# BAYESIAN INFERENCE *for* MODELS OF COLLECTIVE BEHAVIOUR

Jack Walton

June 2020

*Thesis submitted for the degree of*
*Doctor of Philosophy*

*to the*

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

**Acknowledgement**

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

**Abstract**

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

# *Contents*

# 1

# *Introduction*

## 1.1 Motivation

Many of us have been struck by the inherent beauty of animals moving collectively; starlings gathering in huge numbers at dusk to perform the most mesmerising of ballets, the entire flock moving as if some fluid object (Figure 1.1); fish forming tight milling structures in defence against predation (Figure 1.2), changing direction in the blink of an eye and with a flash of silver. Collective motion, broadly defined as the formation of macro-level structures from the interactions of individuals (Camazine et al. 2003), has been observed over many different length scales, and has been exhibited by many different species (Allee 1931). But through all its variations and incantations, the one thing that remains constant is the phenomenon's ability to capture the attention and wonder of the observer.

Through captured imaginations, and over the years, collective behaviour has become a thriving topic of multidisciplinary research, holding captive the minds of physicists, biologists, mathematicians, and many others of the scientific persuasion. Though our understanding has evolved significantly from early suggestions that collective behaviour results from thought-transference and telepathy between individuals (Selous 1931), there is still so much that remains unknown.

In many cases, the *why* of collective behaviour is broadly understood. From an evolutionary standpoint, we can reason about the benefits which collective behaviour brings to the individuals involved. For example, it is known that aggregation can provide an effective defence against predation (Landeau and Terborgh 1986), and that both foraging and

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

migration can benefit from the knowledge of the collective (Simons 2004).

Despite this, much less about the *how* of collective behaviour is known. The mechanisms which lead to the formation and maintenance of aggregations remain more elusive, and are a topic of much interest. The hope is that the study of mathematical models of flocking, and the comparison of these models with real flocking events, will shed light on the mechanics which underlie these phenomena. In recent years modern computing power has made the process of simulating these models relatively pain-free, however, a lack of quality data detailing real flocking events has made the comparison between model and data difficult.

Previously, much work has been invested in developing the 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 (Aoki 1982). Many different 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 thorough quantitative comparison between field data and theory has been lacking. This lack of quantitative comparison between model and data can largely be attributed to the scarcity of empirical data.

However, 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).

Figure 1.2: Mackerel form a milling structure as a defence against predation.

With this data, it is only now that we are in a position to make robust comparison between model prediction and real-world observation.

Using recently collected data, this thesis seeks to compare newly available data of flocking events and theoretical models. The intention is to fit multiple different models to the same dataset. With this we will be in a position to consider which of the fitted models best describes the observed data.

## 1.2 Thesis overview

We make a start in Chapter 2 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.

Bayesian statistics will be introduced to the reader in Chapter 3. Important results, techniques and algorithms from the subject will be outlined as well as any problems that a Bayesian practitioner may encounter, and how they may address these problems.

Proceeding onwards with Appendix A, the reader will be given a short introduction to directional statistics. Here we briefly discuss why one should take a little extra care to avoid pitfalls when handling circular 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, we provide a quick overview of the evolutionary advantages which collective behaviour affords individuals. After this we will discuss Eulerian and Lagrangian models: the two main modelling paradigms used to simulate flocking events. 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 amongst a group (Landeau and Terborgh 1986). In addition to this confusion effect, groups of individuals can take-in more sensory information about their environment than lone individuals are capable of, promoting the early detection of predators (Pitcher and Parrish 1993).

As well as providing defence against predation, behaving as a group can aid in foraging

for food as collections of individuals are able to gather more data about their environment than solitary individuals (Clark 1986). Collective motion is also understood to aid group navigation and migration, with the suggestion that navigational accuracy increases with group size through the 'many wrongs principle' (Simons 2004). For birds, group navigation often brings an additional 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.

Despite the advantages afforded by collective behaviour, it isn't without its dangers. For example, there is an understanding that flocking behaviours may also have the unintended consequence of actually *attracting* the attention of predators (Wittenberger and Hunt 1985). A more dramatic consequence of collective behaviour can be seen in the formation of ant mills. Ant mills occur when a group of foraging army ants become separated from the main column of a raiding swarm (Schneirla 1944). Each ant follows the ant in front of it, eventually causing the separated workers to run in a densely packed circle until they all die from exhaustion (Schneirla 1971). This phenomenon was first recorded in 1921, when William Beebe observed an ant mill with a circumference of 370 m (Beebe 1921). With a mill so large, it took an individual ant 2.5 hours to make a single revolution of the mill (Surowiecki 2005).

As we have seen, much of the *why* of collective behaviour can be understood by considering the evolutionary advantages which group behaviour affords individuals of the group. However, we have still yet to broach the *how* of collective behaviour.

## 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

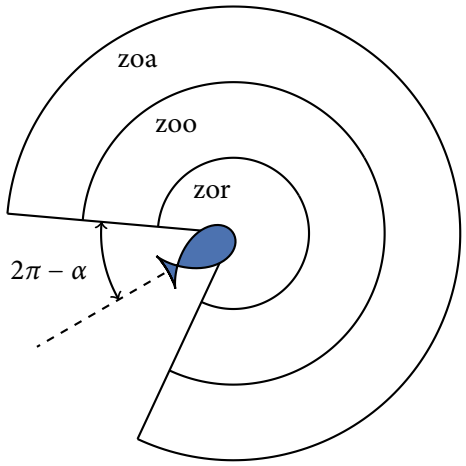



Figure 2.1: An illustration of the area around an agent (here, a fish) partitioned into multiple zones. zor: zone of repulsion, zoo: zone of orientation (or alignment), zoa: zone of attraction. The missing segment behind the agent represents a blind zone into which it cannot see.

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 with which to interact. 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 fish schooling in two-dimensions. Here it was shown that collective behaviour can arise from simple interactions at an individual level, *without* the need of a leader, and *without* each individual having information about the movement of the group as a whole. The model simulated zonal interactions in which the area around each fish was partitioned into zones of repulsion, alignment and attraction (avoid, parallel and approach in the original publication). The partitioning of space in this way is illustrated 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 and Wissel 1992).

Following this, Reynolds (1987) formulated a mathematical model, motivated by the production of computer animations, which described the movement of birds flocking in


Figure 2.2: The animation company Iloura lent heavily on the Massive software to orchestrate army formations, fighting soldiers and horse actions for the dramatic scenes in HBO's Game of Thrones episode 'Battle of the Bastards'.

three-dimensional space. To produce more aesthetically pleasing animations, the software, "Boids", 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), originally developed by Stephen Regelous for Peter Jackson's Lord of the Rings trilogy (Koeppel 2002). This software was used to help generate the striking battle sequences of the trilogy, where each individual orc, elf and other miscellaneous creature of middle-earth was simulated according to the rules of an agent based model (Robbins 2017). In 2004, Regelous received the Scientific and Engineering Award from the Academy of Motion Picture Arts and Sciences for his work on Massive. Since then, Massive has been used in films such as Inception, Harry Potter and the order of the Phoenix, James Bond's Spectre, and HBO's hit TV series Game of Thrones (Figure 2.2). With this legacy in mind, in 2018 Regelous received an Emmy award to recognise his contribution to the entertainment industry.

Not motivated by the lure of an Academy Award, but instead motivated by research within statistical physics, Vicsek et al. (1995) introduced a simple two-dimensional model

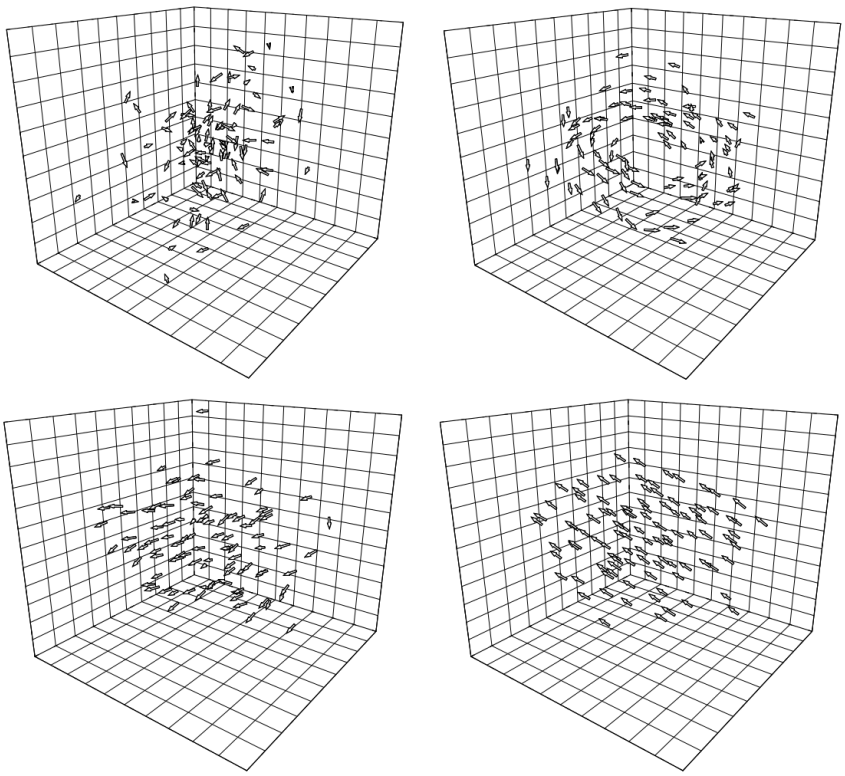


Figure 2.3: Taken from Iain D. Couzin et al. (2002), the different steady-state solutions (swarm, torus, dynamically parallel and highly parallel) obtained by making small changes to model parameters of a three-dimensional flocking model.

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 the "Vicsek Model" (VM). Despite 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 and Wagner 1966). However, VM is out of equilibrium and Mermin-Wagner only applies to systems in equilibrium.

Later, models were developed to explore the movements of mammals and other vertebrate groups. Using a three-dimensional model that follows the zonal approach of Aoki (1982), Iain D. 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 2.3. 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 Ian D. 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 and 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 (Figure 2.4). The large-scale front patterns seen in the aerial photography were reproduced by their model.

Later, Toner and Tu (1998) introduced a quantitative continuum theory of flocking. There are similarities between the hydrodynamic equations introduced by the authors and

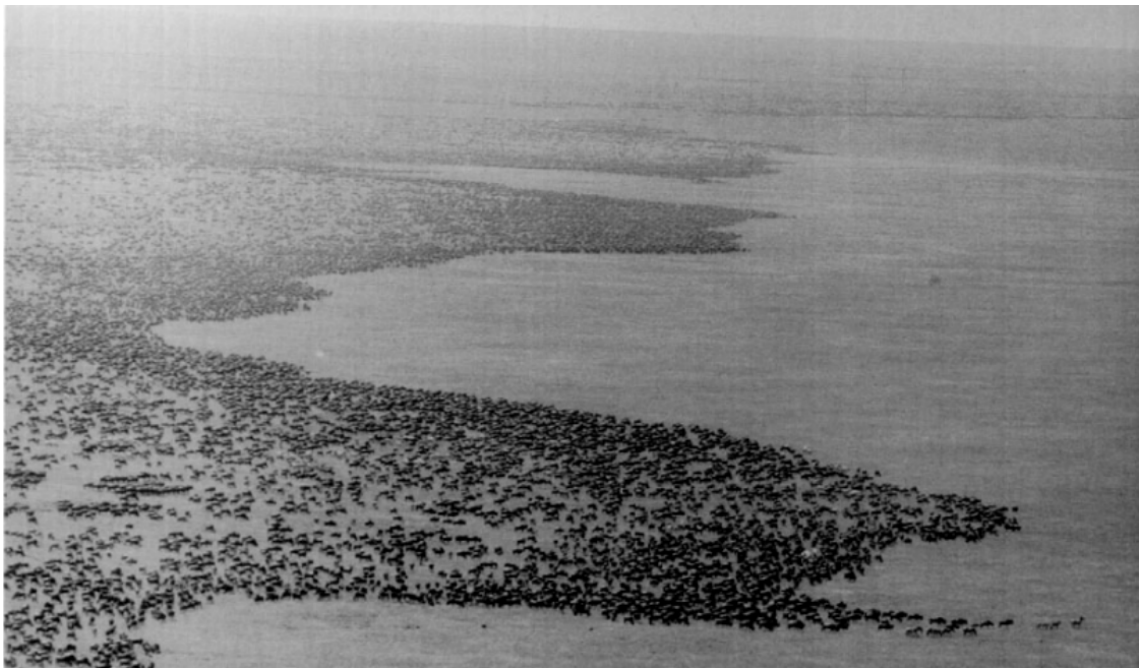


Figure 2.4: Aerial photography of migrating wildebeest showing large-scale front patterns, as presented in the work by Guernon and Levin (1993).

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 and Bertozzi 2004) and stationary clump solutions (Topaz, Bertozzi and Lewis 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 (Giardina 2008). 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 apriorstic 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 correct-

ness. 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 dynamic component. The first breakthrough came from Cullen, Shaw and Baldwin (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; Long, Aoyama and Inagaki 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 and 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 2.5). 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

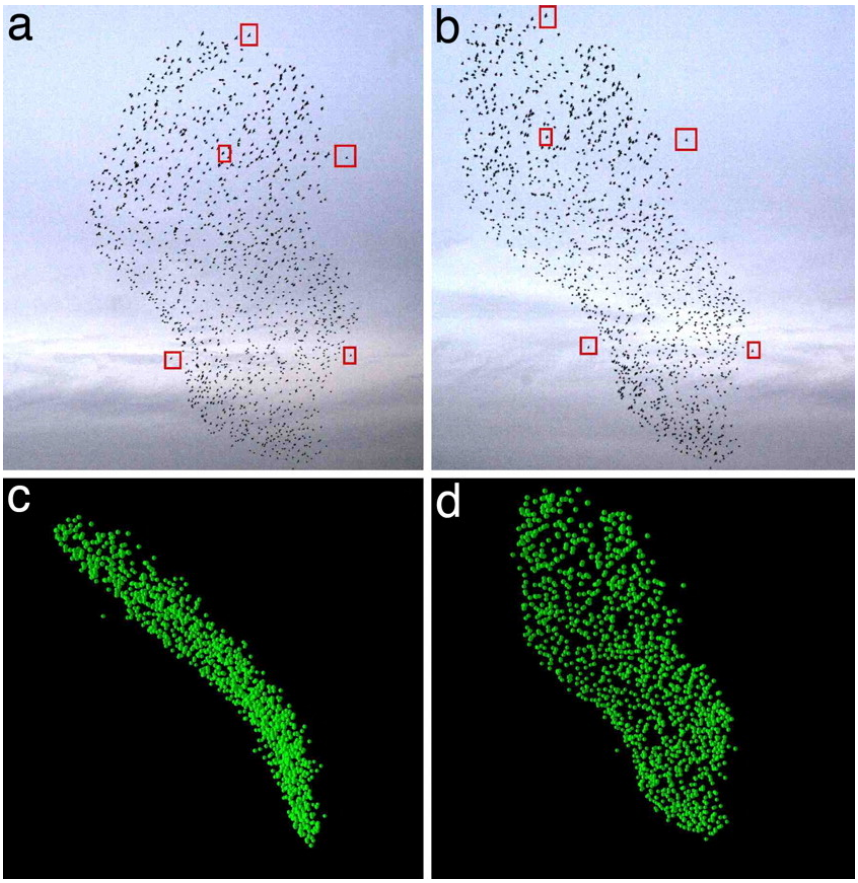



Figure 2.5: A flock of 1246 starlings reconstructed in three dimensions. Photographs taken at the same instant but 25m apart (a–b) are used to reconstruct their three-dimensional positions (c–d). To perform reconstruction Ballerini et al. (2008) needed to match each bird in (a) to its corresponding image in (b). The red squares show five matched pairs of birds.

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, Li and Edelstein-Keshet (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

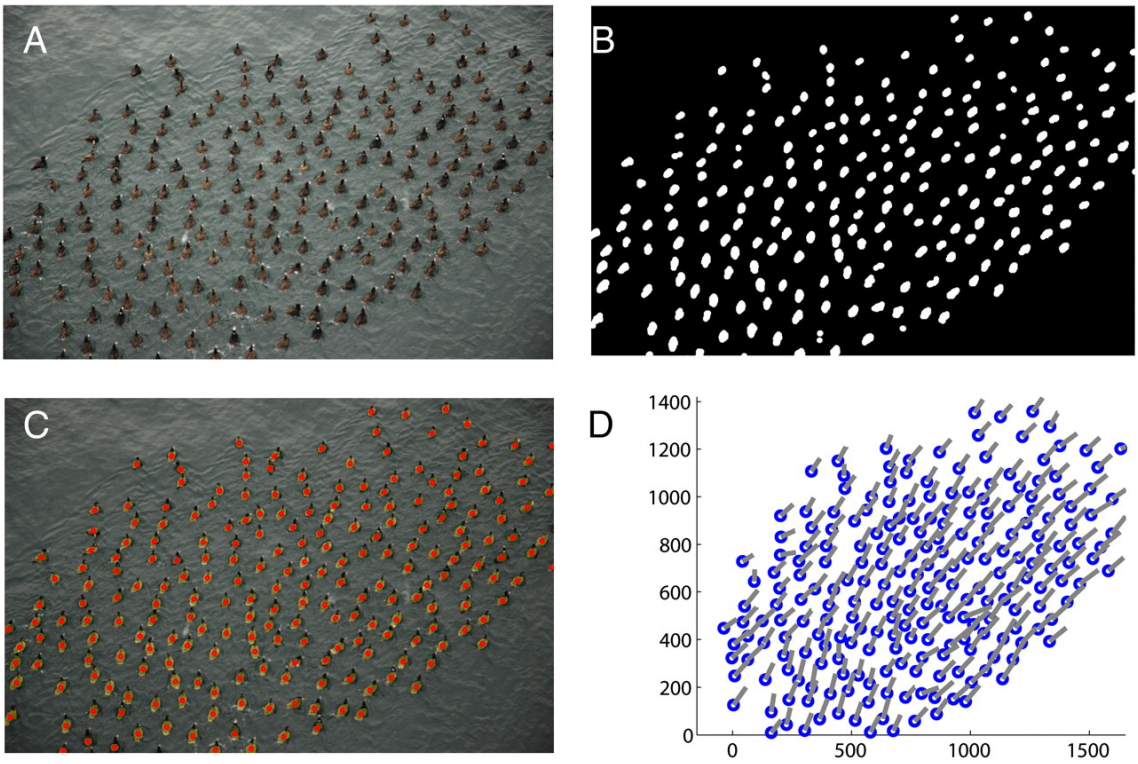



Figure 2.6: Image field data and the process of transformations to extract positions of a flock of surf scoters (Lukeman, Li and Edelstein-Keshet 2010).

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 discussed 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 distributions.

The agents in Mann's model moved under a weighted sum of alignment and attraction. After ten time steps the simulated data transitioned from disordered motion to a steady state rotating mill. The author then compared the ability to infer the weighting parameter, 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 mathematical models of collective behaviour to real data. Bayesian inference allows a 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 newly acquired data to generalisations of a popular agent-based 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 perhaps most importantly, discuss when our methodologies may fail us.

## 3.1 Bayesian inference

Having observed data $x$ we wish to quantify beliefs and uncertainties about parameters $\theta = (\theta_1, \theta_2, \ldots, \theta_d)^T$. Given the observed data, the likelihood function of the parameters is defined as:

$$L(\theta \,|\, x) = f(x \,|\, \theta). \tag{3.1}$$

The likelihood details the probability density of the data in terms of the parameters. We may then specify our prior knowledge about the parameters $\theta$ through the prior distribution $\pi(\theta)$. Bayes Theorem can then be used to form our posterior beliefs from the likelihood

function and our prior beliefs:

$$\pi(\theta \mid x) = \frac{\pi(\theta)L(\theta \mid x)}{\int_\theta \pi(\theta)L(\theta \mid x)\,\mathrm{d}\theta}. \tag{3.2}$$

As the integral in the denominator is not a function of $\theta$ we may consider it a constant of proportionality. With this we can express our posterior beliefs as proportional to the product of the likelihood and our prior beliefs:

$$\pi(\theta \mid x) \propto \pi(\theta) \times L(\theta \mid x),$$
$$\text{posterior} \propto \text{prior} \times \text{likelihood}.$$

## 3.2 Markov chain Monte Carlo (MCMC)

For the most part, the normalising constant (given in the denominator of Equation (3.2)) 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 $\pi(\theta \mid x)$, whilst avoiding evaluating the bothersome normalising constant.

### 3.2.1 *Metropolis–Hastings*

The Metropolis–Hastings algorithm is a popular MCMC scheme. The algorithm was introduced by Metropolis, Rosenbluth et al. (1953) in a now classic paper, and was later generalised by Hastings (1970). The algorithm works by constructing a Markov chain which has stationary distribution equivalent to the target distribution.

The algorithm begins by initialising a Markov chain with parameters $\theta^{(0)}$. Next, the algorithm proposes new parameter values $\theta^\star$ from a proposal distribution $q(\theta^\star \mid \theta^{(i-1)})$. These proposed values are accepted with probability $\alpha(\theta^\star \mid \theta^{(i-1)})$. If the proposal is accepted the next state of the Markov chain is set to the proposed values, otherwise the next state is set to the current values. The acceptance probability depends on a ratio of the posterior density evaluated at the current values and the posterior density evaluated at the proposed values. Because of this, the normalising constants cancel in this ratio and we see that the target distribution only need be known up to a constant of proportionality. This process of proposing and accepting or rejecting proposals continues until a satisfactory number of draws are made. Metropolis–Hastings is described more formally in Algorithm 1.

Algorithm 1: Targeting $\pi(\theta \mid x)$ with $S$ iterations of the Metropolis–Hastings algorithm.

Initialise chain with $\theta^{(0)}$
**for** $i = 1$ **to** $S$ **do**
  Propose $\theta^\star \sim q(\theta^{(i)} \mid \theta^{(i-1)})$
  Construct acceptance probability $\alpha(\theta^\star \mid \theta^{(i-1)})$ as

$$\alpha(\theta^\star \mid \theta^{(i-1)}) = \min\left\{1, \frac{\pi(\theta^\star)\, L(\theta^\star \mid x)}{\pi(\theta^{(i-1)})\, L(\theta^{(i-1)} \mid x)} \frac{q(\theta^{(i-1)} \mid \theta^\star)}{q(\theta^\star \mid \theta^{(i-1)})}\right\}.$$

  Draw $u \sim \text{Uniform}(0, 1)$
  **if** $u \leq \alpha(\theta^\star \mid \theta^{(i-1)})$ **then**
    *# Accept proposal*
    $\theta^{(i)} \leftarrow \theta^\star$
  **else**
    *# Reject proposal*
    $\theta^{(i)} \leftarrow \theta^{(i-1)}$
  **end if**
**end for**

## *Choosing a Proposal Distribution*

The practitioner must choose a suitable proposal distribution $q(\theta^\star \mid \theta)$. Ideally the choice of proposal distribution will give rapid convergence to $\pi(\theta \mid x)$ and efficiently explore the support of $\pi(\theta \mid x)$. A special case of Metropolis–Hastings arises when the proposal distribution is symmetric, that is

$$q(\theta^\star \mid \theta) = q(\theta \mid \theta^\star).$$

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

$$\alpha(\theta^\star \mid \theta^{(i-1)}) = \min\left\{1, \frac{\pi(\theta^\star)}{\pi(\theta^{(i-1)})} \frac{L(\theta^\star \mid x)}{L(\theta^{(i-1)} \mid x)}\right\}.$$

The random walk sampler is a popular implementation of Metropolis–Hastings which makes use of symmetric proposals. With this sampler proposals are realised as

$$\theta^\star = \theta^{(i-1)} + \omega^{(i-1)},$$

where the $\omega$ are sampled as

$$\omega^{(i-1)} \sim \mathcal{N}_d(0, \Sigma),$$

and $\mathcal{N}_d$ denotes a $d$-dimensional multivariate normal distribution. The parameter $\Sigma$ 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 $\Sigma$ can be used to control the mixing of chains. So, naturally, we wish to identify some 'optimum' $\Sigma$ to try and improve mixing. Such a tuning parameter should allow rapid convergence to $\pi(\theta \mid x)$ and facilitate exploration of the entire support of the target. If the target distribution is Gaussian, it has been shown that 0.234 is an optimum acceptance probability to try achieve (Roberts and Rosenthal 2001). In an attempt to tune $\Sigma$ to obtain the optimum acceptance probability, a common technique is to use

$$\Sigma = \frac{2.38^2}{d} \widehat{\text{Var}}(\theta \mid x).$$

However, even with strategies to try select some optimum innovation structure, random walk samplers tend to perform poorly in high-dimensional spaces. Consider that as the dimension of a problem increases, the probability of proposing a point out in the tails of the target distribution increases. As a result the acceptance probability becomes small and produces a Markov chain which rarely moves. The acceptance probability can be increased by choosing a $\Sigma$ which results in smaller innovations. However, this has the consequence of producing a Markov chain which explores the sample space slowly, and converges to the target distribution slowly.

Fortunately, there exist more sophisticated proposal mechanisms which perform better than random walk samplers in higher dimensional problems. One such sampler is represented by Hamiltonian Monte Carlo, which seeks to utilise information about the gradient of the target distribution to inform innovations. With a problem of dimension $d$, the computational expense of a random-walk sampler is $O(d^2)$, whereas the cost of Hamiltonian Monte Carlo is roughly $O(d^{5/4})$ (Creutz 1988).

### 3.2.2 *Hamiltonian Monte Carlo (HMC)*

Hamiltonian Monte Carlo, originally Hybrid Monte Carlo, was first introduced by Duane et al. (1987). In this now landmark paper, the HMC algorithm was detailed and used for numerical simulation of Lattice Quantum Chromodynamics. Following this, Radford Neal recognised the potential statistical applications of HMC, and used it in his work on Bayesian neural network models (Neal 1995). However, it wasn't really until Neal's 2011 review (Neal 2011) that HMC received mainstream attention in statistical computing (Betancourt 2017).

Hamiltonian Monte Carlo is a realisation of the Metropolis–Hastings algorithm. Here, new parameter values are proposed by computing trajectories of motion according to Hamiltonian dynamics. With this proposal mechanism it is possible to propose parameter values which are distant from the current state, but which retain a high probability of acceptance. As a result, this proposal mechanism represents an efficient method of traversing a parameter space, and circumvents the slow exploration of the parameter space typically experienced by random walk samplers in higher-dimensions.

## *Mathematical formulation*

Hamiltonian mechanics represents a reformulation of classical mechanics. In Hamiltonian mechanics a system is described by a $d$-dimensional position vector, $\theta$, and a $d$-dimensional momentum vector, $p$. This system then evolves through time according to Hamilton's equations:

$$\begin{aligned} \frac{\partial p_i}{\partial t} &= -\frac{\partial \mathcal{H}}{\partial \theta_i}, \\ \frac{\partial \theta_i}{\partial t} &= \frac{\partial \mathcal{H}}{\partial p_i}, \end{aligned} \tag{3.3}$$

where $i = 1, \ldots, d$ and $\mathcal{H}(\theta, p)$ is the Hamiltonian. The Hamiltonian is often interpreted to represent the total energy of a system, which can be considered as the sum of the kinetic energy, $T$, and potential energy, $V$:

$$\mathcal{H}(\theta, p) = T(p) + V(\theta). \tag{3.4}$$

We wish to explore our target distribution (typically the posterior distribution) as if evolving some Hamiltonian system. This can be achieved if we expand our $d$-dimensional parameter space into $2d$-dimensional phase space. Our current state can be considered as the position vector, $\theta$. Introducing auxiliary momentum variables, $p$, expands our parameter space into phase space, as desired.

With our parameter space extended to phase space, we must also expand our target distribution to phase space. To do so we formulate the canonical distribution, a joint density function over phase space:

$$\pi(\theta, p) = \pi(p \,|\, \theta)\, \pi(\theta). \tag{3.5}$$

The momentum is typically introduced as:

$$p \mid \theta \sim \mathcal{N}_d(0, M), \tag{3.6}$$

where $M$ is a positive-definite "mass matrix", often chosen as the identity matrix or some scalar multiple of the identity matrix. See that marginalising out the momentum in Equation (3.5) recovers the target distribution.

To proceed, we consider expressing the canonical distribution as the negative exponent of a Hamiltonian:

$$\pi(\theta, p) = \exp\{-\mathcal{H}(\theta, p)\}. \tag{3.7}$$

Taking the logarithm of Equation (3.7) and using Equation (3.5) we see

$$\mathcal{H}(\theta, p) = -\log \pi(p \mid \theta) - \log \pi(\theta). \tag{3.8}$$

Recall from Equation (3.4) that the total energy in a system can be considered as the sum of the system's kinetic energy and potential energy. If we compare Equation (3.4) and Equation (3.8) we can see that we have constructed a system with kinetic energy given by the negative logarithm of the momentum density, and potential energy given by the negative logarithm of the target density, that is:

$$T(p) = -\log \pi(p \mid \theta) \quad \text{and} \quad V(\theta) = -\log \pi(\theta).$$

### *Computer implementation, NUTS & Stan*

Now that we have described HMC we are in a position to consider its implementation *in silico*. For computer implementation we must first be able to approximate solutions to Hamilton's equations. Such approximations can be achieved by discretising time using some small time step $\epsilon$. Next, the practitioner must also choose the number of steps $L$ for which to simulate Hamilton's equations. With this in place, the practitioner typically implements the leapfrog method to solve Hamilton's equations (Neal 2011). Algorithm 2 details a realisation of HMC in practice, and demonstrates the leapfrog method to simulate Hamiltonian mechanics.

As we have seen, in implementing HMC it is left to the practitioner to choose appropriate values for $L$ and $\epsilon$. Unfortunately, making a poor choice for either of these parameters can result in a significant decrease in the performance of HMC (Hoffman and Gelman 2014). Actually, $\epsilon$ can be tuned during the algorithm's implementation, using ideas from

Algorithm 2: Targeting $\pi(\theta \mid x)$ with $S$ iterations of Hamiltonian Monte Carlo, using $L$ steps and discretisation $\epsilon$.

Initialise chain with $\theta^{(0)}$
**for** $i = 1$ **to** $S$ **do**
  Draw momentum $p^{(i-1)} \sim \mathcal{N}_d(0, M)$
  $p^\star \leftarrow p^{(i-1)}$
  $\theta^\star \leftarrow \theta^{(i-1)}$
  *# Simulate Hamiltonian with leapfrog method for L steps and discretisation $\epsilon$*
  **for** $j = 1$ **to** $L$ **do**
    $\theta^\star, p^\star \leftarrow \text{Leapfrog}(\theta^\star, p^\star, \epsilon)$
  **end for**
  Construct acceptance probability $\alpha(\theta^\star, p^\star \mid \theta^{(i-1)}, p^{(i-1)})$ as

$$\alpha(\theta^\star, p^\star \mid \theta^{(i-1)}, p^{(i-1)}) = \min\left\{1, \exp\left[\mathcal{H}(\theta^{(i-1)}, p^{(i-1)}) - \mathcal{H}(\theta^\star, p^\star)\right]\right\}$$

  Draw $u \sim \text{Uniform}(0, 1)$
  **if** $u \leq \alpha(\theta^\star, p^\star \mid \theta^{(i-1)}, p^{(i-1)})$ **then**
    *# Accept proposal*
    $\theta^{(i)} \leftarrow \theta^\star$
  **else**
    *# Reject proposal*
    $\theta^{(i)} \leftarrow \theta^{(i-1)}$
  **end if**
**end for**

**function** Leapfrog($\theta$, $p$, $\epsilon$)
  $p \leftarrow p + \frac{\epsilon}{2}\frac{\partial V}{\partial \theta}$
  $\theta \leftarrow \theta + \epsilon M^{-1} p$
  $p \leftarrow p + \frac{\epsilon}{2}\frac{\partial V}{\partial \theta}$
  **return** $\theta$, $p$

the adaptive MCMC literature. However, there is no easy way to select a value of $L$ *a priori*. Typically, a practitioner will have to make multiple costly tuning runs in order to select an appropriate value of $L$.

It was with this tuning problem in mind that Hoffman and Gelman (2014) introduced the No-U-Turn Sampler (NUTS). This algorithm extends HMC and eliminates the need for the parameter $L$. Using primal-dual averaging the authors were also able to tune $\epsilon$ during the implementation of the algorithm. Altogether then, NUTS represents an implementation of HMC where the practitioner is left free of the obligation of choosing any tuning parameters.

Although NUTS relieves the practitioner the obligation of selecting parameters $\epsilon$ and $L$,

its implementation remains far from trivial. Here enters Stan. Stan, named after Stanislaw Ulam, one of the original pioneers of Monte Carlo methods (Metropolis and Ulam 1949), is a probabilistic programming language implemented in C++ (Gelman, Lee and Guo 2015). Stan requires the user to construct a Stan program, which specifies how to compute the log-posterior density for their model of interest (Stan Development Team 2015). With this, Stan implements the NUTS algorithm, and returns to the user samples from the desired posterior density.

### 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.

## 3.3 Model selection

Ideally, to test the predictive accuracy of our fitted models, we would wait for out-of-sample data (new data, distinct from that used in our fitting). However, this is often not viable. One way around this problem is to use leave-one-out cross-validation (LOO-CV). The idea here is to split the data set into training data and test data, perform the model fitting on the training data, and assess fit with the test data. Unfortunately, this method comes at computational expense; LOO-CV can necessitate performing up to $n$ model fits (where $n$ is the number of data points). To avoid this expense it is common to assess predictive accuracy using within-sample data. There exist a variety of information criteria which do exactly this.

### 3.3.1 *Information criteria*

An appealing idea is to assess predictive accuracy using within-sample data. Established methods to do this include AIC, DIC and WAIC: Akaike, Deviance and Widely Available

Information Criterion, respectively. To compute AIC or DIC it is necessary to evaluate the posterior density conditioning on a point estimate. However, WAIC has the more desirable property of averaging over the entire posterior distribution. Because of this reason Gelman, Carlin et al. (2013) find WAIC more appealing than AIC and DIC.

To compute WAIC for a given fit it is first necessary to compute the log pointwise predictive density (lppd), defined as:

$$\text{lppd} = \sum_{j=1}^{n} \log\left(\frac{1}{S}\sum_{i=1}^{S} \pi(\theta^{(i)} \mid x_j)\right),$$

where $S$ is the number of posterior samples. The log pointwise predictive density is a biased estimator of elppd, the expected log pointwise predictive density for a new dataset. WAIC accounts for this bias by adding a correction for the effective number of parameters in the fit, computed as:

$$p_{\text{WAIC}} = \sum_{j=1}^{n} V_{i=1}^{S} \log\big(\pi(\theta^{(i)} \mid x_j)\big),$$

where $V_{i=1}^{S} \log\big(\pi(\theta^{(i)} \mid x_j)\big)$ represents the posterior variance of the log predictive density for data point $x_j$. With this correction WAIC gives an approximately unbiased estimate of elppd. It is common to work with information criteria on the deviance scale, with this Watanabe (2009) defines WAIC as:

$$\text{WAIC} = -2\,\text{lppd} + 2p_{\text{WAIC}}.$$

# A

# *Directional statistics*

Circular data arises naturally in the study of collective behaviour; most commonly, in describing the direction of motion of individuals. Given some dataset, the first instinct of the scientist 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.

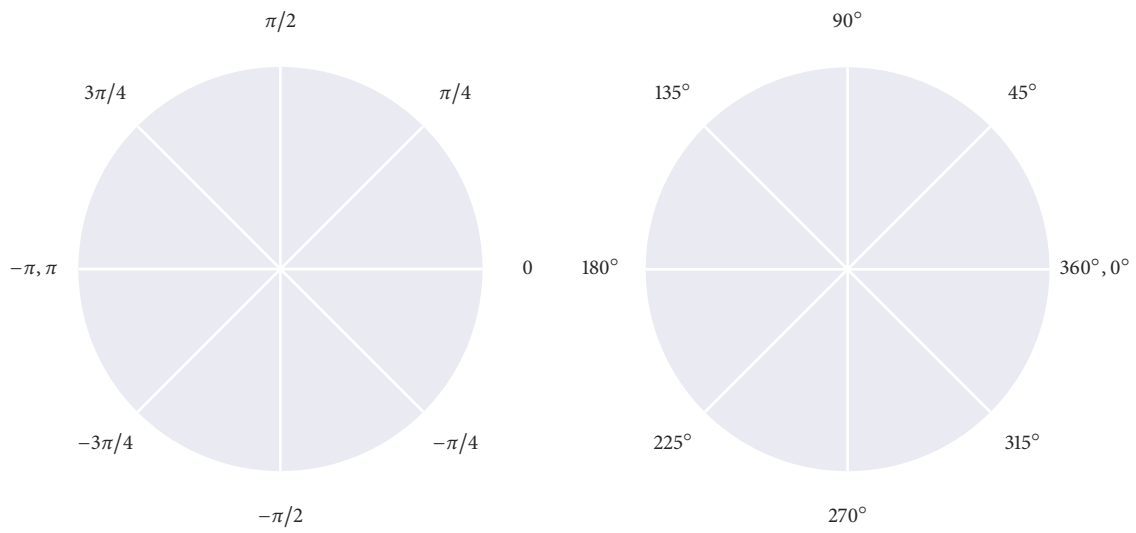
In this appendix we shall consider why standard techniques, methods and summaries are inappropriate to use with circular data. After this realisation, we proceed to introduce some useful techniques which can be used to handle and visualise directional data.

## A.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 $\pi/180°$.

In this thesis we define the zero-direction as the direction from the point $(0, 0)$ 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 A.1(a). Occasionally, we shall appeal to degrees and their comparative intuitiveness, and in these cases we shall use the setup illustrated in Figure A.1(b).




(a) (b)

Figure A.1: Visualising the conventions used in this thesis to measure (a): radians and (b): degrees.

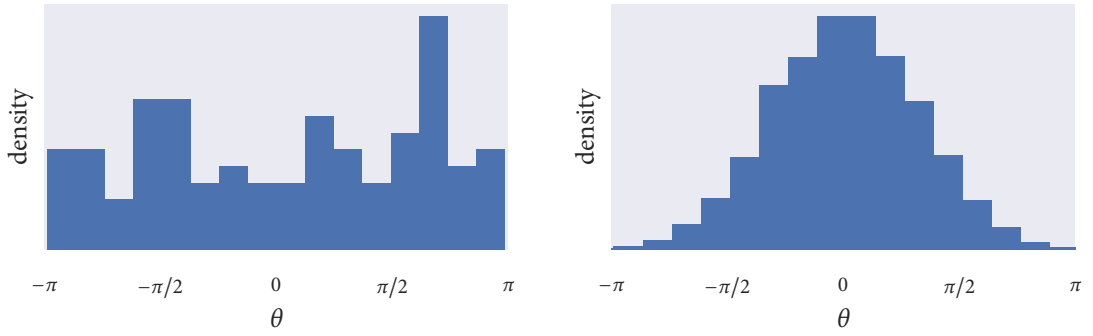
## A.2 Visualisation

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; consider, for example, interpreting the directions plotted in Figure A.2. Polar histograms (sometimes known as rose plots) make for more intuitive representations of angles. Instead of using bars, as the histogram does, the rose plot bins data into sectors of a circle. Here, the *area* of each sector is constructed to be proportional to the frequency of data points in the corresponding bin (Mardia and Jupp 2009).

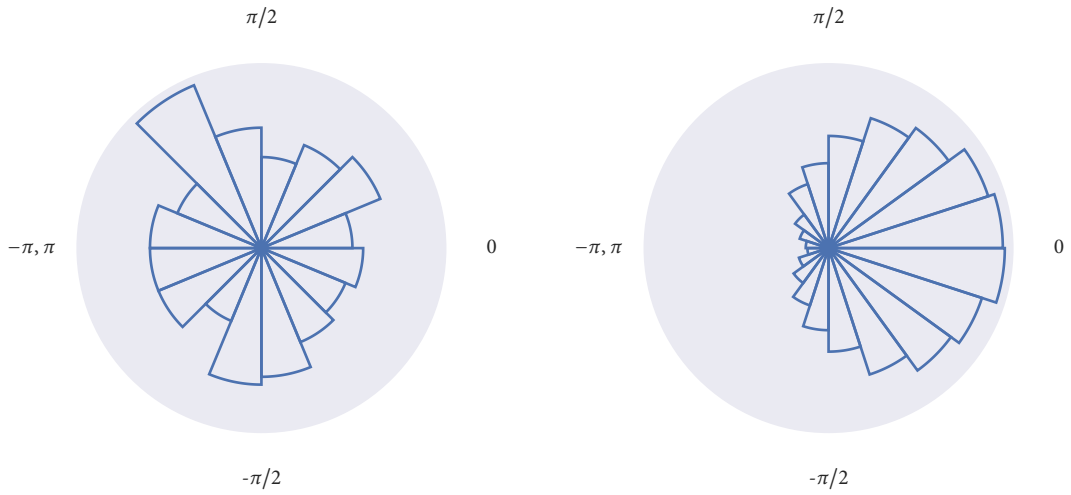
To advocate the advantages of the rose plot we shall visualise two randomly generated datasets. The first dataset consists of one hundred realisations from a uniform $U(-\pi, \pi)$ distribution, and the second dataset consists of ten thousand draws from a normal $N(0, 1)$ distribution.

In Figure A.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. In Figure A.3 we visualise the same data. Here we also get a good idea of how the directions are distributed. However, using the rose plot means we get a very intuitive representation of direction.




(a) (b)

Figure A.2: Using histograms to visualise (a): one hundred samples drawn from $U(-\pi, \pi)$ and (b): ten thousand samples drawn from $N(0, 1)$.




(a) (b)

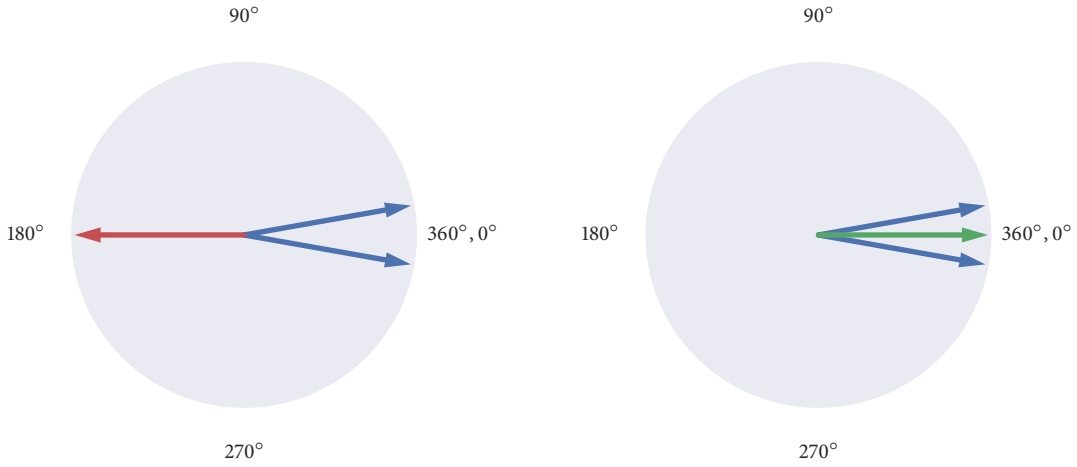
Figure A.3: Using polar histograms to visualise (a): one hundred samples drawn from $U(-\pi, \pi)$ and (b): ten thousand samples drawn from $N(0, 1)$.

## A.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 statistic to use with circular data.

Consider that we wish to take an average of the angles 10° and 350°. Using the arithmetic mean we compute an average of 180°. However, this average points in the opposite direction to which we intuitively expect. In Figure A.4(a) we visualise this result.

Before introducing the circular mean it is first necessary to introduce the atan2 function.




(a) Arithmetic mean (red).

(b) Circular mean (green).

Figure A.4: Computing the average of 10° and 350° (represented by the blue arrows), using two different mean functions. The green and red arrows show the average computed by each method.

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. Consider that 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; that is, it cannot distinguish between directions which differ by $\pi$ radians (see that $\tan^{-1}(\theta + \pi) = \tan^{-1}(\theta)$). As an example, consider calculating the direction from the $x$-axis to the ray extending from the origin to the point $(1, 1)$. 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 the origin to the point $(-1, -1)$. 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 so by adding a correction term which depends on the quadrant which contains our point of interest $(x, y)$. The correction term applied in each of the four quadrants is visualised in Figure A.5. With these considerations, atan2 can be realised by the piecewise function:

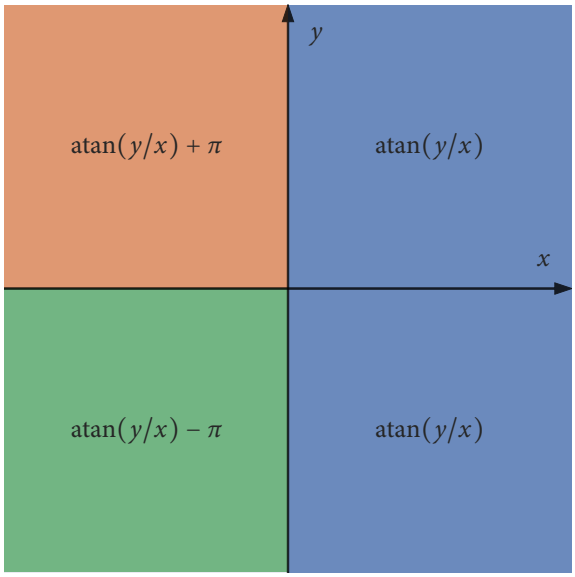



Figure A.5: An illustration of the quadrant corrections made by atan2.

$$\text{atan2}(y,x) = \begin{cases} \text{atan}(y/x) & \text{if } x > 0, \\ \text{atan}(y/x) + \pi & \text{if } x < 0 \text{ and } y \geq 0, \\ \text{atan}(y/x) - \pi & \text{if } x < 0 \text{ and } y < 0, \\ \pi/2 & \text{if } x = 0 \text{ and } y > 0, \\ -\pi/2 & \text{if } x = 0 \text{ and } y < 0, \\ \text{undefined} & \text{if } x = 0 \text{ and } y = 0. \end{cases} \tag{A.1}$$

As we saw in Figure A.4(a), averaging a set of angles with the arithmetic mean does not give the desired result. Instead, we must refer to the circular mean. Given a set of angles $\theta = (\theta_1, \ldots, \theta_n)^T$, we may compute their circular mean as:

$$\langle \theta \rangle = \text{atan2}\left( \frac{1}{n} \sum_{j=1}^{n} \sin(\theta_j), \frac{1}{n} \sum_{j=1}^{n} \cos(\theta_j) \right), \tag{A.2}$$

where the atan2 function is defined in Equation (A.1) (Fisher 1995).

The definition of the circular mean given in equation Equation (A.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.

In practice, the $1/n$ which occurs in Equation (A.2) is superfluous. Referring to Equation (A.1), see that all of the cases involving atan require the quotient of $x$ and $y$. Because of this ratio, the $1/n$ terms will always cancel, and so aren't strictly necessary.

## A.4 von Mises distribution

The von Mises distribution, sometimes simply referred to as the circular normal distribution, is a continuous probability density function defined on the circle, with support $[-\pi, \pi)$. The distribution is parameterised by two parameters: $\mu \in [-\pi, \pi)$ and $\kappa > 0$. The parameter $\mu$ is a measure of location and the parameter $\kappa$ is a measure of spread. These parameters, $\mu$ and $\kappa$, are analogous to $\mu$ and $1/\sigma^2$ of the normal distribution.

For the angle $\theta$, the von Mises distribution has probability density:

$$f(\theta \,|\, \mu, \kappa) = \frac{e^{\kappa \cos(\theta - \mu)}}{2\pi I_0(\kappa)}, \tag{A.3}$$

where the normalising constant, $I_0(\kappa)$, is the modified Bessel function of the first kind and order zero (Jammalamadaka and Sengupta 2001).

In this thesis we do *not* use the von Mises distribution. Instead, we continue to use a normal distribution to model circular data. This approximation is appropriate when $\kappa$ is large (that is, when there is little dispersion). For large $\kappa$ it is known that $I_0(\kappa) \approx e^k/\sqrt{2\pi\kappa}$. Using this, and the Taylor expansion $\cos(\alpha) \approx 1 - \alpha^2/2$, from Equation (A.3) we have:

$$\begin{aligned} f(\theta \,|\, \mu, \kappa) &\approx \frac{e^{\kappa[1 - \frac{1}{2}(\theta - \mu)^2]}}{2\pi e^k/\sqrt{2\pi\kappa}} \\ &= \frac{e^{-\frac{\kappa}{2}(\theta - \mu)^2}}{\sqrt{2\pi/\kappa}}, \end{aligned}$$

which is just the probability density of the normal distribution with mean $\mu$ and precision $\kappa$. So we see, for distributions with small dispersion, it is appropriate to approximate the von Mises distribution with a normal distribution.

# *Bibliography*


Allee, Warder C. (1931). *Animal aggregations: a study in general sociology*. University of Chicago Press.

Aoki, Ichiro (1982). 'A simulation study on the schooling mechanism in fish'. *Bulletin of the Japanese society of scientific fisheries* 48.8, 1081–1088.

Ballerini, Michele, Nicola Cabibbo, Raphael Candelier, Andrea Cavagna, Evaristo Cisbani, Irene Giardina, Vivien Lecomte, Alberto Orlandi, Giorgio Parisi, Andrea Procaccini, Massimiliano Viale and Vladimir Zdravkovic (2008). 'Interaction ruling animal collective behavior 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.

Beebe, William (1921). *Edge of the jungle*. New York: Henry Holt and Co., 291–294.

Betancourt, Michael (Jan. 2017). 'A conceptual introduction to Hamiltonian Monte Carlo'. *Arxiv e-prints*.

Budgey, Richard (1998). 'Three dimensional bird flock structure and its implications for birdstrike tolerence in aircraft'. *International bird strike proceedings committee* 24, 207–220.

Camazine, Scott, Jean-Louis Deneubourg, Nigel R. Franks, James Sneyd, Eric Bonabeau and Guy Theraula (2003). *Self-organization in biological systems*. Princeton university press.

Clark, Colin (1986). 'The evolutionary advantages of group foraging'. *Theoretical population biology* 30, 45–75.

Couzin, Iain D., Jens Krausew, Richard Jamesz, Graeme D. Ruxtony and Nigel R. Franks (2002). 'Collective memory and spatial sorting in animal groups'. *Journal of theoretical biology* 218, 1–11.

Couzin, Ian D., Jens Krause, Nigel R. Franks and Simon A. Levin (2005). 'Effective leadership and decision making in animal groups on the move'. *Nature* 433, 513–516.

Creutz, Michael (1988). 'Global Monte Carlo algorithms for many-fermion systems'. *Physical review D* 38.4, 1228–1238.

Cullen, John M., Evelyn Shaw and Howard A. Baldwin (1965). 'Methods for measuring the three-dimensional structure of fish schools'. *Animal behaviour* 13.4, 534–536.

Duane, Simon, Anthony D. Kennedy, Brian J. Pendleton and Duncan Roweth (1987). 'Hybrid Monte Carlo'. *Physics letters B* 195.2, 216–222.

Fisher, Nicholas I. (1995). *Statistical analysis of circular data*. cambridge university press.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari and Donald B. Rubin (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

Gelman, Andrew, Daniel Lee and Jiqiang Guo (2015). 'Stan: a probabilistic programming language for Bayesian inference and optimization'. *Journal of educational and behavioral statistics* 40.5, 530–543.

Giardina, Irene (2008). 'Collective behavior in animal groups: theoretical models and empirical studies'. *HFSP journal* 2.4, 205–219.

Guernon, Shay and Simon A. Levin (1993). 'Self-organization of front patterns in large wildebeest herds'. *Journal of theoretical biology* 165.4, 541–552.

Hastings, Wilfred K. (1970). 'Monte Carlo sampling methods using Markov chains and their applications'. *Biometrika* 57.1, 97–109.

Hoffman, Matthew D. and Andrew Gelman (2014). 'The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo'. *Journal of machine learning research* 15.1, 1593–1623.

Huth, Andreas and Christian Wissel (1992). 'The simulation of the movement of fish schools'. *Journal of theoretical biology* 156, 365–385.

Jammalamadaka, Sreenivasa R. and Ambar Sengupta (2001). *Topics in circular statistics*. Vol. 5. World Scientific.

Katz, Yael, Kolbjørn Tunstrøm, Christos C. Ioannou, Cristián Huepe and Iain D. Couzin (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, 18 720–18 725.

Koeppel, Dan (2002). 'Massive attack'. *Popular science* 261.6, 38–44.

Landeau, Laurie and John Terborgh (1986). 'Oddity and the 'confusion effect' in predation'. *Animal behaviour* 34.5, 1372–1380.

Long, Le V., Tsuneo Aoyama and Tadashi Inagaki (1985). 'A stereo photographic method for measuring the spatial position of fish'. *Bulletin of the Japanese society of scientific fisheries* 51.2, 183–190.

Lukeman, Ryan (2009). 'Modelling collective behaviour in animal groups: from mathematical analysis to field work'. University of British Columbia.

Lukeman, Ryan, Yue-Xian Li and Leah Edelstein-Keshet (2010). 'Inferring individual rules from collective behavior'. *Proceedings of the national academy of sciences of the United States of America* 107.28, 12 576–12 580.

Major, Peter F. and Lawrence M. Dill (1978). 'The three-dimensional structure of airborne bird flocks'. *Behavioral ecology and sociobiology* 4.2, 111–122.

Mann, Richard P. (2011). 'Bayesian inference for identifying interaction rules in moving animal groups'. *PLOS ONE* 6.8.

Mardia, Kanti V. and Peter E. Jupp (2009). *Directional statistics*. Vol. 494. John Wiley & Sons.

Mermin, Nathaniel D. and Herbert Wagner (1966). 'Absence of ferromagnetism or antiferromagnetism in one- or two-dimensional isotropic Heisenberg models'. *Physical review letters* 17.22, 1133–1136.

Metropolis, Nicholas, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller and Edward Teller (1953). 'Equation of state calculations by fast computing machines'. *Journal of chemical physics* 21.6, 1087–1092.

Metropolis, Nicholas and Stanislaw Ulam (1949). 'The Monte Carlo method'. *Journal of the American statistical association* 44.247, 335–341.

Neal, Radford M. (1995). 'Bayesian learning for neural networks'. University of Toronto.

Neal, Radford M. (2011). 'MCMC using Hamiltonian dynamics'. In: *Handbook of Markov Chain Monte Carlo*. Ed. by Steve Brooks, Andrew Gelman, Galin Jones and Xiao-Li Meng. Chapman & Hall/CRC. Chap. 5.

Okubo, Akira (1986). 'Dynamical aspects of animal grouping: swarms, schools, flocks, and herds'. *Advanced biophysics* 21, 1–94.

Organick, Elliott I. (1966). *A Fortran IV primer*. Addison-Wesley, 42.

Partridge, Brian L., Tony Pitcher, John M. Cullen and John Wilson (1980). 'The three-dimensional structure of fish schools'. *Behavioral ecology and sociobiology* 6.4, 277–288.

Pitcher, Tony J. and Julia K. Parrish (1993). *Behaviour of telost fishes*. Ed. by Tony J. Pitcher. Chapman and Hall. Chap. Functions of Shoaling Behaviour in Teleosts, 379–400.

Reynolds, Craig W. (1987). 'Flocks, herds, and schools: a distributed behavioral model'. *Computer graphics* 24.4, 25–34.

Robbins, Jim (2017). *The wonder of birds: what they tell us about ourselves, the world, and a better future*. Spiegel & Grau, 49–50.

Roberts, Gareth O. and Jefferey S. Rosenthal (2001). 'Optimal scaling for various Metropolis-Hastings algorithms'. *Statistical science* 16.4, 351–367.

Schneirla, Theodore C. (1944). *A unique case of circular milling in ants, considered in relation to trail following and the general problem of orientation*. American Museum of Natural History.

Schneirla, Theodore C. (1971). *Army ants: a study in social organization*. WH Freeman.

Selous, Edmund (1931). *Thought-transference (or what?) in birds*. Constable & Co.

Simons, Andrew M. (2004). 'Many wrongs: the advantage of group navigation'. *Trends in ecology and evolution* 19.9, 453–455.

Stan Development Team (2015). 'Stan modeling language: user's guide and reference manual'. *Version 2.12*.

Surowiecki, James (2005). *The wisdom of crowds*. Knopf Doubleday Publishing Group.

Toner, John and Yuhai Tu (1998). 'Flocks, herds, and schools: a quantitative theory of flocking'. *Physical review E* 58.4, 4828–4858.

Topaz, Chad M. and Andrea L. Bertozzi (2004). 'Swarming patterns in a two-dimensional kinematic model for biological groups'. *SIAM journal on applied mathematics* 65.1, 152–174.

Topaz, Chad M., Andrea L. Bertozzi and Mark A. Lewis (2006). 'A nonlocal continuum model for biological aggregation'. *Bulletin of mathematical biology* 68, 1601–1623.

Vicsek, Tamás, András Czirók, Eshel Ben-Jacob, Inon Cohen and Ofer Shochet (1995). 'Novel type of phase transition in a system of self-driven particles'. *Physical review letters* 75.6, 1226–1229.

Watanabe, Sumio (2009). *Algebraic geometry and statistical learning theory*. Vol. 25. Cambridge University Press.

Weimerskirch, Henri, Julien Martin, Yannick Clerquin, Peggy Alexandre and Sarka Jiraskova (2001). 'Energy saving in flight formation'. *Nature* 413, 697–698.

Wittenberger, James F. and George L. Hunt (1985). *Avian biology*. Ed. by Donald Farner. Vol. 8. Academic Press. Chap. The adaptive significance of coloniality in birds, 1–78.