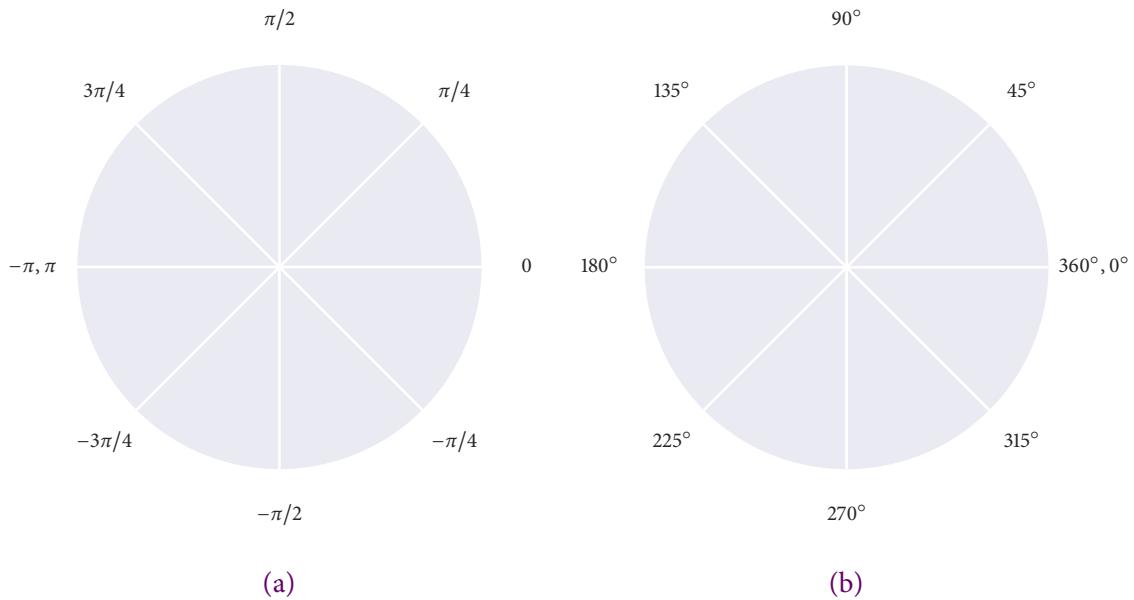


Figure 4.1: Comparing (a): radian and (b): degree measuring conventions adopted in this thesis.

Figure 4.1(a). Occasionally, we shall appeal to degrees and their comparative intuitiveness, and in these cases we shall use the setup illustrated in Figure 4.1(b).

4.2 VISUALISATION

In possession of a dataset, one of the first instincts of the scientist is to visualise their data. The researcher is undoubtedly familiar with a large number of graph types. Yet choosing the most suitable graph to display a given dataset is crucial in making an informative plot.

Traditional histograms are not very good for visualising directional data. Polar histograms (sometimes known as rose plots) make the visualisation of such data easier. Instead of using bars, as the histogram does, the rose plot bins data into sectors of a circle. The area of each sector is proportional to the frequency of data points in the corresponding bin.

To advocate the advantages of the rose plot we shall visualise two randomly generated datasets. The first dataset consists of 100 realisations from a uniform $U(-\pi, \pi)$ distribution, and the second dataset consists of 10 000 draws from a normal $N(0, 1)$ distribution.

In Figure 4.2 we visualise the two datasets using traditional histogram plots. From this figure we get a good idea of the distribution of the data, however we get no sense of direction. The traditional histogram leaves us to interpret the directions ourselves.

In Figure 4.3 we visualise the same data as before. Here we also get a good idea of how the directions are distributed. However, using the rose plot means we get a very intuitive

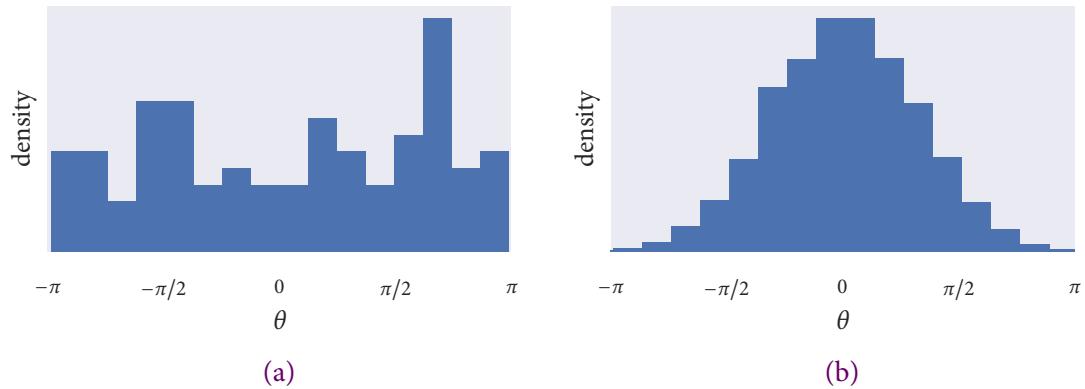


Figure 4.2: Using histograms to visualise (a): 100 samples from $U(-\pi, \pi)$ and (b): 10 000 draws from $N(0, 1)$.

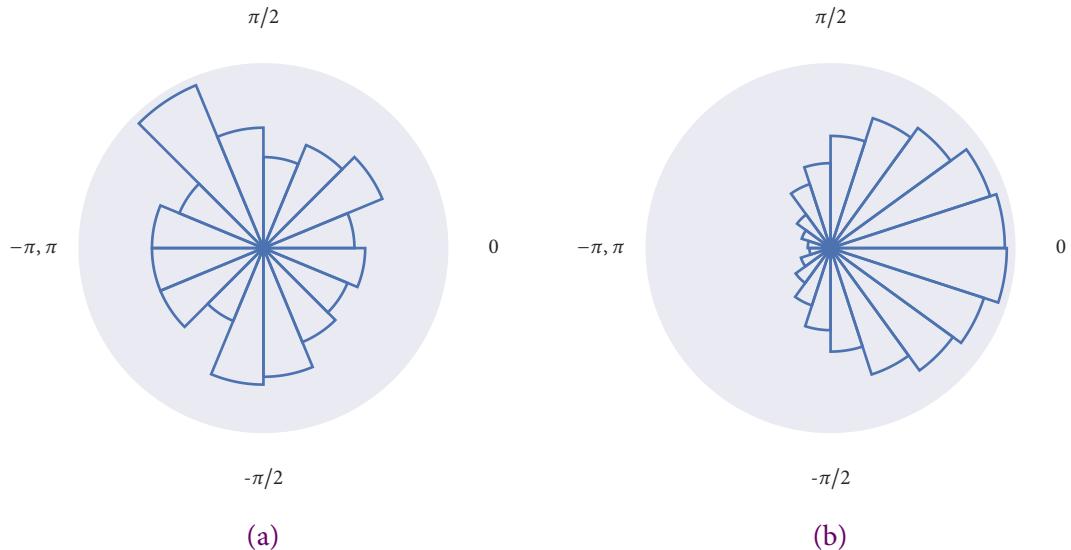


Figure 4.3: Using polar histograms to visualise the dataset generated from (a): $U(-\pi, \pi)$ and (b): $N(0, 1)$.

representation of direction. We therefore consider the rose plot a very useful method of displaying circular data which is easy to interpret.

4.3 SUMMARY STATISTICS

Summary statistics are a useful tool to give an idea of the general characteristics of a dataset. Probably the first statistic which we learn to compute is the arithmetic mean. The arithmetic mean, however, is not an appropriate measure to use with circular data.

Consider that we wish to take an average of the angles 10° and 350° . Using the arith-

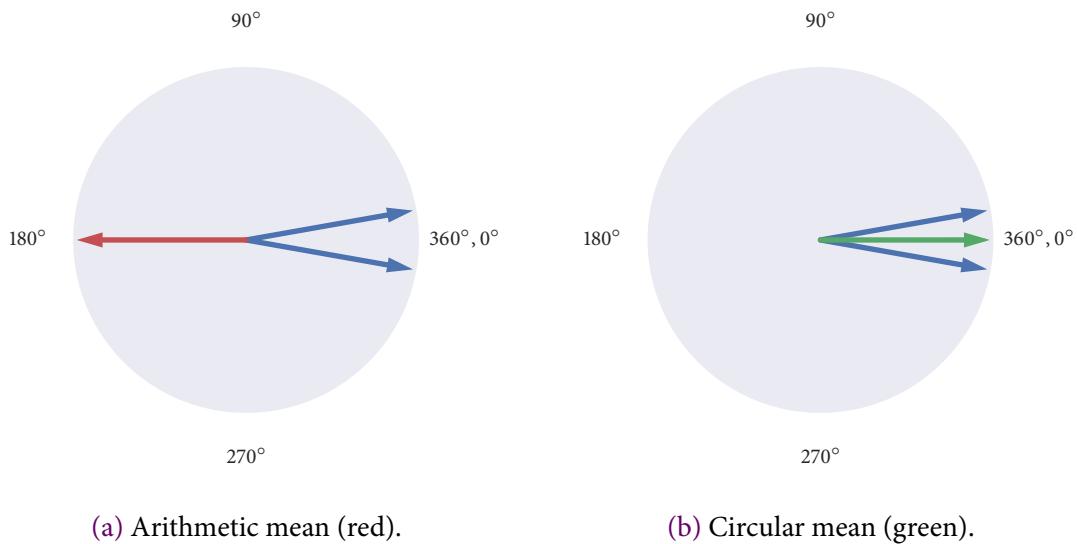


Figure 4.4: Computing the average of two angles with two different means. The blue arrows (10° and 350°) represent the directions to be averaged and the green and red arrows show the average computed by the method.

metic mean we compute an average of 180°— however, this average points in the opposite direction to which we intuitively expect. In Figure 4.4(a) we visualise this surprising result.

Before introducing the circular mean it is first necessary to introduce the atan2 function. The atan2 function dates back to the Fortran programming language (Organick, 1966). It was introduced to overcome some of the inconveniences inherent in the atan (or \tan^{-1}) function. For starters, the inverse tangent function has codomain $(-\pi/2, \pi/2)$, though we are often interested in directions in the range $(-\pi, \pi]$. In addition to this, the arctan function is not quadrant-aware — it cannot distinguish between opposite directions (directions which differ by π radians). As an example, consider calculating the direction from the x -axis to the ray extending from the origin to $(1, 1)$. Naturally, we'd reach for \tan^{-1} to compute the angle as $\tan^{-1}(1/1) = \pi/4$, as expected. Now, consider that we wish to calculate the direction from the x -axis to the ray extending from $(0, 0)$ to $(-1, -1)$. By inspection, or intuitively, we expect an answer of $-3\pi/4$ — however, we compute the answer as $\tan^{-1}(-1/-1) = \pi/4$. The angle calculated using the inverse tangent function points in the opposite direction to what we expect.

The atan2 function, however, does not have these shortcomings. The function is constructed to be quadrant-aware: correcting the computations of \tan^{-1} to return the directions we intuitively expect. It does this by adding a correction term that depends on which quadrant contains the point (x, y) . The handling of the four quadrants by atan2 is visualised

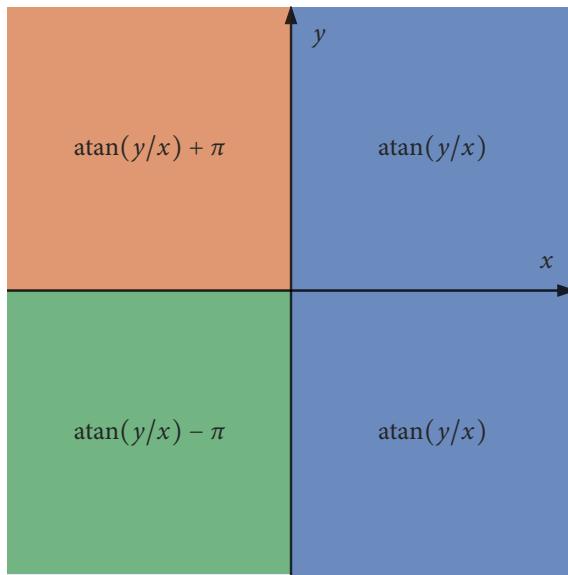


Figure 4.5: An illustration of the quadrant corrections made by atan2.

in Figure 4.5. With these considerations, atan2 can be realised as:

$$\text{atan2}(y, x) = \begin{cases} \text{atan}(y/x) & \text{if } x > 0, \\ \text{atan}(y/x) + \pi & \text{if } x < 0 \text{ and } y \geq 0, \\ \text{atan}(y/x) - \pi & \text{if } x < 0 \text{ and } y < 0, \\ \pi/2 & \text{if } x = 0 \text{ and } y > 0, \\ -\pi/2 & \text{if } x = 0 \text{ and } y < 0, \\ \text{undefined} & \text{if } x = 0 \text{ and } y = 0. \end{cases} \quad (4.1)$$

When averaging a set of angles we must not use the arithmetic mean. Instead, we must refer to the circular mean. Given a set of angles $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)^T$, we may compute their circular mean as:

$$\langle \boldsymbol{\theta} \rangle = \text{atan2}\left(\frac{1}{n} \sum_{j=1}^n \sin(\theta_j), \frac{1}{n} \sum_{j=1}^n \cos(\theta_j)\right), \quad (4.2)$$

where the atan2 function is defined in Equation (4.1).

The definition of the circular mean given in equation Equation (4.2) works by converting the angles into Cartesian co-ordinates — representing the directions as points on the unit circle. The centre of mass of the Cartesian co-ordinates is then computed, and the resulting position is converted back to a direction, resulting in our mean angle.

5

Model fitting

The task of a researcher working in the field of collective behaviour has to be in assessing the validity of proposed theoretical models. Models must be compared to real data. In this chapter we shall do exactly that: we shall generalise a popular model from the literature before proceeding to fit it to empirical data. Fitting is made in a Bayesian framework, which allows the capture of uncertainty in parameter estimation.

Though some work has been made at fitting theoretical models to field data, these attempts have typically used rudimentary methods to fit to some epiphenomena of the data, such as neighbour-density plots. Additionally, previous efforts have not made a systematic attempt at capturing uncertainty in parameter fitting — capturing this uncertainty is especially important since typical datasets often have high levels of noise.

5.1 SURF SCOTER DATASET

We were fortunate enough to receive a high quality dataset from a researcher in Canada (Lukeman *et al.*, 2010). The dataset consists of 25 sequences which describe the movements of surf scoters (a large sea duck) interacting on the surface of a lake. We are interested in this dataset because it provides tracking of individuals *between* frames. Tracking individuals between frames is necessary to reproduce the trajectories of individuals.

5.1.1 Visualisation

The dataset was captured by fixing a camera at a vantage point above a lake and, in the presence of surf scoters, taking photographs every 3 seconds. From the images, Lukeman *et al.* (2010) extracted the positions (recorded in pixel units) of agents in every frame. Every bird is assigned an ID number which allows the tracking of individuals between frames. The data extraction process is visualised in [Figure 2.4](#).

We received the dataset as 25 arrays detailing the positions of every agent in every frame. Manually inspecting an array, however, does not give us much knowledge about the motion of the birds in that sequence. To counter this we produce animations of every sequence. (Digital) In [Figure 5.1](#) we see an animation of the data contained in sequence 5. The position of every agent in every frame is shown, along with its identification number.

Of course, animations aren't always an appropriate way to visualise sequences. Animations are particularly unhelpful when trying to showcase data in a static medium, such as a printed thesis, or a research paper. As such, we construct trajectory plots which attempt to show the same information as an animation does, but in a single image. Example trajectory plots are given in [Figure 5.2](#). For completeness, plots of all the sequences can be found in [Appendix A](#). These plots give us a general idea of the motion of the group in each sequence. We may also use these plots to highlight particular trajectories of interest, as in [Figure 5.2\(b\)](#).

5.1.2 Missingness

At any given frame in a sequence there are individuals captured in the camera's field of vision, however, there are also individuals in the group which are not captured in the camera's view. We know that these uncaptured individuals exist as they may enter the field of vision at a later frame in the sequence, or we may have observed these individuals before they move out of the field of vision. As such, we know that there are individuals which may be influencing the group, which we don't observe in every frame. When an agent is missing from a frame, we say that this data is missing.

We categorise an agent's missingness into two distinct categories: data missing in the *beginning* of a sequence and data missing in the *end* of a sequence. We say that an agent is missing in the beginning of a sequence if it is not observed in frame 0, but is observed in some later frame number. Similarly, we say that an agent is missing in the end of a sequence if it is not observed in the final frame of the sequence, but was observed in some earlier frame. It is possible for an agent to have missingness both at the beginning *and* the end of

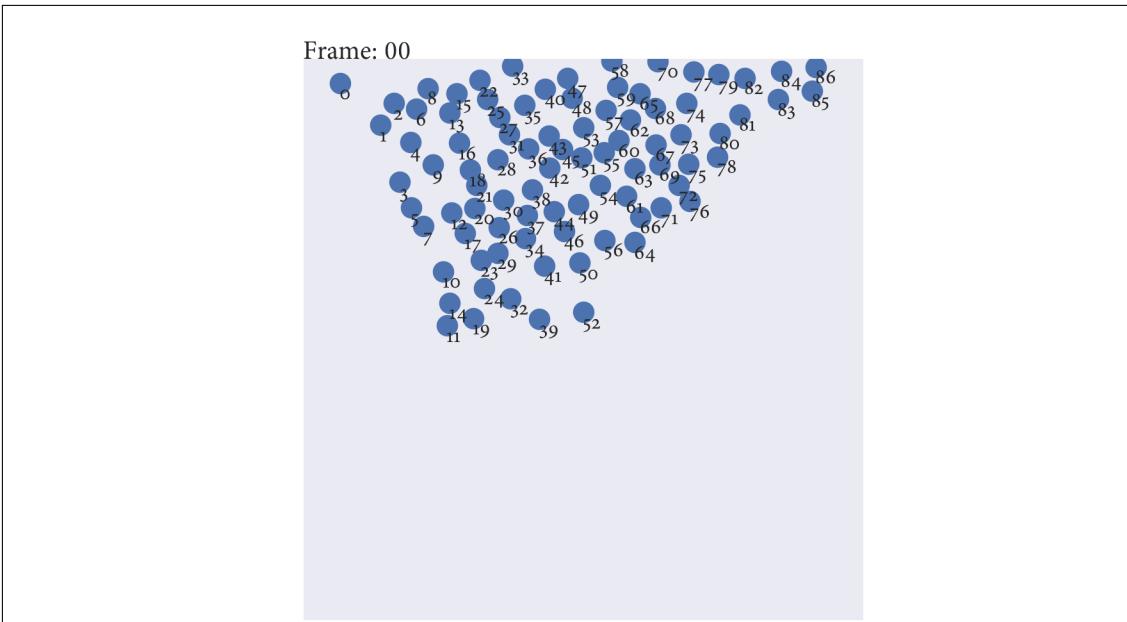


Figure 5.1: Animation of the data contained in sequence 5. Each agent is annotated by its identifying number.

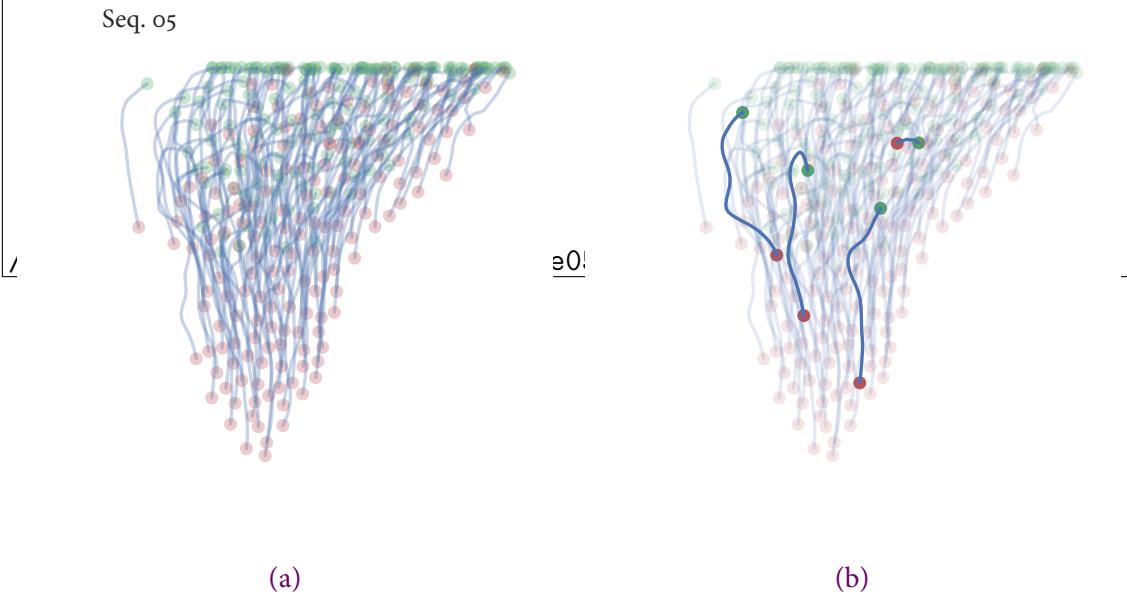


Figure 5.2: Inspecting the trajectories of the agents in sequence 5. The first observed positions of agents are represented by green markers. The last observed positions of agents are shown by red markers. (a): The observed trajectories of all agents. (b): Highlighting the observed trajectories of agents with ID 1, 20, 50, 63, 120 and 155.

Sequence	Agents	Frames	Data missing (%)		
			Beginning	End	Total
1	199	24	8	7	15
2	193	34	8	10	18
3	186	34	9	15	24
4	82	46	20	5	25
5	171	31	25	3	28
6	32	45	20	8	28
7	152	30	28	0	28
8	260	29	14	18	32
9	121	26	23	9	32
10	162	34	30	3	33
11	171	35	22	13	35
12	86	41	9	28	37
13	178	35	28	9	37
14	167	39	12	29	41
15	185	29	5	36	41
16	117	29	1	42	43
17	211	21	13	35	48
18	78	30	7	42	49
19	115	44	3	47	50
20	209	45	34	17	51
21	121	27	7	45	52
22	165	33	8	44	52
23	280	44	51	3	54
24	292	45	6	52	58
25	132	31	1	61	62

Table 5.1: Summary of the sequences in the dataset. Detailing the number of agents and the number of frames in every sequence, as well as the amount of missing data.

a sequence.

To get a general feel for the dataset and the information contained in each sequence, we summarise the data in [Table 5.1](#). The sequences are labelled from sequence 1 to sequence 25, in order of increasing missingness. For each sequence we indicate the number of agents observed, the number of frames, and the percentage of data missing.

Though [Table 5.1](#) gives a general idea of how much data is missing in each sequence, it doesn't tell us how the missingness is distributed throughout the flock. For example, we still don't know whether the missingness is distributed uniformly throughout the flock or whether the majority of the missingness belongs to a minority of the flock. To help understand

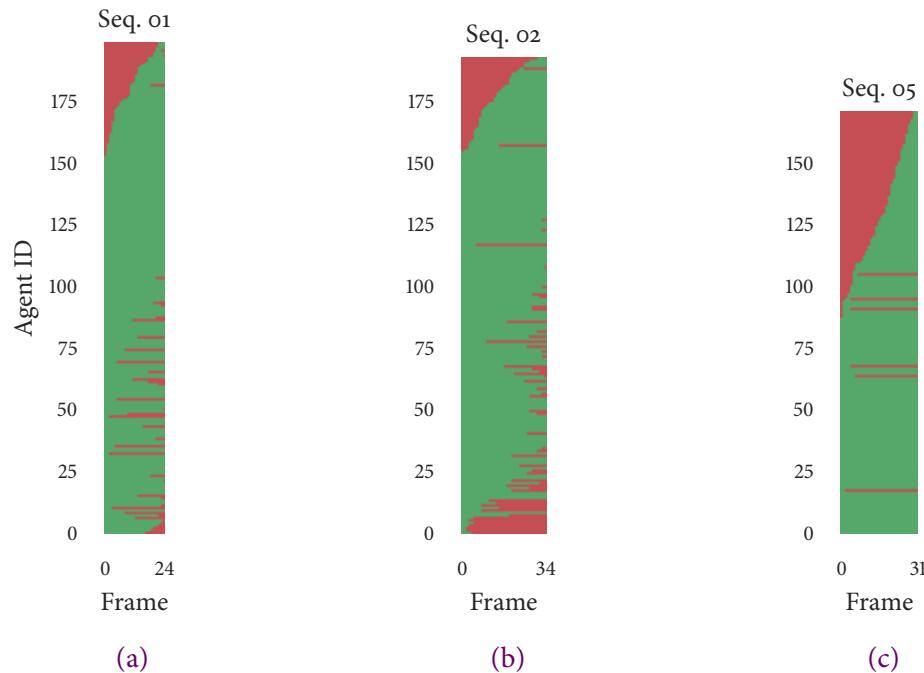


Figure 5.3: Visualisation of missingness in (a) Sequence 1 (b) Sequence 2 (c) Sequence 5. A green grid point tells us that agent i was observed at frame t , and a red point tells us that agent i was missing at time t .

the distribution of missingness we created plots to visualise at which times every individual is and isn't observed. A sample of these constructed plots are shown in Figure 5.3, and as before, plots of all the sequences are included in Appendix A. A green grid point in the plot informs us that the bird with ID number i was observed at frame t . Similarly, a red point shows us that bird i was missing at frame t .

5.2 MODEL DEVELOPMENT

Now that we have a feel for our dataset, we are ready to consider theoretical models which we can try fit to the data. There are many models which exist in the literature. In this section we shall describe one of the most popular models in the literature. Once described we shall consider how we can generalise this model to make it more flexible, and more biologically realistic. In this position, we are ready to fit this model to data.

5.2.1 Vicsek Model

Vicsek *et al.* (1995) devised a model which has now become a cornerstone of agent-based modelling. In their paper, Vicsek *et al.* (1995) showed that a simple model in which agents interact with neighbours only through alignment can produce movement reminiscent of real flocking behaviours. Furthermore, they showed the output of this model undergoes phase transitions as the amount of noise in the system is varied.

The Vicsek model simulates the movements of N agents over T timesteps. In the model agents seek to align their direction with that of their neighbours. Agents identify their neighbours as individuals which lie less than some fixed distance away. More formally, agent i 's neighbours at time t are given by the set $n_{i,t} = \{ j \mid d_{ij,t} < r, j = 1, \dots, N \}$, where r is the interaction radius and $d_{ij,t}$ is the Euclidean-distance between agents i and j at time t .

Initially, agents are assigned a random position, $\mathbf{x}_{i,t=0}$, within a square cell of length L and a random direction of motion $\theta_{i,t=0}$. Agents move with fixed speed v and update their positions as

$$\mathbf{x}_{i,t+1} = \mathbf{x}_{i,t} + \mathbf{v}_{i,t}, \quad (5.1)$$

where $\mathbf{v}_{i,t}$ is constructed to have direction $\theta_{i,t+1}$ and magnitude $|\mathbf{v}_{i,t}| = v$.

An agents directional update is computed as the average direction of its neighbours plus some noise term. To compute the average direction, we refer to circular mean defined in [Equation \(4.2\)](#). With this definition, we can compute the average direction of agent i 's neighbours at time t as

$$\langle \theta \rangle_{i,t} = \text{atan2}\left(\frac{1}{|n_{i,t}|} \sum_{j=1}^N \mathbb{1}_{n_{i,t}}(j) \sin \theta_{j,t}, \frac{1}{|n_{i,t}|} \sum_{j=1}^N \mathbb{1}_{n_{i,t}}(j) \cos \theta_{j,t}\right), \quad (5.2)$$

where the indicator function, $\mathbb{1}$, is defined

$$\mathbb{1}_{n_{i,t}}(j) = \begin{cases} 1 & \text{if } j \in n_{i,t} \\ 0 & \text{if } j \notin n_{i,t}. \end{cases}$$

The noise in Vicsek's model was drawn from a uniform $U(-\eta/2, \eta/2)$ distribution and summed with the average direction. Putting this all together allows us to realise agent i 's directional update as

$$\theta_{i,t+1} \mid \langle \theta \rangle_{i,t}, \eta \sim U(\langle \theta \rangle_{i,t} - \eta/2, \langle \theta \rangle_{i,t} + \eta/2). \quad (5.3)$$

5.2.2 Hierarchical Vicsek

Here we shall seek to alter the Vicsek model so that it becomes more biologically realistic. One of the ways we do this is by introducing hierarchy into the model, which allows variation within groups.

In all of the agent-based models encountered in the literature, there is no inter-agent variation — all the agents in a group are identical. In the context of the Vicsek model, all the agents have the same interaction radius r , and are all subjected to noise generated from a uniform $U(-\eta/2, \eta/2)$ distribution. In effect, the individuals in the groups respond identically to the same stimulus.

In reality, there will be biological and behavioural variation within groups of individuals. Some individuals may be more prone to noise, others less. Similarly, some agents may pay close attention to the movements of their neighbours, whereas other individuals may pay less attention. These differing responses may be due to physiological differences between the individuals, but they also may be linked to social hierarchy, or correlated with the age and experience of an individual. With these considerations, we conclude that we should allow variation within groups. In Vicsek, we do so by assigning every agent its own interaction radius r_i , and allowing individuals to be more or less prone to noise, assigning each agent its own parameter η_i .

In considering the Vicsek model, we decide that the hard cut-off imposed by the interaction radius r_i is biologically unrealistic. In this paradigm, interactions are binary: agents interact fully, or not at all. For example, consider that the distance between agent i and j at time t is $r_i + |\epsilon|$, for $|\epsilon| \ll 1$. In this case agent j makes no contribution to agent i 's direction. However, if $d_{ij,t} = r_i - |\epsilon|$ for $|\epsilon| \ll 1$, then agent j makes a full contribution to agent i 's direction. This case is visualised in [Figure 5.4\(a\)](#). Similarly, because of the interaction, agents distanced $|\epsilon|$ and $r_i - |\epsilon|$ away from agent i , both influence agent i equally.

Our belief is that the averaging process should be weighted by distance. That way, a neighbour very close to agent i will have more influence than a neighbour much further away. We decide that a bivariate normal distribution provides an appropriate weighting function. We calculate $\omega_{ij,t}$, the weighting which agent i will give to neighbour j 's direction at time t , defined by the bivariate normal distribution with mean $(x_{i,t}, y_{i,t})^T$ and covariance

$$\Sigma_i = \begin{bmatrix} \sigma_{X_i}^2 & 0 \\ 0 & \sigma_{X_i}^2 \end{bmatrix}.$$

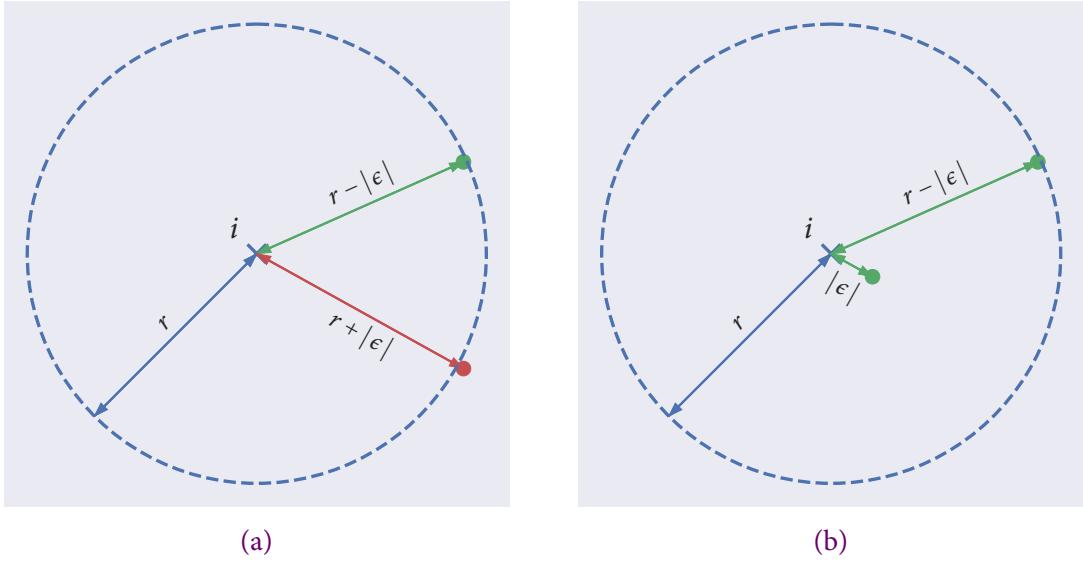


Figure 5.4: Agent i , its interaction radius and neighbours (green). (a): The hard cut-off at r . (b): Without weighting both neighbours influence i equally.

Now, we can adapt Equation (5.2) to define the weighted average:

$$\langle \theta \rangle_{i,t} = \text{atan2}\left(\frac{1}{N} \sum_{j=1}^N \omega_{ij,t} \sin \theta_{j,t}, \frac{1}{N} \sum_{j=1}^N \omega_{ij,t} \cos \theta_{j,t}\right). \quad (5.4)$$

The final refinement which we make is to alter the distribution from which the noise is generated. In Vicsek noise is generated from a uniform distribution, so agents are just as likely to be perturbed by large amounts of noise as small amounts. We consider that the likelihood of experiencing a particular noise term should be weighted by the magnitude of the noise, and that small noise terms should be encountered more frequently than large noise terms. To account for this we generate noise from a normal $N(0, \sigma_{Y_i})$. In our modified model, agents update their direction according to

$$\theta_{i,t+1} \mid \langle \theta \rangle_{i,t}, \sigma_{Y_i} \sim N(\langle \theta \rangle_{i,t}, \sigma_{Y_i}). \quad (5.5)$$

In our adapted version of Vicsek's model, agent i 's behaviour is controlled by the parameters σ_{X_i} and σ_{Y_i} . The weighting which agent i gives to its neighbours depends on the σ_{X_i} ; a large value of σ_{X_i} corresponds to an agent which weights its neighbours strongly, and a small value of σ_{X_i} corresponds to an agent which pays little attention to its neighbours. Similarly, the parameter σ_{Y_i} controls how much noise an agent experiences. A large value of σ_{Y_i} corresponds to an agent which experiences large amounts of noise. In a similar manner,

a small value of σ_{Y_i} tells us that agent i is only perturbed by small amounts.

With the parameters of the model decided, we are left to choose appropriate hyperparameters. In both cases our parameters are standard deviations, and so are strictly positive. This should be reflected in the distributions over these parameters. We decide to use gamma distributions in both cases as it has support $(0, \infty)$.

Typically, the gamma distribution is parameterised as $\text{Ga}(g, h)$. Here, however, we opt to parameterise the gamma distribution by its mean and variance. Given that $X|g, h \sim \text{Ga}(g, h)$, it is known that $E[X] = g/h$ and $\text{Var}(X) = g/h^2$. Letting $m = E[X]$ and $v = \text{Var}(X)$, we see that $g = m^2/v$ and $h = m/v$. As such, we see that an alternative, yet equivalent parameterisation is $X|m, v \sim \text{Ga}(m^2/v, m/v)$. We prefer this parameterisation as constructing prior distributions over m and v is more intuitive than working with g and h . Putting all this together, we decide that the interaction parameter should be distributed as

$$\sigma_X | m_X, v_X \sim \text{Ga}(m_X^2/v_X, m_X/v_X). \quad (5.6)$$

Similarly, for the noise parameter we stipulate the distribution

$$\sigma_Y | m_Y, v_Y \sim \text{Ga}(m_Y^2/v_Y, m_Y/v_Y). \quad (5.7)$$

5.3 INFERRING PARAMETERS

With the model setup, we are now ready to fit this model to the dataset described in Section 5.1 and perform parameter inference. To use the inference methods introduced in Chapter 3, we need to able to express the likelihood of the parameters given some data.

Let us begin by considering the likelihood of observing a single agent update its direction, in the presence of neighbours, from $\theta_{i,t}$ to $\theta_{i,t+1}$. Using Equation (5.5), σ_{X_i} and $\{\theta_{j,t}, \mathbf{x}_{j,t} \mid j = 1, \dots, N\}$ to compute the mean direction $\langle \theta \rangle_{i,t}$, we may realise the likelihood:

$$L(\sigma_{X_i}, \sigma_{Y_i} \mid \theta_{i,t+1}, \{\theta_{j,t}, \mathbf{x}_{j,t} \mid j = 1, \dots, N\}) = \left(\frac{1}{2\pi\sigma_{Y_i}^2} \right)^{1/2} \exp \left\{ -\frac{1}{2\sigma_{Y_i}^2} \left(\theta_{i,t+1} - \langle \theta \rangle_{i,t} \right)^2 \right\}. \quad (5.8)$$

Let us now consider that we wish to compute the likelihood of observing agent i 's directions $\theta_{i,t}$, $\theta_{i,t+1}$ and $\theta_{i,t+2}$, again in the presence of neighbours. Using that successive noise

terms are independent, we can express the likelihood as a product

$$\begin{aligned} & L(\sigma_{X_i}, \sigma_{Y_i} | \theta_{i,t+2}, \{ \theta_{j,t'}, \mathbf{x}_{j,t'} | j = 1, \dots, N, t' = t, t+1 \}) \\ &= L(\sigma_{X_i}, \sigma_{Y_i} | \theta_{i,t+2}, \{ \theta_{j,t+1}, \mathbf{x}_{j,t+1} | j = 1, \dots, N \}) \cdot L(\sigma_{X_i}, \sigma_{Y_i} | \theta_{i,t+1}, \{ \theta_{j,t}, \mathbf{x}_{j,t} | j = 1, \dots, N \}). \end{aligned}$$

We may notice that the two terms in the product are of the same form as in [Equation \(5.8\)](#). Using this observation, we rewrite the likelihood more compactly

$$\begin{aligned} & L(\sigma_{X_i}, \sigma_{Y_i} | \theta_{i,t+2}, \{ \theta_{j,t'}, \mathbf{x}_{j,t'} | j = 1, \dots, N, t' = t, t+1 \}) \\ &= \prod_{t'=t}^{t+1} L(\sigma_{X_i}, \sigma_{Y_i} | \theta_{i,t+1}, \{ \theta_{j,t}, \mathbf{x}_{j,t} | j = 1, \dots, N \}) \\ &= \prod_{t'=t}^{t+1} \left(\frac{1}{2\pi\sigma_{Y_i}^2} \right)^{1/2} \exp \left\{ -\frac{1}{2\sigma_{Y_i}^2} \left(\theta_{i,t+1} - \langle \theta \rangle_{i,t} \right)^2 \right\} \end{aligned}$$

Using the product notation makes it much easier to express and evaluate likelihoods of data over many timesteps. For example, suppose that we now observe the movements of N agents from timestep $t = 1$ to timestep $t = T$. The likelihood of observing agent i 's trajectory can be neatly expressed

$$\begin{aligned} & L(\sigma_{X_i}, \sigma_{Y_i} | \theta_{i,T}, \{ \theta_{j,t}, \mathbf{x}_{j,t} | j = 1, \dots, N, t = 1, \dots, T-1 \}) \\ &= \prod_{t=1}^{T-1} L(\sigma_{X_i}, \sigma_{Y_i} | \theta_{i,t+1}, \{ \theta_{j,t}, \mathbf{x}_{j,t} | j = 1, \dots, N \}) \\ &= \prod_{t=1}^{T-1} \left(\frac{1}{2\pi\sigma_{Y_i}^2} \right)^{1/2} \exp \left\{ -\frac{1}{2\sigma_{Y_i}^2} \left(\theta_{i,t+1} - \langle \theta \rangle_{i,t} \right)^2 \right\}. \end{aligned}$$

Realising the likelihood of observing the movements of multiple agents is a similar process to expressing the likelihood of a single agent over multiple steps. First, we must recognise that the noise terms experienced by agents i and $i + 1$ are independent and that this allows the likelihood to be factorised as a product over the agents. With this acknowledgement, we determine that the likelihood of observing the movements of N

agents over T timesteps, given the parameters σ_X and σ_Y , to be

$$\begin{aligned}
 L(\sigma_X, \sigma_Y | \{\theta_{i,T} | i = 1, \dots, N\}, \{\theta_{j,t}, \mathbf{x}_{j,t} | j = 1, \dots, N, t = 1, \dots, T-1\}) \\
 &= \prod_{i=1}^N \prod_{t=1}^{T-1} L(\sigma_{X_i}, \sigma_{Y_i} | \theta_{i,t+1}, \{\theta_{j,t}, \mathbf{x}_{j,t} | j = 1, \dots, N\}) \\
 &= \prod_{i=1}^N \prod_{t=1}^{T-1} \left(\frac{1}{2\pi\sigma_{Y_i}^2} \right)^{1/2} \exp \left\{ -\frac{1}{2\sigma_{Y_i}^2} \left(\theta_{i,t+1} - \langle \theta \rangle_{i,t} \right)^2 \right\}.
 \end{aligned} \tag{5.9}$$

So, we can now express the likelihood of the data, given the parameters σ_X and σ_Y . Recall, however, that there is still the matter of the hyperparameters to be considered. Revisiting Equations (5.6) and (5.7) allows us to formulate the likelihood of the hyperparameters, given the parameters, as a product of successive evaluations of the Gamma distribution:

$$L(m_X, v_X, m_Y, v_Y | \sigma_X, \sigma_Y) = \prod_{i=1}^N \frac{h_X^{g_X}}{\Gamma(g_X)} \sigma_{X_i}^{g_X-1} \exp\{-h_X \sigma_{X_i}\} \prod_{i=1}^N \frac{h_Y^{g_Y}}{\Gamma(g_Y)} \sigma_{Y_i}^{g_Y-1} \exp\{-h_Y \sigma_{Y_i}\}, \tag{5.10}$$

where $g_X = m_X^2/v_X$, $h_X = m_X/v_X$, and similarly for g_Y and h_Y .

We are now in a position to express the likelihood of observing data describing the movements of N agents over T timesteps, given parameters σ_X , σ_Y and hyperparameters m_X , v_X , m_Y and v_Y , combining Equations (5.9) and (5.10) to see

$$\begin{aligned}
 L(m_X, v_X, m_Y, v_Y, \sigma_X, \sigma_Y | \{\theta_{i,T} | i = 1, \dots, N\}, \{\theta_{j,t}, \mathbf{x}_{j,t} | j = 1, \dots, N, t = 1, \dots, T-1\}) \\
 &= L(m_X, v_X, m_Y, v_Y | \sigma_X, \sigma_Y) \\
 &\quad \cdot L(\sigma_X, \sigma_Y | \{\theta_{i,T} | i = 1, \dots, N\}, \{\theta_{j,t}, \mathbf{x}_{j,t} | j = 1, \dots, N, t = 1, \dots, T-1\}) \\
 &= \prod_{i=1}^N \frac{h_X^{g_X}}{\Gamma(g_X)} \sigma_{X_i}^{g_X-1} \exp\{-h_X \sigma_{X_i}\} \prod_{i=1}^N \frac{h_Y^{g_Y}}{\Gamma(g_Y)} \sigma_{Y_i}^{g_Y-1} \exp\{-h_Y \sigma_{Y_i}\} \\
 &\quad \cdot \prod_{i=1}^N \prod_{t=1}^{T-1} \left(\frac{1}{2\pi\sigma_{Y_i}^2} \right)^{1/2} \exp \left\{ -\frac{1}{2\sigma_{Y_i}^2} \left(\theta_{i,t+1} - \langle \theta \rangle_{i,t} \right)^2 \right\}.
 \end{aligned} \tag{5.11}$$

Now, to write down the full posterior distribution we need only combine the likelihood of Equation (5.11) with prior beliefs $\pi(m_X, v_X, m_Y, v_Y)$. Again, because we are considering the mean and variance of standard deviations, we chose to represent our prior beliefs with Gamma distributions. We consider that that our prior beliefs between mean and variance

are independent, and so express the hyperprior as

$$\pi(m_X, v_X, m_Y, v_Y) = \pi(m_X)\pi(v_X)\pi(m_Y)\pi(v_Y).$$

We then write down our prior beliefs on each of the hyperparameters separately

$$\begin{aligned} m_X | \alpha_X, \beta_X &\sim \text{Ga}(\alpha_X, \beta_X) \\ v_X | \lambda_X, \gamma_X &\sim \text{Ga}(\lambda_X, \gamma_X) \\ m_Y | \alpha_Y, \beta_Y &\sim \text{Ga}(\alpha_Y, \beta_Y) \\ v_Y | \lambda_Y, \gamma_Y &\sim \text{Ga}(\lambda_Y, \gamma_Y), \end{aligned}$$

where the parameters α , β , λ and γ are to be chosen later.

5.4 INFERRING MISSINGNESS

In [Section 5.1.2](#) we considered that there are missing data points in the sequences. In total, 28% of the dataset is missing from sequence 5. Of this missing data, 3% is missing at the end of the sequence and 25% is missing in the beginning of the sequence. In total there are a possible 5 301 datapoints in the sequence, of which 1 468 points are missing.

We handle the missingness by integrating over all the possible paths an agent could have made when unobserved. The handling of the missing data is more computationally intensive than inferring the parameters of the model as the dimensionality of the missing data is ($\approx 5\times$) bigger. As such, this part of the inference scheme is inferring a very large amount of data. *In the future I will detail the proposal mechanisms I used to propose missing directions at the beginning and end of the sequences (the cases are handled separately), and the resulting corrections to the acceptance ratio.*

5.5 OUTPUT

Inference shall be made only on sequence 5 (*in time we will extend our inference to work on all the sequences*). We chose this sequence as there is more perturbation in the trajectories of this sequence than some of the other sequences. We understand that the more the trajectories of the agents depart from a straight line, the more information that sequence contains.

The behaviour of every agent in the model is specified by two parameters. In total,

sequence 5 tracks the movements of 171 agents — this means there are 342 parameters to fit in this sequence. In addition to the parameters, there are a further 4 hyperparameters. Our scheme will therefore need to explore a parameter space with 346 dimensions. For now, we shall concentrate only on the parameters in the model, and not consider the missingness.

We run Metropolis-Hastings as described in [Algorithm 2](#), for 10^6 iterations, thinning the output by a factor of 100. We perform a normal random walk to explore the parameter space. To help the mixing of the chains over such a large parameter space, the space is divided into three blocks, containing the hyperparameters, σ_X and σ_Y parameters separately.

A run of this scheme took approximately 11 hours to complete. Evaluating the likelihood is computationally expensive as to explore the parameter space of σ_X we are required to recompute $\langle \theta \rangle_{i,t}$ for $i = 1, \dots, 175$, $t = 1, \dots, 31$ at every iteration — of which there were 10^6 . Of the proposed parameters 27% of the hyperparameters, 19% of the σ_X parameters, and 29% of σ_Y parameters were accepted.

A selection of the output from the scheme is visualised in [Figures 5.5](#) to [5.12](#). In these plots we see that the chains over the hyperparameters have converged. Of particular note is ν_X — we see from [Figure 5.6](#) that this inferred value is very large, suggesting that there is a lot of variation in the σ_X parameters and how the individuals in the group interact with their neighbours. It appears that some agents are interacting more strongly with neighbours than others. This appears an interesting result, contradicting the assumption of many models in the literature that all agents behave identically.

From [Figures 5.7](#) and [5.8](#) we see that our chains over the σ_Y parameters have almost converged — though our scheme would benefit from a slightly longer run with more thinning. Recall that σ_{Y_i} determines how much noise is experienced by agent i . To help interpret this parameter, we take the mean value over each of the first four σ_Y chains. In [Figure 5.9](#) we plot normal distributions with standard deviations set by the mean of these chains. We immediately get a sense that the noise terms which agents experience are very small. To further help us interpret the noise terms we plot polar histograms of draws from these distributions in [Figure 5.10](#).

Looking at the chains over the first four σ_X parameters and the corresponding posterior beliefs, given in [Figures 5.11](#) and [5.12](#) respectively, we see that our chains have not finished converging. To reach convergence our scheme needs running for longer with more thinning applied. (Alternatively we could attempt to tune the random walk better, after all the acceptance rate of the σ_X chains was 0.19, which is a little low.) Despite this, we can see that our chains are beginning to converge. *It is not obvious how to interpret the σ_X parameters — we have tried a few approaches. I shall omit them for now until we are completely satisfied*

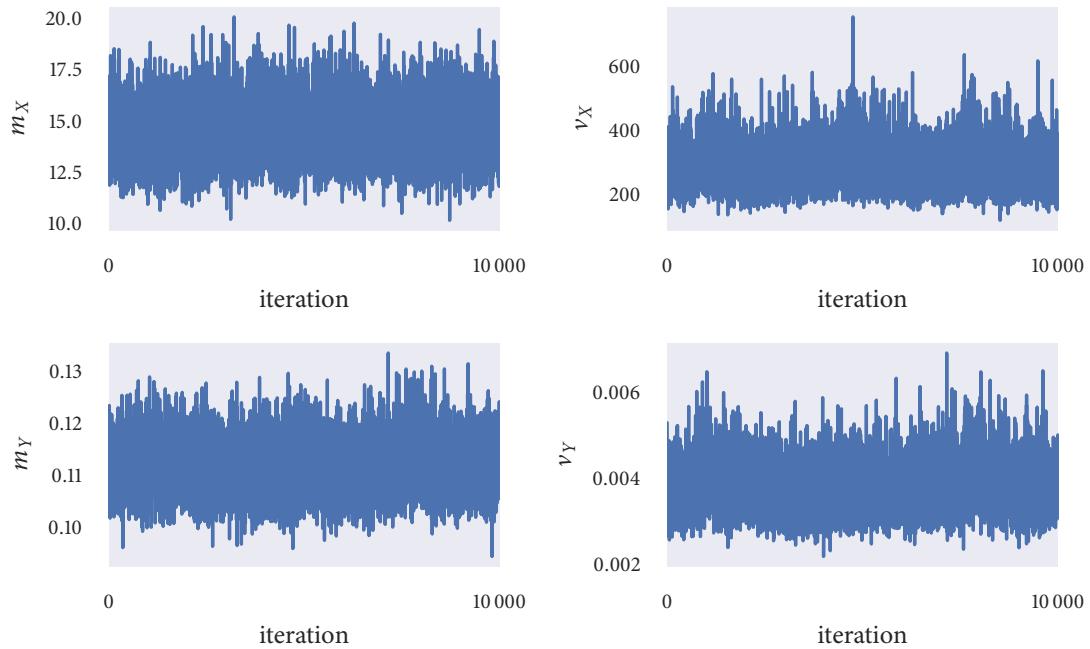


Figure 5.5: Trace plots over the hyperparameters.

with the measures we use.

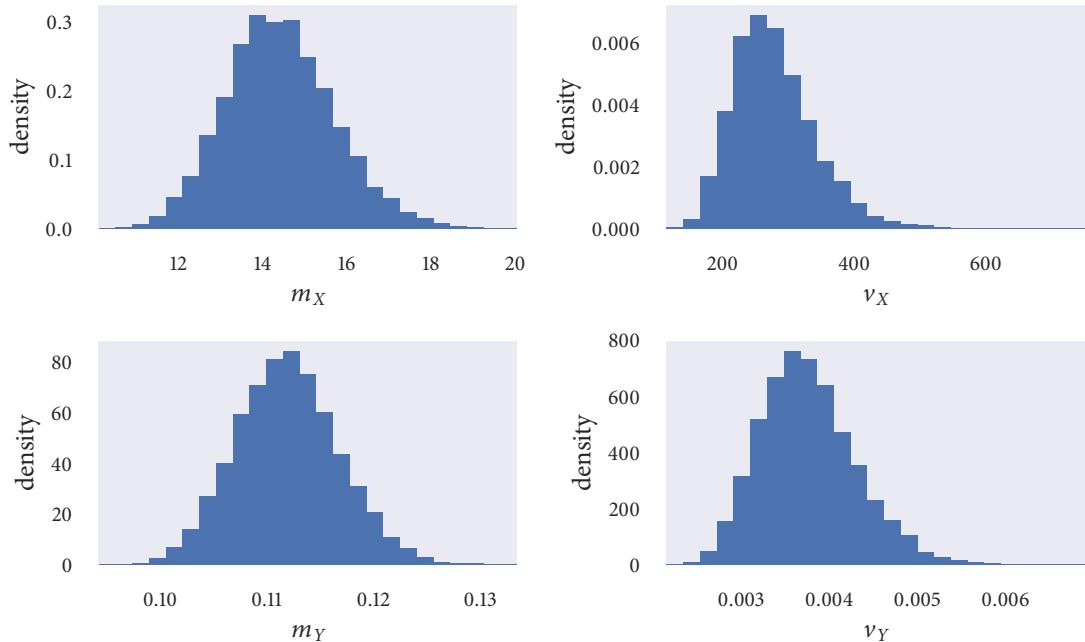


Figure 5.6: Histogram plots of our posterior beliefs about the hyperparameters m_X , v_X , m_Y and v_Y .

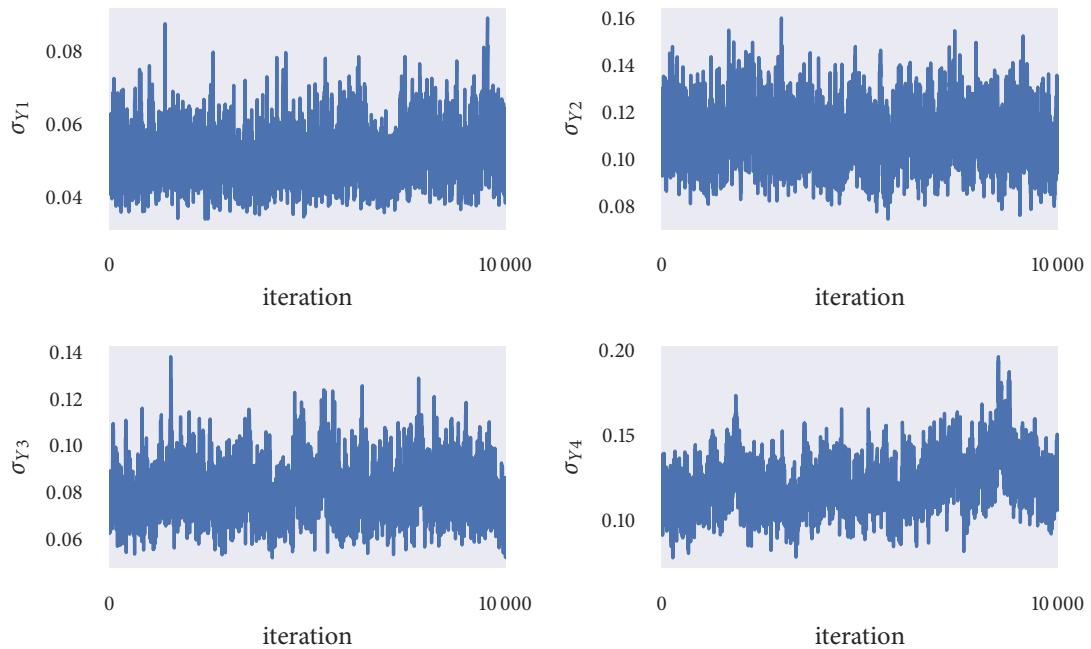


Figure 5.7: Trajectories of the chains over the first four σ_Y parameters.

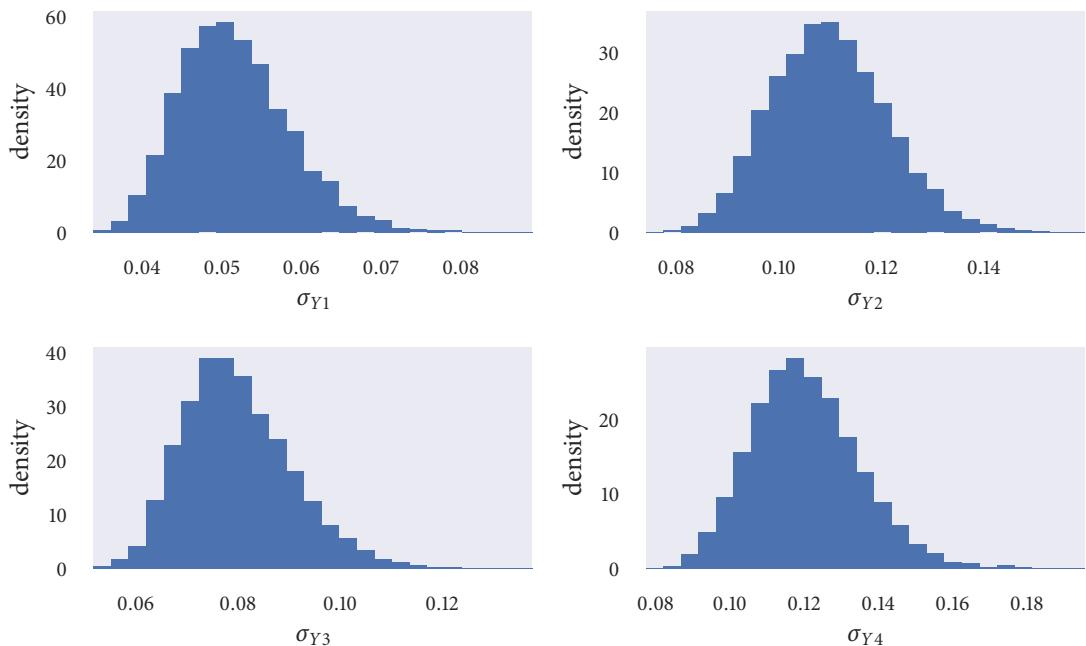


Figure 5.8: Posterior beliefs about the parameters σ_{Y_1} , σ_{Y_2} , σ_{Y_3} and σ_{Y_4} .

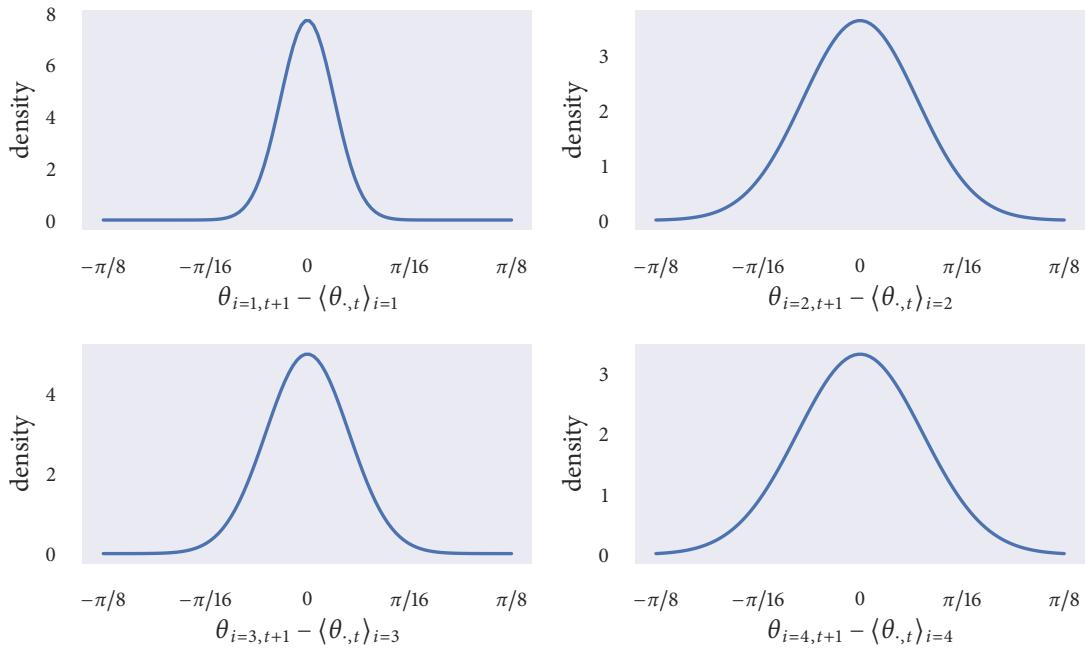


Figure 5.9: Taking the mean values of the chains over σ_{Y_1} , σ_{Y_2} , σ_{Y_3} and σ_{Y_4} , here we plot the noise experienced by agents 1, 2, 3 and 4.

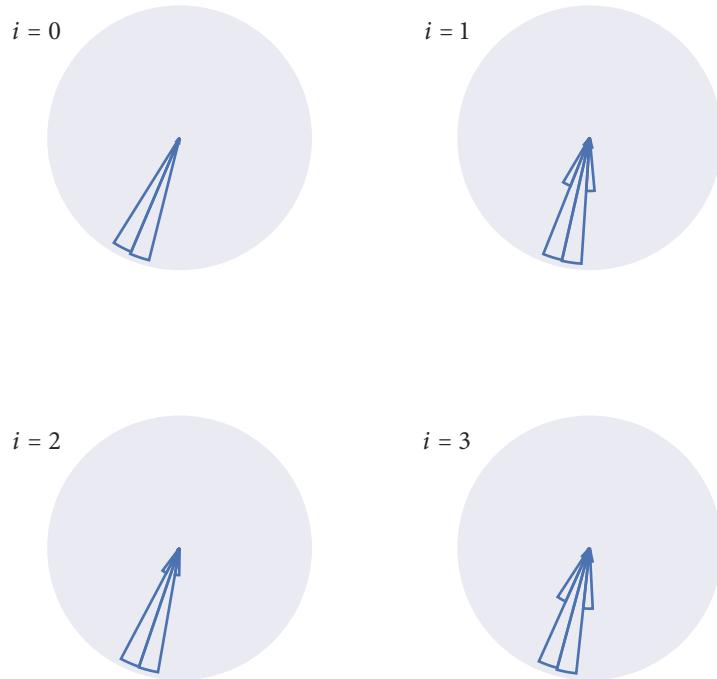


Figure 5.10: The noise experienced by agents 1, 2, 3 and 4, plotted on polar histograms, rotated by the average direction of each agent throughout the sequence.

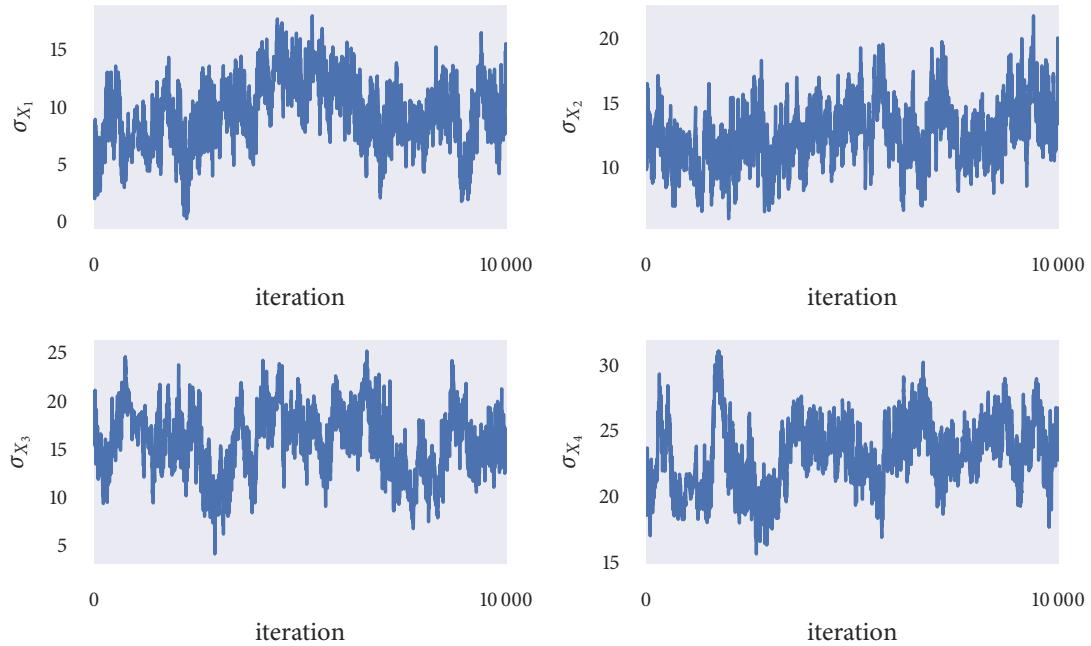


Figure 5.11: Chains over the parameters σ_{X_1} , σ_{X_2} , σ_{X_3} and σ_{X_4} .

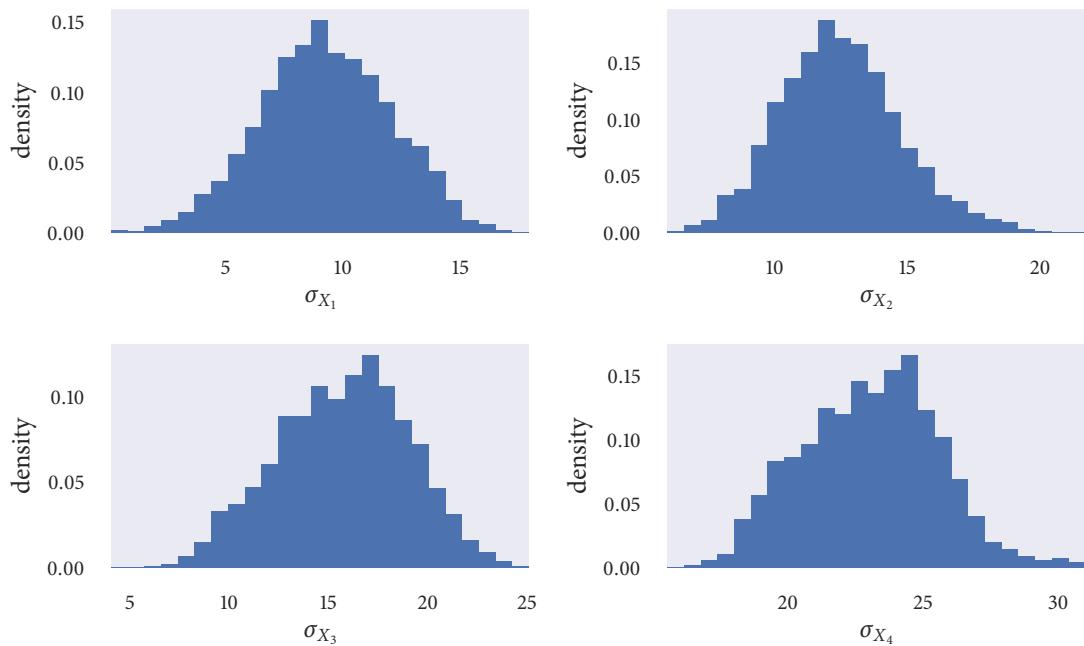


Figure 5.12: Samples from the posterior distribution over the first four σ_X parameters.

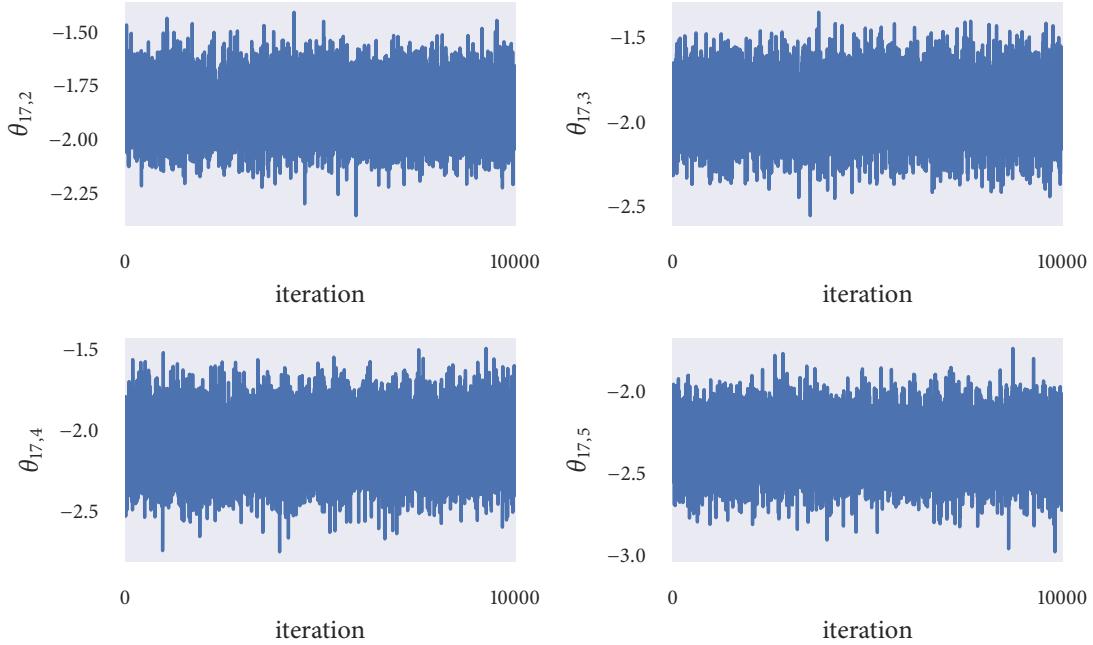


Figure 5.13: Chains over the missing directions of agent 17, missing from frame 2 until the end of the sequence.

Of the paths proposed at the end of the sequence 100% are accepted. However, for the paths proposed at the beginning of the scheme only 17% are accepted. The scheme took a little over 8 hours to run.

Because we are inferring such a large amount of data it is infeasible to assess all of the output visually. Here we shall inspect a small amount of the inferred data. Figures 5.13 to 5.15 visualise inferred data at the end of the sequence. In particular, these plots assess the inferred directions of bird 17 over frames 1, 2, 3 and 4. The chains shown in Figure 5.13 show that our scheme has here converged, and in Figures 5.14 and 5.15 we inspect our posterior beliefs about the missing directions.

In Figures 5.16 to 5.18 we inspect output relating to data missing at the beginning of the sequence. In fact, we inspect inferred directions for agent 97 over frames 0, 1, 2 and 3. *Note, to be consistent with model setup we should label the frames from 1 and not 0. As such the below plots need altering.* What we observe in these plots is unusual — we see that the scheme looks to be working fine, before it proceeds to stick for large numbers of iterations, we are in the process of examining why this happened and hope to resolve the problem shortly.

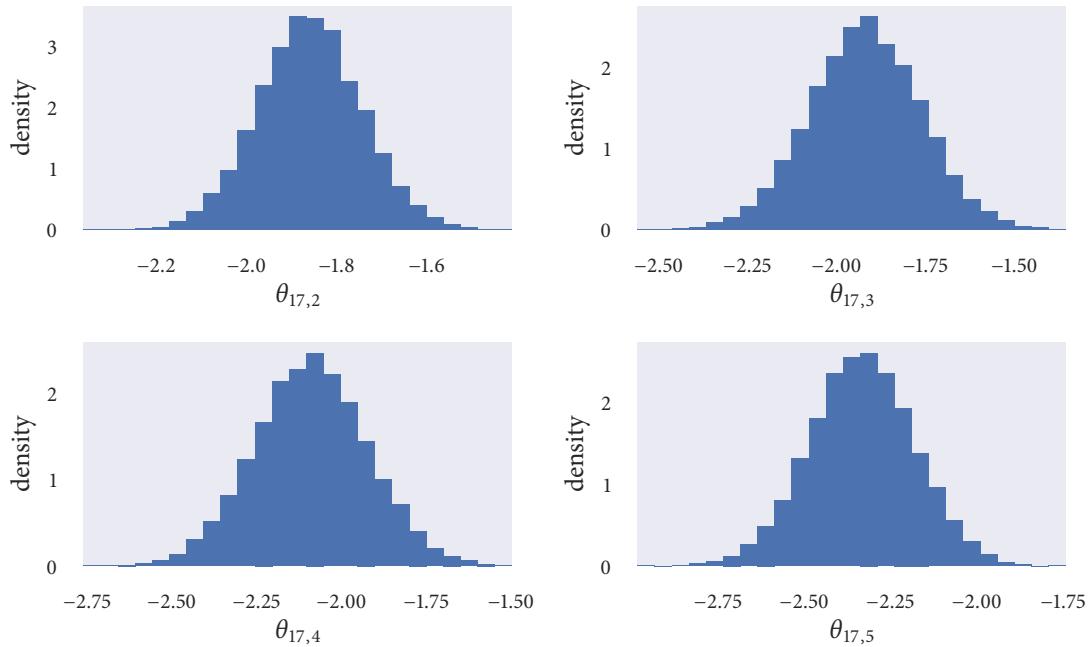


Figure 5.14: Our posterior beliefs about the missing directions of agent 17, over frames 2, 3, 4 and 5.

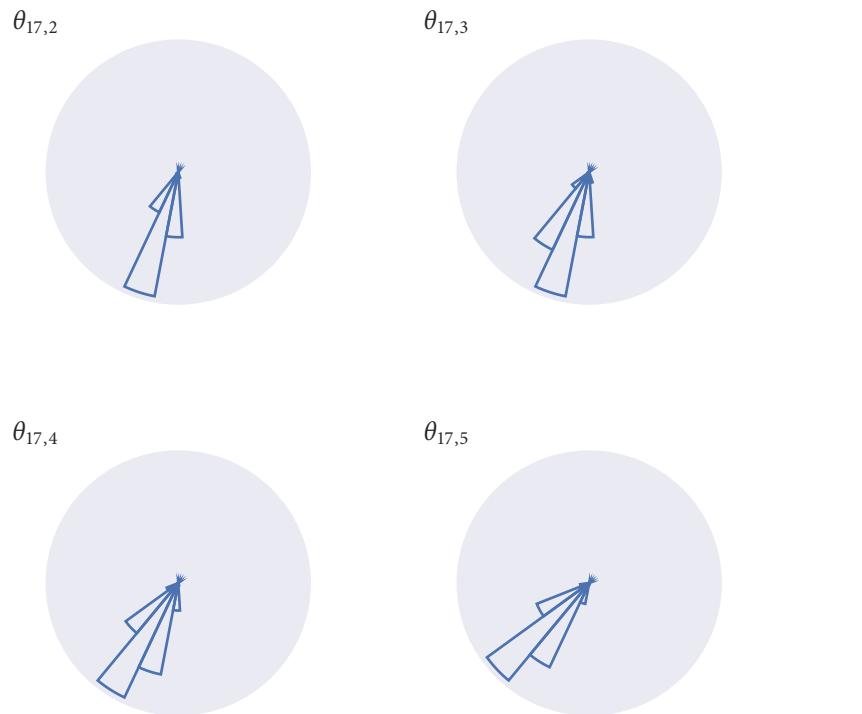


Figure 5.15: Polar histograms showing our posterior beliefs about the missing directions of agent 17, over frames 2, 3, 4 and 5.

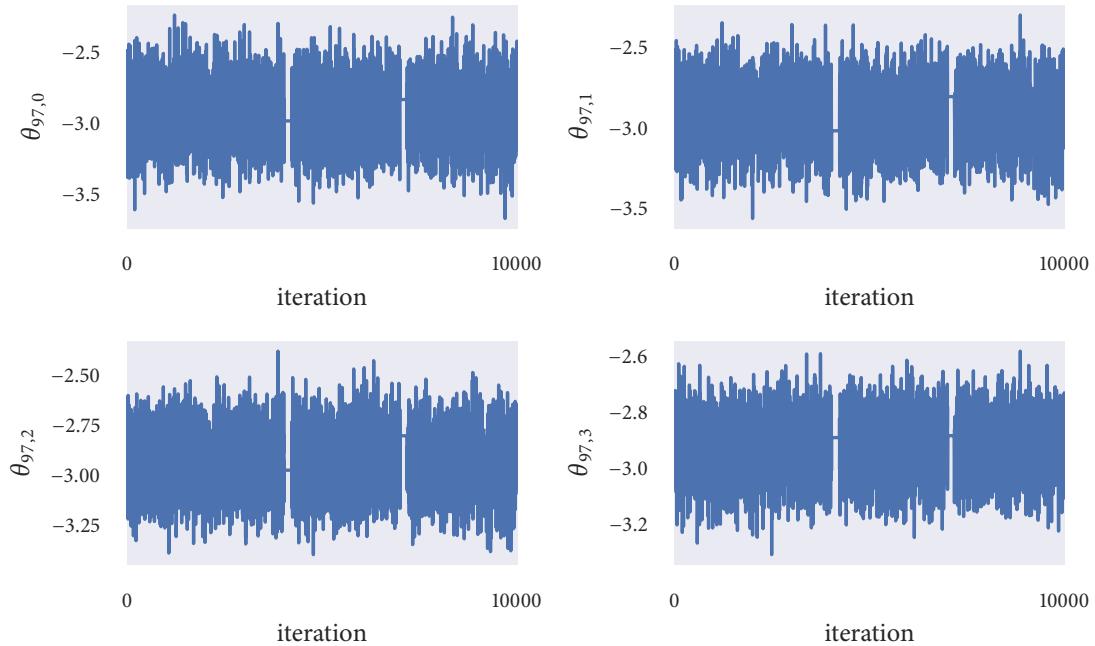


Figure 5.16: Inferring directions missing at the beginning of a sequence. In particular we show trace plots over agent 97's missingness.

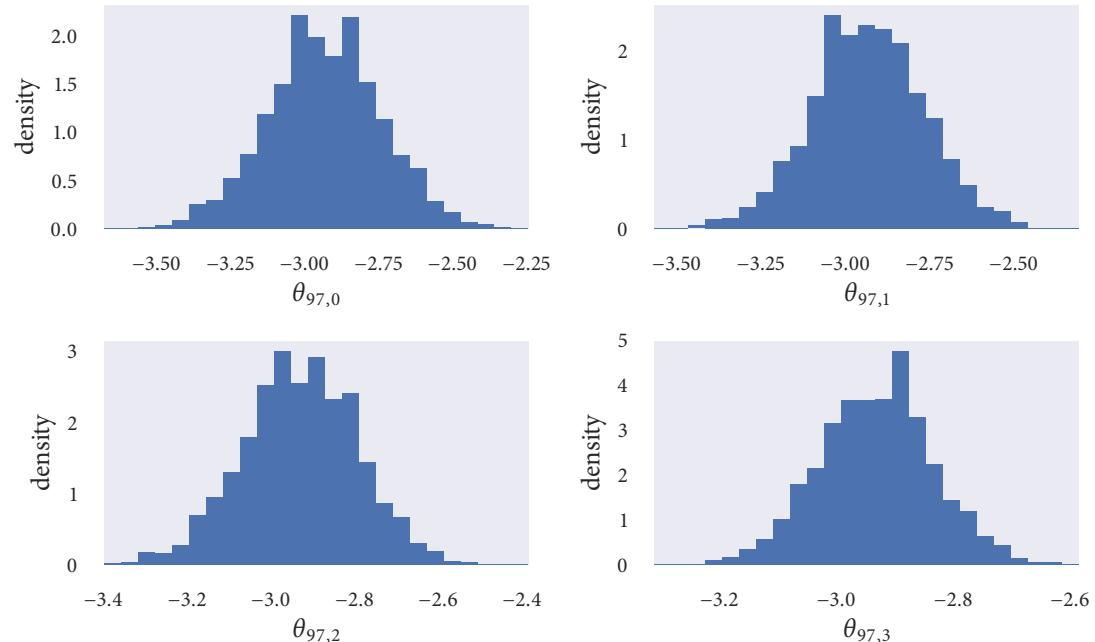


Figure 5.17: Histogram plots of our posterior beliefs about the missing trajectories of agent 97.

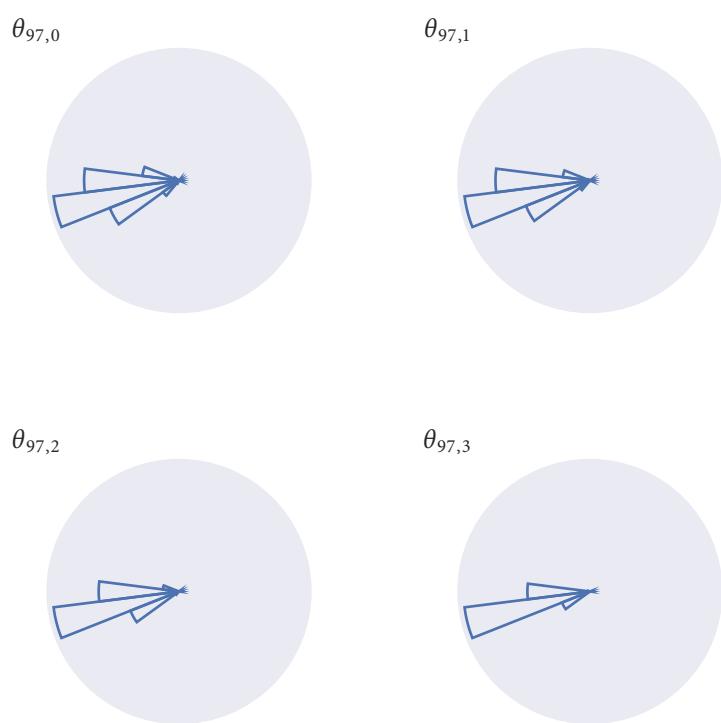


Figure 5.18: Polar histogram showing our posterior beliefs about the missing directions of agent 97 at the beginning of the sequence.

A

Data plots

Here we include, in full, summary plots of each sequence. In particular, we observe how missing data is distributed in each sequence. We do this by creating a grid where every square represents a data point describing the position of a single agent at a single frame. Each gridpoint is coloured green or red — green if the corresponding datapoint is observed and red if the datapoint is missing.

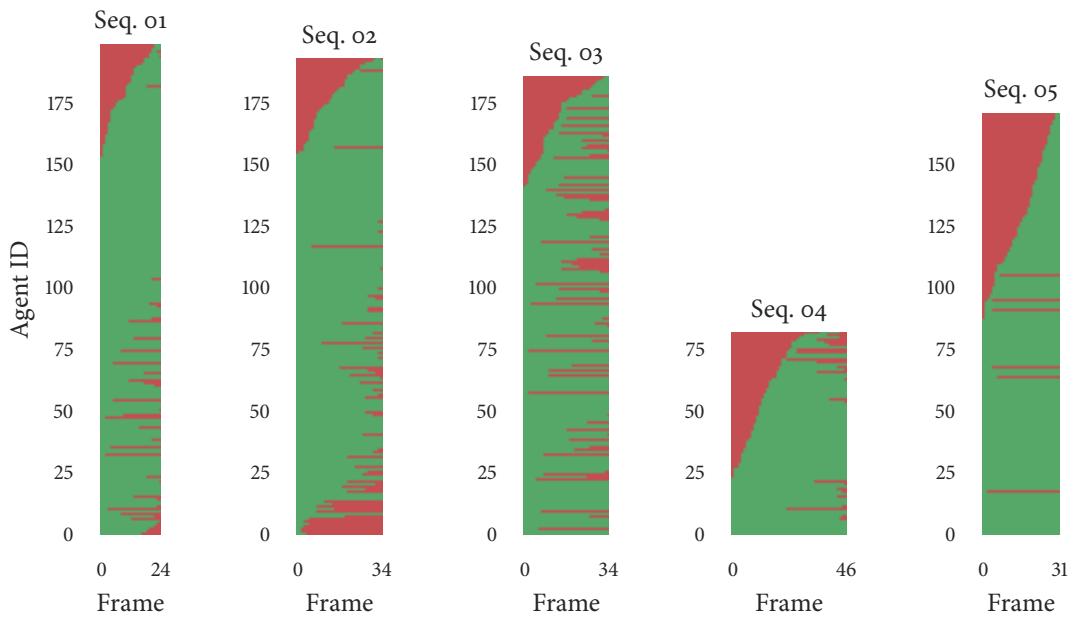


Figure A.1: Distribution of missingness over agents and time, in sequences 1–5 (left to right).

50 APPENDIX A

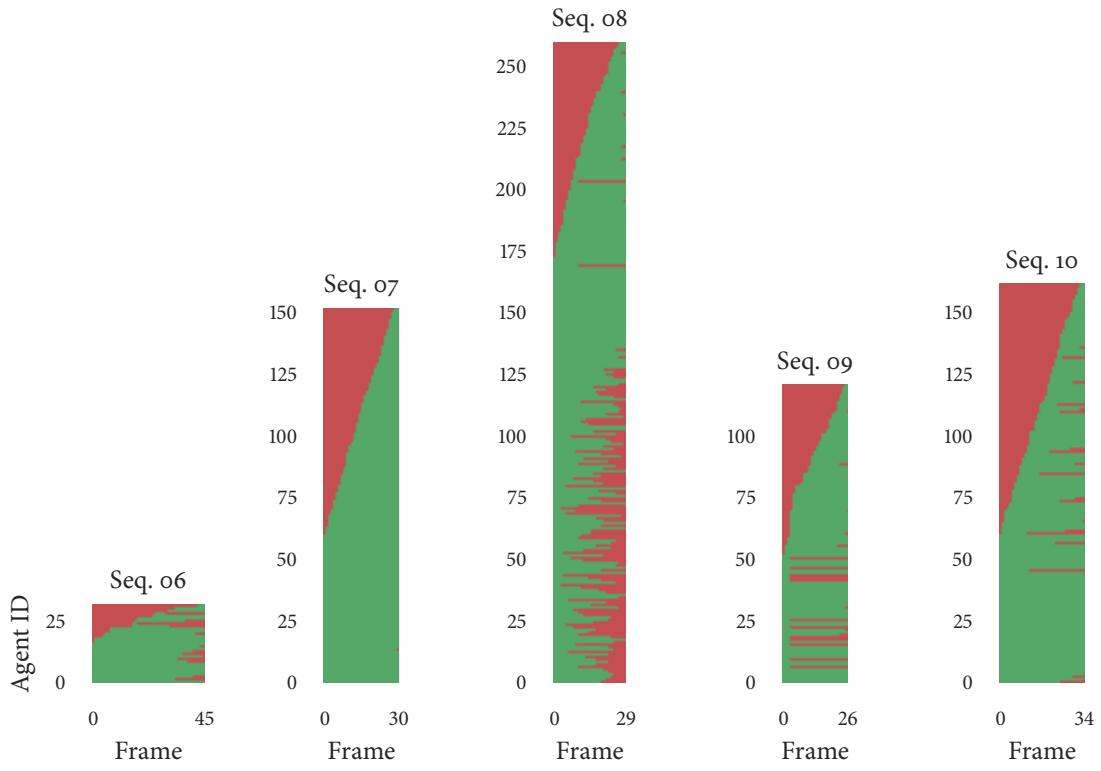


Figure A.2: A plot showing the missing datatpoints in sequences 6–10 (left to right).

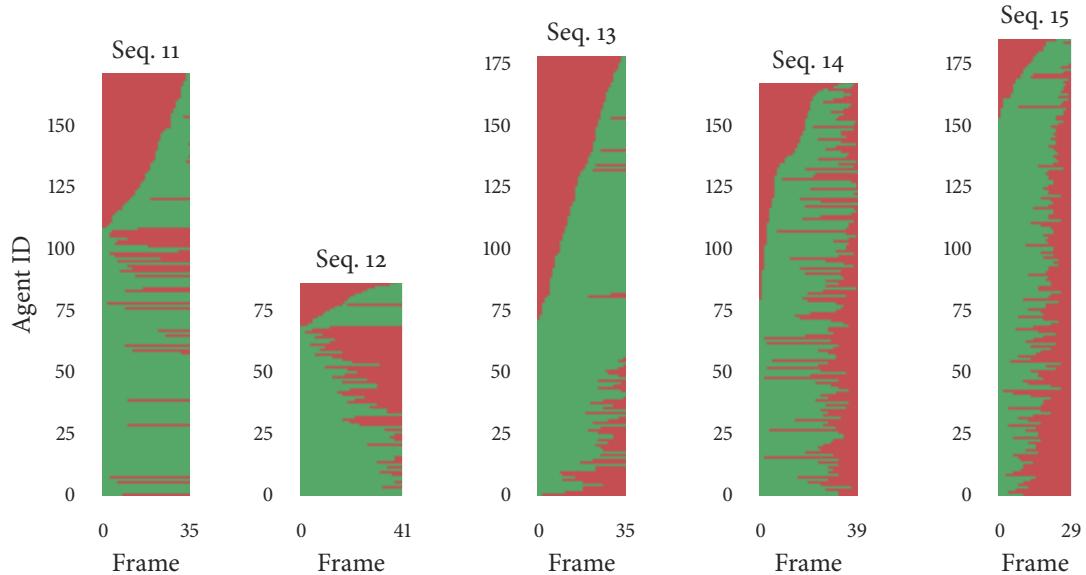


Figure A.3: Missingness in sequences 11–15 (left to right).

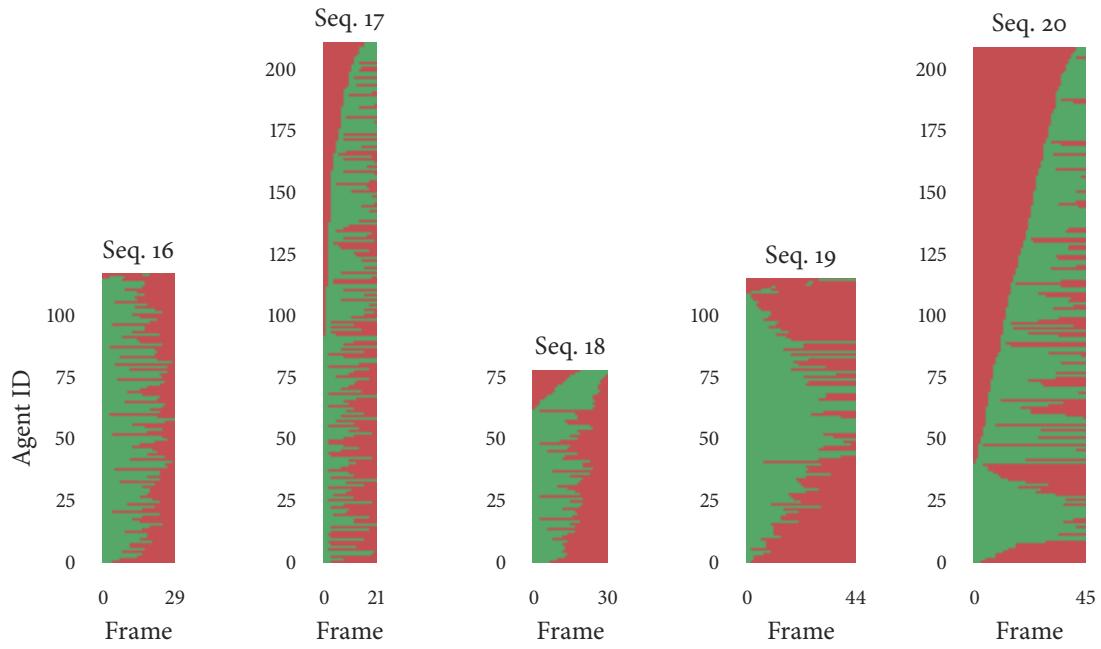


Figure A.4: Visualising the missingness in sequences 16–20 (left to right).

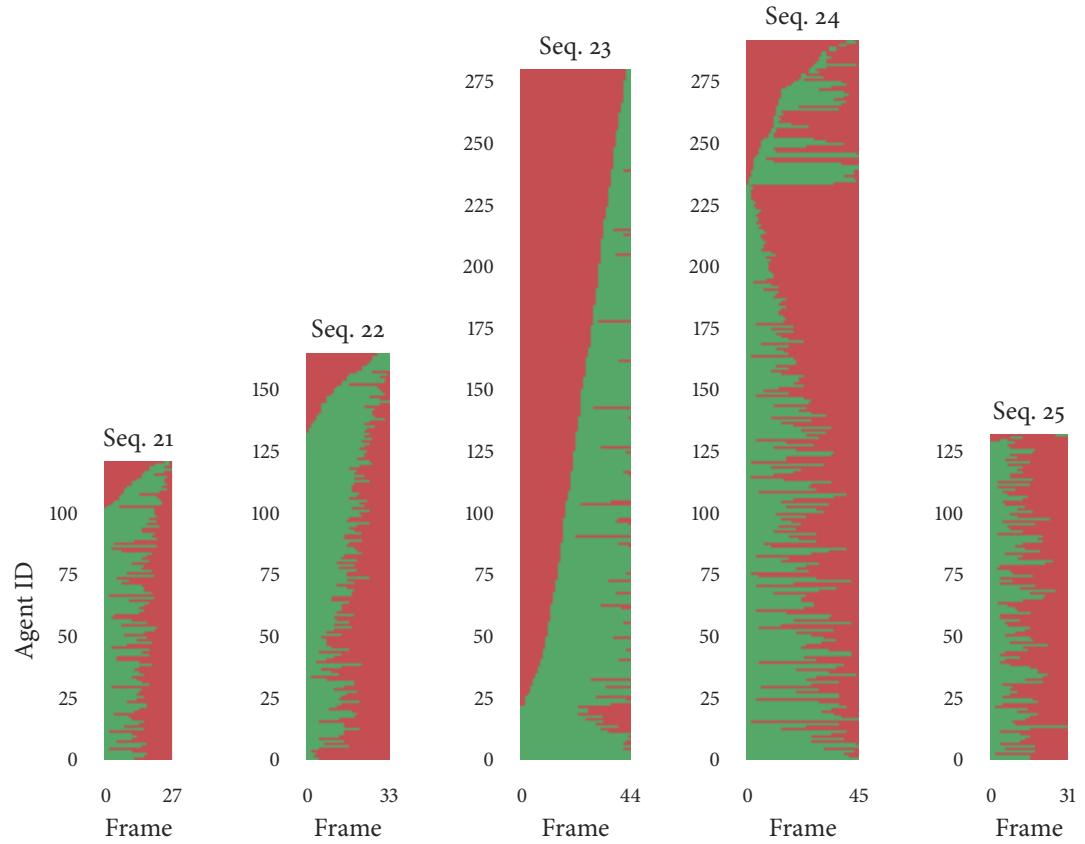
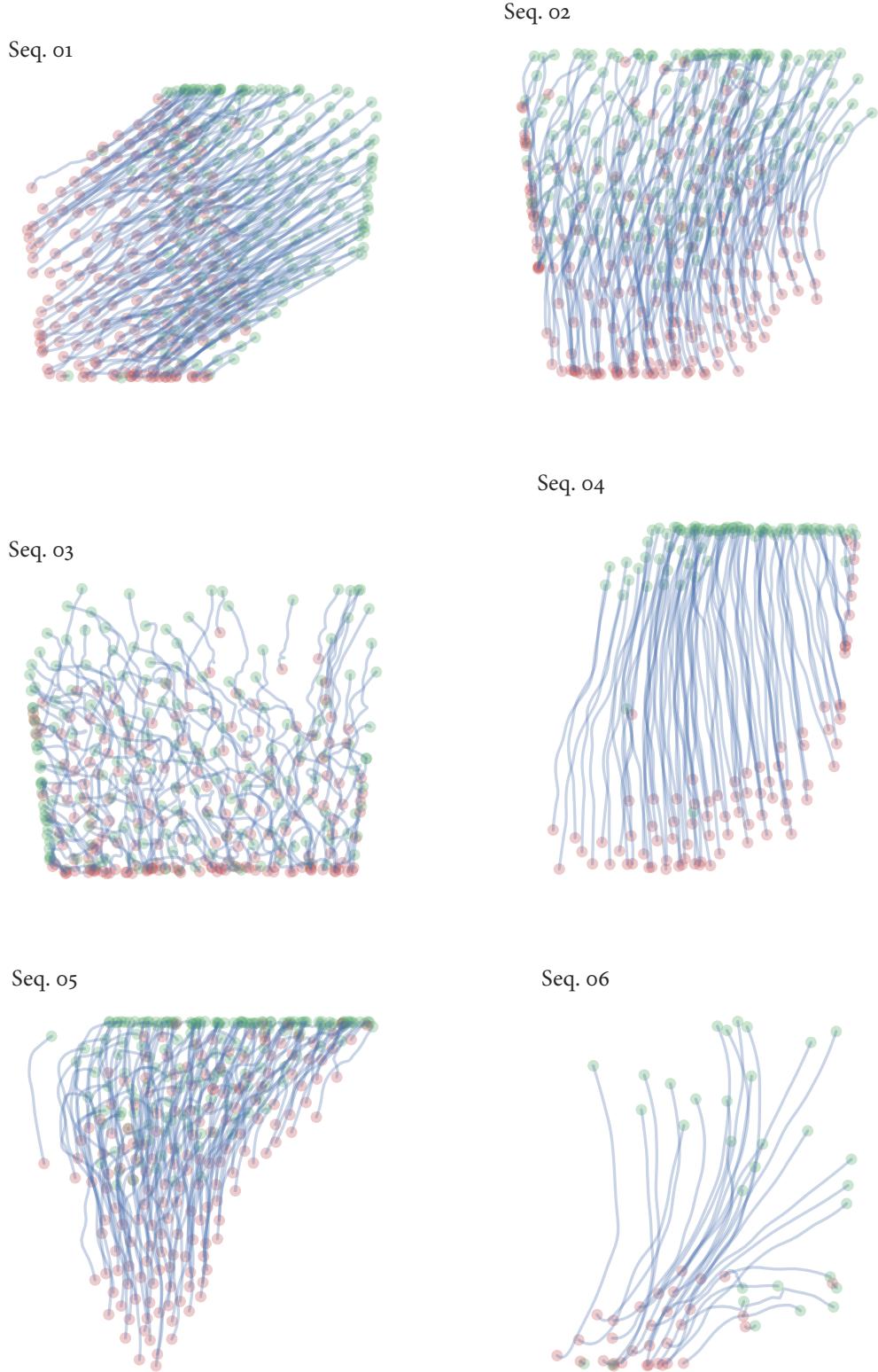
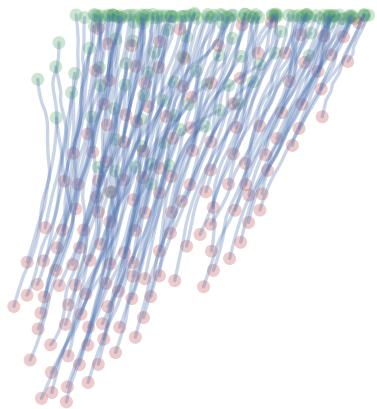


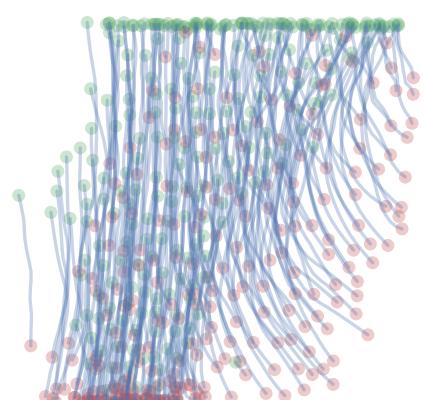
Figure A.5: Sequences 21–25 (left to right), and the distribution of missing data points.



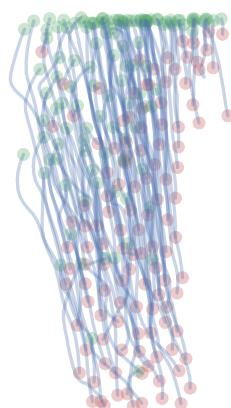
Seq. 07



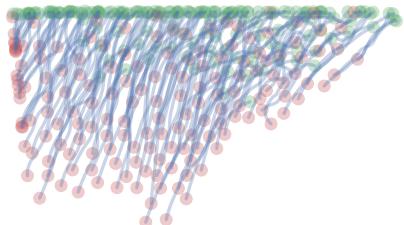
Seq. 08



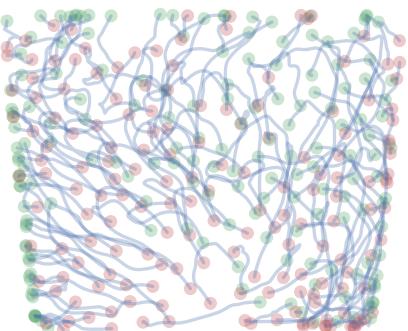
Seq. 09



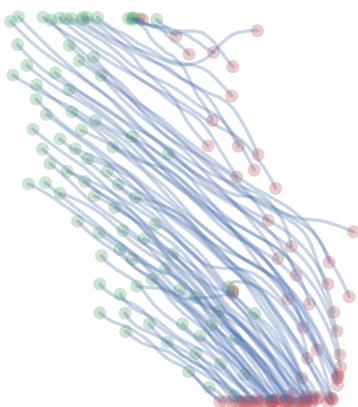
Seq. 10

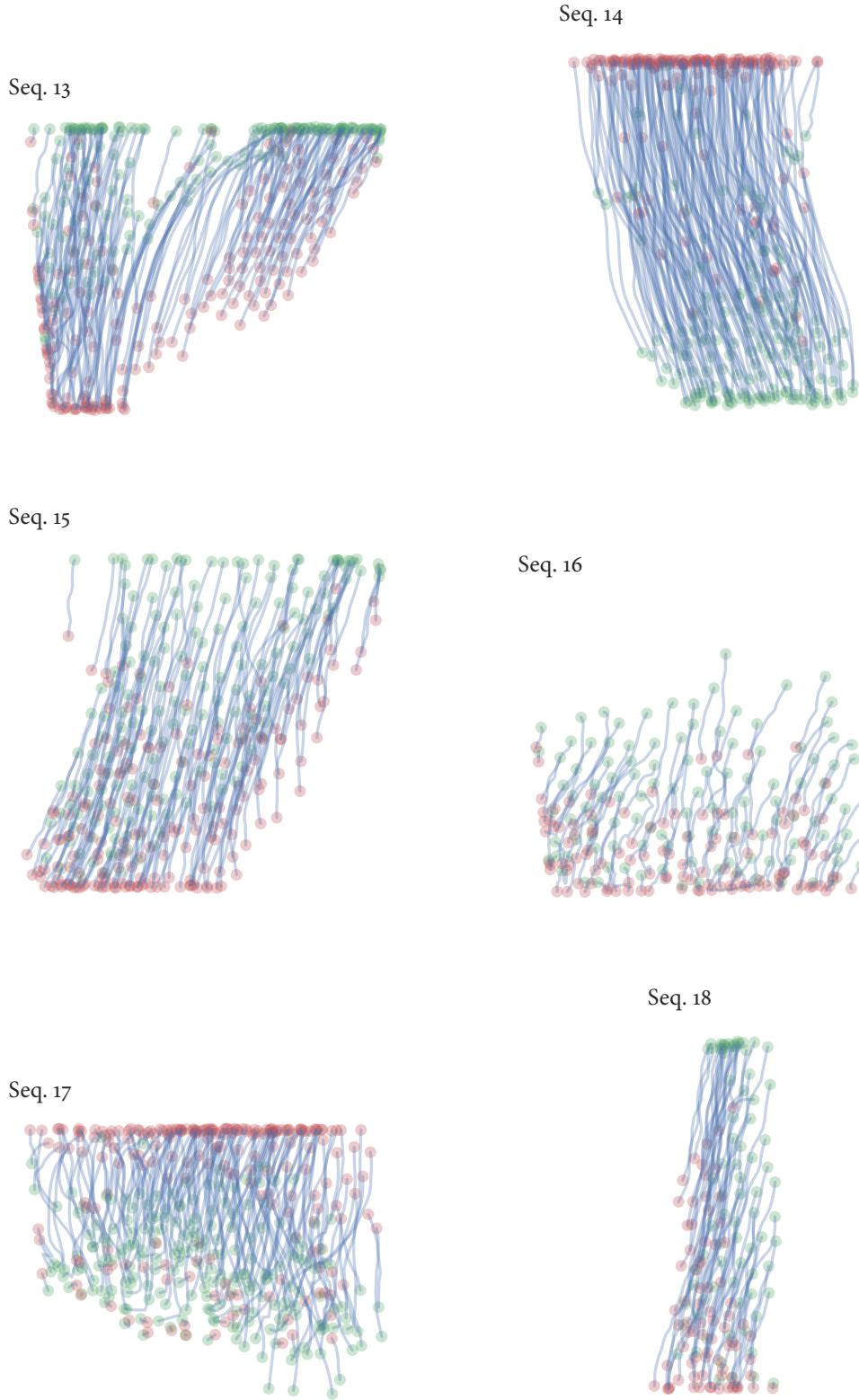


Seq. 11

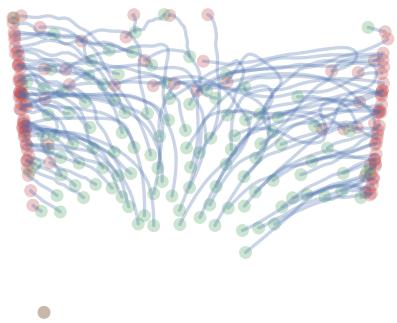


Seq. 12

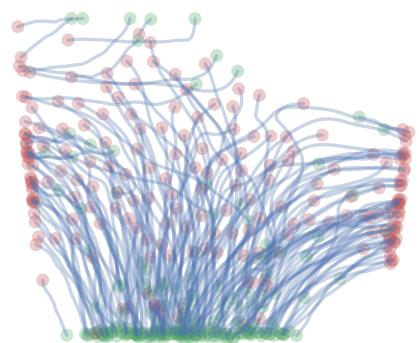




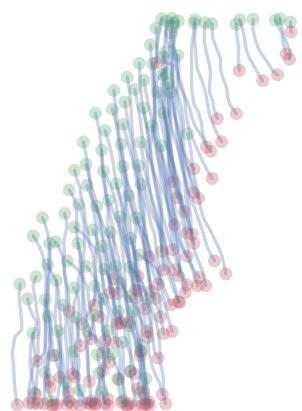
Seq. 19



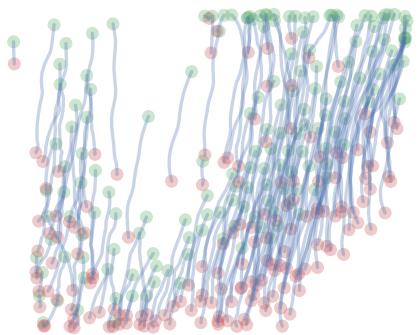
Seq. 20



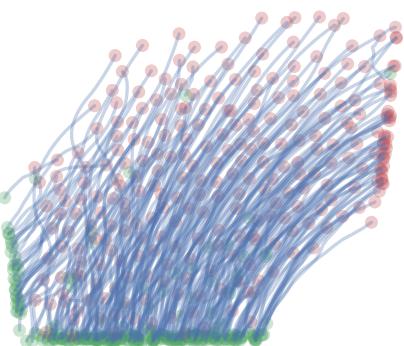
Seq. 21



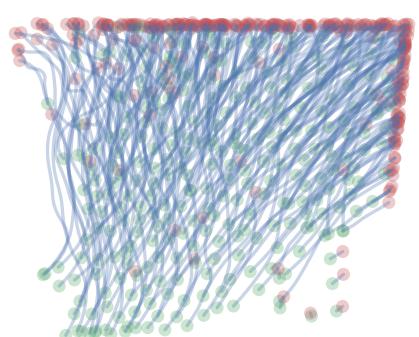
Seq. 22



Seq. 23



Seq. 24



Bibliography

- AOKI, I. 1982 A simulation study on the schooling mechanism in fish. *Bulletin of the Japanese Society of Scientific Fisheries* **48** (8), 1081–1088.
- BALLERINI, M., CABIBBO, N., CANDELIER, R., CAVAGNA, A., CISBANI, E., GIARDINA, I., LECOMTE, V., ORLANDI, A., PARISI, G., PROCACCINI, A., VIALE, M. & ZDRAVKOVIC, V. 2008 Interaction ruling animal collective behaviour depends on topological rather than metric distance: Evidence from a field study. *Proceedings of the National Academy of Sciences of the United States of America* **105** (4), 1232–1237.
- BAZAZI, S., BUHL, J., HALE, J. J., ANSTEY, M. L., SWORD, G. A., SIMPSON, S. J. & COUZIN, I. D. 2008 Collective motion and cannibalism in locust migratory bands. *Current Biology* **18**, 735–739.
- BUDGEY, R. 1998 Three dimensional bird flock structure and its implications for birdstrike tolerance in aircraft. *International Bird Strike Proceedings Committee* **24**, 207–220.
- COUZIN, I. D., KRAUSE, J., FRANKS, N. R. & LEVIN, S. A. 2005 Effective leadership and decision making in animal groups on the move. *Nature* **433**, 513–516.
- COUZIN, I. D., KRAUSEW, J., JAMESZ, R., RUXTON, G. D. & FRANKSZ, N. R. 2002 Collective memory and spatial sorting in animal groups. *Journal of Theoretical Biology* **218**, 1–11.
- CULLEN, J. M., SHAW, E. & BALDWIN, H. A. 1965 Methods for measuring the three-dimensional structure of fish schools. *Animal Behaviour* **13** (4), 534–536.
- DUNKEL, J., HEIDENREICH, S., DRESCHER, K., WENSINK, H. H., BÄR, M. & GOLDSTEIN, R. E. 2013 Fluid dynamics of bacterial turbulence. *Physical Review Letters* **110**.
- GIARDINA, I. 2008 Collective behavior in animal groups: theoretical models and empirical studies. *HFSP Journal* **2** (4), 205–219.

- GUERNON, S. & LEVIN, S. A. 1993 Self-organization of front patterns in large wildebeest herds. *Journal of Theoretical Biology* **165** (4), 541–552.
- HASTINGS, W. K. 1970 Monte carlo sampling methods using markov chains and their applications. *Biometrika* **57** (1), 97–109.
- HUTH, A. & WISSEL, C. 1992 The simulation of the movement of fish schools. *Journal of theoretical Biology* **156**, 365–385.
- KATZ, Y., TUNSTRØM, K., IOANNOUA, C. C., HUEPE, C. & COUZIN, I. D. 2011 Inferring the structure and dynamics of interactions in schooling fish. *Proceedings of the National Academy of Sciences of the United States of America* **108** (46), 18720–18725.
- LANDEAU, L. & TERBORGH, J. 1986 Oddity and the fhconfusion effectfh in predation. *Animal Behaviour* **34** (5), 1372–1380.
- LUKEMAN, R. 2009 Modelling collective behaviour in animal groups: From mathematical analysis to field work. PhD thesis, University of British Columbia.
- LUKEMAN, R., LI, Y.-X. & EDELSTEIN-KESHET, L. 2010 Inferring individual rules from collective behavior. *Proceedings of the National Academy of Sciences of the United States of America* **107** (28), 12576–12580.
- MAJOR, P. F. & DILL, L. M. 1978 The three-dimensional structure of airborne bird flocks. *Behavioral Ecology and Sociobiology* **4** (2), 111–122.
- MANN, R. P. 2011 Bayesian inference for identifying interaction rules in moving animal groups. *PLoS ONE* **6** (8).
- MERMIN, N. D. & WAGNER, H. 1966 Absence of ferromagnetism or antiferromagnetism in one- or two-dimensional isotropic heisenberg models. *Physical Review Letters* **17** (22), 1133–1136.
- METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. & TELLER, E. 1953 Equation of state calculations by fast computing machines. *Journal of Chemical Physics* **21** (6), 1087–1092.
- OKUBO, A. 1986 Dynamical aspects of animal grouping: swarms, schools, flocks, and herds. *Advanced Biophysics* **21**, 1–94.
- ORGANICK, E. I. 1966 *A Fortran IV Primer*. Addison-Wesley.

- PARTRIDGE, B. L., PITCHER, T., CULLEN, J. M. & WILSON, J. 1980 The three-dimensional structure of fish schools. *Behavioral Ecology and Sociobiology* **6** (4), 277–288.
- PITCHER, T. J. & PARRISH, J. K. 1993 *Behaviour of Teleost Fishes*. Chapman and Hall.
- REYNOLDS, C. W. 1987 Flocks, herds, and schools: A distributed behavioral model. *Computer Graphics* **21** (4), 25–34.
- ROBBINS, J. 2017 *The Wonder of Birds: What They Tell Us About Ourselves, the World, and a Better Future*. Spiegel & Grau.
- ROBERTS, G. O. & ROSENTHAL, J. S. 2001 Optimal scaling for various metropolis-hastings algorithms. *Statistical Science* **16** (4), 351–367.
- SELOUS, E. 1931 *Thought-Transference (Or What?) in Birds*. Constable & Co.
- SIMONS, A. M. 2004 Many wrongs: the advantage of group navigation. *TRENDS in Ecology and Evolution* **19** (9), 453–455.
- TONER, J. & TU, Y. 1998 Flocks, herds, and schools: A quantitative theory of flocking. *Physical Review E* **58** (4), 4828–4858.
- TOPAZ, C. M. & BERTOZZI, A. L. 2004 Swarming patterns in a two-dimensional kinematic model for biological groups. *SIAM Journal on Applied Mathematics* **65** (1), 152–174.
- TOPAZ, C. M., BERTOZZI, A. L. & LEWIS, M. A. 2006 A nonlocal continuum model for biological aggregation. *Bulletin of Mathematical Biology* **68**, 1601–1623.
- VAN LONG, L., AOYAMA, T. & INAGAKI, T. 1985 A stereo photographic method for measuring the spatial position of fish. *Bulletin of the Japanese Society of Scientific Fisheries* **51** (2), 183–190.
- VICSEK, T., CZIRÓK, A., BEN-JACOB, E., COHEN, I. & SHOCNET, O. 1995 Novel type of phase transition in a system of self-driven particles. *Physical Review Letters* **75** (6), 1226–1229.
- WEIMERSKIRCH, H., MARTIN, J., CLERQUIN, Y., ALEXANDRE, P. & JIRASKOVA, S. 2001 Energy saving in flight formation. *Nature* **413**, 697–698.
- WITTENBERGER, J. F. & HUNT, G. L. 1985 *Avian Biology*, , vol. 8. Academic Press.