# BAYESIAN INFERENCE *for* MODELS OF COLLECTIVE BEHAVIOUR

Jack Walton

December 2020

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

*to the*

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

"*Data! data! data!*" he cried impatiently.

"*I can't make bricks without clay.*"

— *Sherlock Holmes*
The Adventure of the Copper Beeches

**Acknowledgements**

This thesis wouldn't have been possible without the saintly patience of my supervisors Andrew Baggaley, Andrew Fletcher and Colin Gillespie, who were tasked with the unenviable job of keeping me on-topic. Thank you for your guidance, expertise, and endless creative flair in getting me out of faculty workshops and the dreaded "MAGIC" lecture series.

Within the department I'd like to thank my friends and office-mates for their unwavering companionship and *insatiable* desire for distraction and procrastination. I must also express my gratitude to Michael Beaty and George Stagg for helping with my (sometimes sensible) computing requests.

Of course it is only with the financial support of EPSRC that my last few years of study have been viable. So, EPSRC, if you're reading this: thank you.

Lastly, dear reader, thank you for your keen eye for typographical detail, and for looking at the pictures, but not reading the text too closely…

**Abstract**

The study of collective behaviour—broadly defined as the formation of macro-level structures from the interactions between individuals—has in recent years become a thriving topic of multi-disciplinary research. Under consideration of biological-fitness the scientist has been able to reason about *why* these structures form, and the advantages which the collective can afford the individual. However, much less is known about *how* these structures are formed and maintained in the first place.

Much work has been invested in the development of mathematical models which seek to explain the formation and maintenance of animal aggregations. Research has shown that behaviour reminiscent of real flocking events can arise from simple mathematical models which describe how individuals interact with one another. However, much of this modelling relies on aprioristic assumptions about how individuals behave and interact, with little-to-no verification against real observation.

In this work we examine mathematical models popular in the literature and suggest modifications motivated by considerations of biological-realism. In particular we advocate adoption of continuous interaction rules, and consider how behavioural and biological variation can be accounted for by imposing hierarchical structure.

We proceed to fit these models of collective behaviour to observations of real and simulated flocking events. Model fitting is performed in a Bayesian framework, allowing the quantification of parameter uncertainty. Fitting models to simulated data provides opportunity to assess the effectiveness and accuracy of our inference schemes, before attempting the same inference on real observation. Multiple competing models are fit to the same data, with the predictive performance of these models ranked using ideas from the model-selection literature. We are then able to recommend a subset of the candidate models as providing the best performance.

Finally, consideration is made for datasets which exhibit missing observations. Such missingness occurs naturally owing to the fixed-location recording equipment used to record flocking events. We argue that this missingness *cannot* be ignored, and must be accounted for during any model-fitting process. Techniques are outlined which allow the researcher to account for missingness. Simulation studies are performed which demonstrate the efficacy of the outlined approach, before the techniques are demonstrated on a real dataset exhibiting missingness.

# *Contents*

# 1

# *Introduction*

## 1.1 Motivation

It is perfectly possible to live our everyday lives without giving a second thought to the world around us. In fact, it can be rather easy to do so. Nature becomes something secondary: an inconvenience of the past, something tamed by modern amenity. But every once in a while comes an event which *demands* our attention. These displays of nature—the great ones—have the power to invoke awe and wonder, offering a renewed appreciation of the natural world, and our place within it.

One of these great displays of nature is represented by the collective motion of animals. The canonical example of collective motion being the starling murmuration (Figure 1.1). In these events starlings gather in huge numbers and perform the most mesmerising of ballets; the entire flock moving as if some fluid object. Coordinated, yet unpredictable, no two such murmurations are ever the same.

Another striking example of collective behaviour is given by the dense milling structures sometimes formed by shoals of fish (Figure 1.2). In these structures the individuals form something greater than the sum of their parts: their shoal taking on a life of its own. With a flash of silver, and in the blink of an eye, the shoal can change direction, offering an effective defence against predation.

Collective motion, broadly defined as the formation of macro-level structures from the interactions between individuals (Camazine et al. 2003), has been observed over many different length scales, and has been exhibited by many different species (Allee 1931). But

**Figure 1.1:** A group of starlings, here consisting of *thousands* of individual members, gather to flock in the evening sky. These flocking events (often called murmurations), represent one of the most eye-catching displays of collective behaviour. This particular event was captured near Gretna, in the Scottish Borders. Photograph: Owen Humphreys/PA.

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 many years, collective behaviour has become a thriving topic of multidisciplinary research, holding captive the minds of physicists, biologists, mathematicians, and statisticians. 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 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 migration can benefit from the knowledge of the collective (Simons 2004).

Despite this, much less is known about the *how* of collective behaviour. The mechanisms which lead to the formation and maintenance of aggregations remain elusive, and are a topic of much continued interest. The hope is that the study of mathematical models of flocking, and the comparison of these models with real flocking events, will lead to insights about the mechanics which underlie these phenomena. In recent years modern computing has made the process of simulating such models (relatively) pain-free. However, a lack of data

**Figure 1.2:** Mackerel forming a tight milling structure in open water. The individuals behave as one entity, and can swell, change direction, deform and reform in fractions of a second. These structures form spontaneously in the presence of danger, and act as a defence against predation.

describing 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 data and theory has been lacking. The lack of quantitative comparison between model and data can largely be attributed to the scarcity of appropriate empirical data.

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

This thesis considers and develops theoretical models of collective behaviour popular in the literature. This development is guided by considerations of biological realism. Simulation studies are performed which show that the developed models can be fit to simulated data, after which the same models are fit to real data. The efficacy of each fitted model is compared, and a subclass of the candidate models are presented as offering the best fit.

## 1.2 Thesis overview

In Chapter 2 the reader is given a review of the literature surrounding the study of collective behaviour. Important ideas and results from the field are introduced, summarised and discussed. After relaying the main results from the literature, open problems and the future of research in the field are discussed.

Bayesian statistics is introduced to the reader in Chapter 3. The underlying philosophy of the Bayesian paradigm, as well as important results, techniques and algorithms are outlined. Consideration is given for common problems that the Bayesian practitioner may encounter, as well as how one might address these problems.

In the proceeding chapter, Chapter 4, a class of model popular in the literature of collective behaviour is introduced. The shortcomings of this model are then considered and summarised, with alterations proposed to address the outlined concerns. The proposed changes are motivated by considerations of biological-realism.

Following on from this, Chapter 5 details the results of a number of so-called simulation studies. In these studies a model is forward simulated for known parameter values. A practitioner is then tasked with capturing these known parameter values with statistical inference. A number of simulation studies are performed on variations of a previously considered model.

The statistical machinery built in Chapter 5 is then repurposed in Chapter 6, as our developed models are fit to real observation. Fitting a number of different models to the same data, the predictive performance of these models are compared, and a subclass of these models are identified as providing the best fit.

Chapter 7 considers the problem of missing data. It is argued that the nature of flocking events mean that missing data becomes an inevitable issue. Algorithms are outlined which allow the scientist to account for additional uncertainty introduced by missingness. Simulation studies are performed which demonstrate the efficacy of this approach. Following this a dataset of a real flocking event with missing observations is introduced. The limitations of previous studies of this data are outlined. The analysis presented in this thesis addresses the shortcomings of this earlier work; accounting for the additional uncertainty introduced by the missing observations.

For the uninitiated, a review of circular statistics is presented in Appendix A. This review motivates the study of circular statistics, and argues its relevance to the scientist involved in the study of collective behaviour.

# 2

# *Literature review*

There exists 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 of collective behaviour.

In this chapter we shall discuss some of the most important ideas and results from the literature surrounding the study of collective behaviour. First, we provide an overview and discussion 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. Finally we review previous work which focused on recording and utilising empirical data to assess predictive performance of theoretical models.

## 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 (the confusion effect) (Landeau and Terborgh 1986). In addition to this confusion effect, groups of individuals can absorb and process 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 catastrophic 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 an individual in front of it, causing the separated workers to run in a densely packed circle until they eventually succumb to 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). It would take an individual ant 2.5 hours to circumnavigate a mill of this size (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 the individuals involved. However, we have still yet to broach the *how* of collective behaviour.

## 2.2 Mathematical approaches

There has long existed an appeal of using mathematical models as a tool to investigate collective behaviour (Aoki 1982; Okubo 1986; Reynolds 1987; Huth and Wissel 1992; Guernon and Levin 1993; Vicsek, Czirók et al. 1995; Couzin, Krause, James et al. 2002). These models of collective behaviour are broadly partitioned into two paradigms: the Lagrangian and Eulerian approaches. 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 analogous models 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 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 neighbours (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 (correspondingly: 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 (Huth and Wissel 1992; Vicsek, Czirók et al. 1995; Couzin, Krause, James et al. 2002; Couzin, Krause, Franks et al. 2005). This model also accounted for fish having incomplete fields of vision. The simulation of some unobserved region 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 three-dimensional space. To produce more aesthetically pleasing animations, the software, "Boids", implemented additional sophistications such as banking during turns. The focus on developing simulations to produce visually-pleasing behaviours rendered rigorous scientific study unbefitting. Interestingly, Tim Burton's 1992 Batman Returns used a modified version of the Boids software to simulate animations of bat swarms and marching penguins.

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

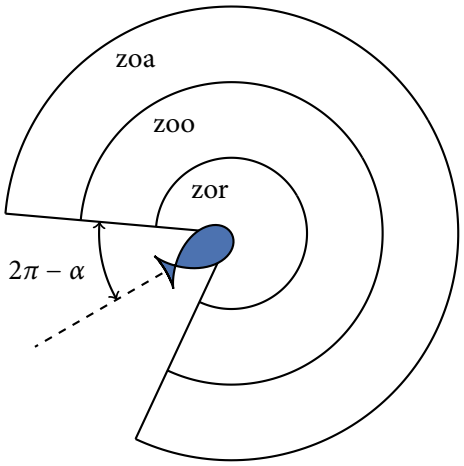



**Figure 2.1:** An illustration, based on the work of Aoki (1982), showing the area around an agent (here, a fish) partitioned into three zones; zor: zone of repulsion; zoo: zone of orientation (or alignment), and zoa: zone of attraction. The missing segment behind the individual represents a blind spot, of angle $2\alpha$, into which it cannot see.

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 by research within statistical physics, Vicsek, Czirók 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 some interaction radius. This model is commonly referred to as the "Vicsek Model". Despite its simplicity this model produces complex behaviour resembling that of a real biological system. Vicsek, Czirók et al. (1995) investigated the phase transition between ordered and disordered motion as the density of particles and noise in the system was varied. The 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 converges upon some arbitrary direction of travel. 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, Mermin-Wagner only applies to systems in equilibrium and the Vicsek model is out of equilibrium.


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

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), Couzin, Krause, James 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, suggesting that previous group structure influences future behaviour as interactions change.

Further research was made by Couzin, Krause, Franks et al. (2005) which investigated how leaders influence the motion of travelling groups. A zonal repulsion-alignment-attraction 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.

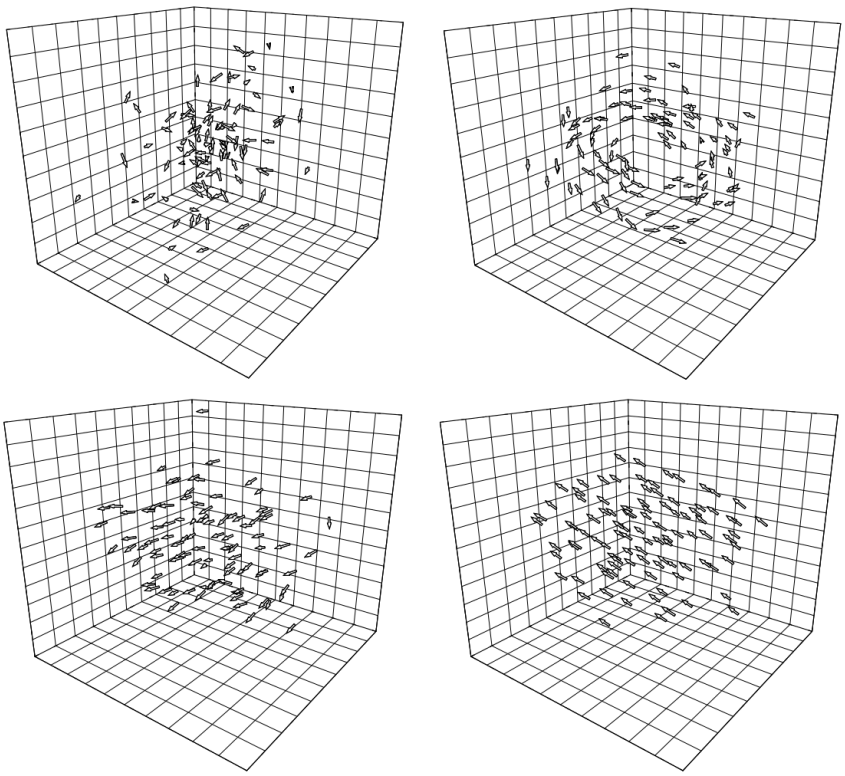


**Figure 2.3:** Taken from Couzin, Krause, James et al. (2002), different steady-state solutions (swarm, torus, dynamically parallel and highly parallel) obtained by making small changes to model parameters of a three-dimensional flocking model.

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 approach 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 group properties develop 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

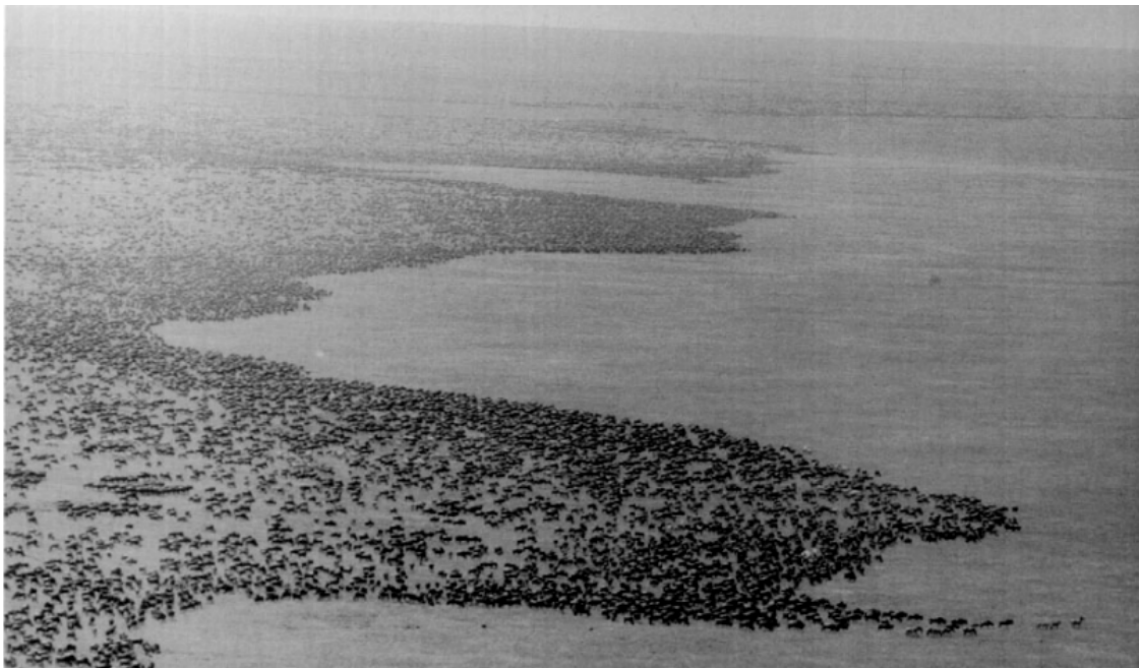


**Figure 2.4:** Aerial photography of migrating wildebeest showing large-scale front patterns, as presented in the work of Guernon and Levin (1993).

large groups of wildebeest. Their model predictions were compared with aerial observations of wildebeest migrating through 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 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 renormalisation 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 and stationary clump solutions (Topaz and Bertozzi 2004; 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). Additionally, these models require more involved computer implementation than their Lagrangian counterparts necessitate. 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. It is understood that the emergence of a biologically realistic pattern from model simulation *is not* sufficient evidence of model correctness. That is, the emergence of a desired pattern is not sufficient evidence that a model is correctly capturing the interactions between individuals. This observation is further compounded by the understanding that models implementing different local interactions can produce similar looking behaviour at the group level; see, for instance, the similar behaviours (swarms, undirected mills and moving aligned groups) exhibited by the zonal repulsion-alignment-attraction model of Couzin, Krause, James et al. (2002) and the attraction-blind-angle model of Strömbom (2011).

As such, real data describing the dynamics of animal aggregations is *essential* to assess the validity and efficacy of theoretical models and the assumptions they make. With such data it becomes possible to compare and rank the predictive performance of competing models.

Thorough comparison between model and data 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 measuring equipment, not to mention the additional difficulty imposed by 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 behaviour 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 between 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. With these considerations, further research also concentrated on fish (Partridge et al. 1980; Long, Aoyama and Inagaki 1985).

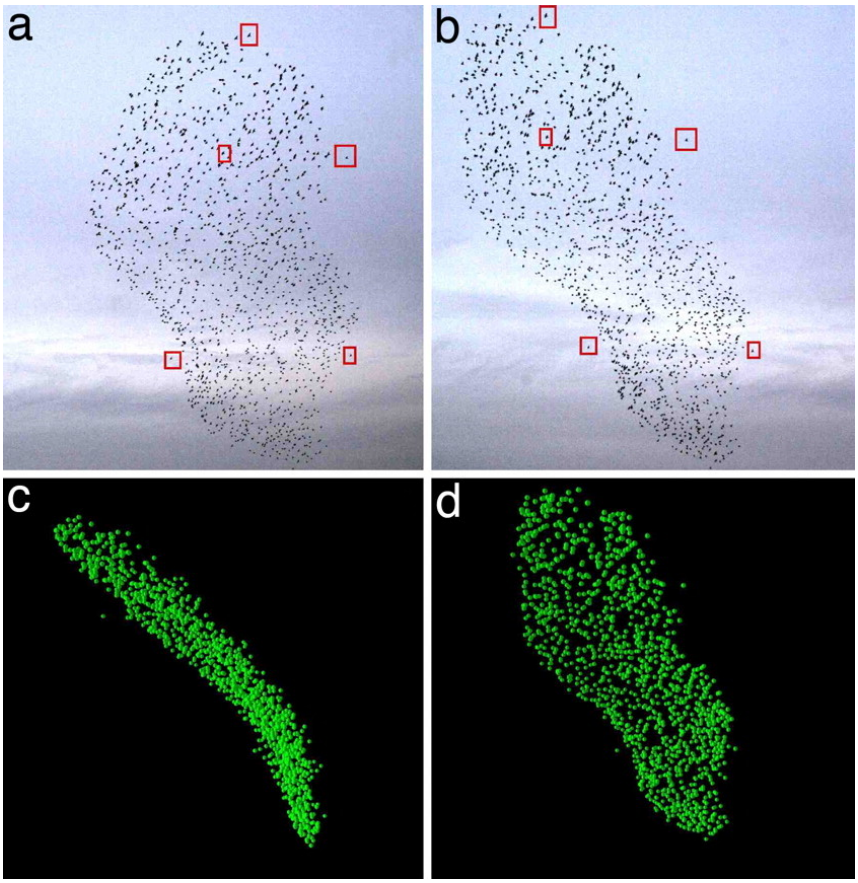



**Figure 2.5:** A flock of 1246 starlings reconstructed in three-dimensions. Photographs taken at the same instant but 25 m apart (a–b) are used to reconstruct their three-dimensional positions (c–d). To perform reconstruction Ballerini et al. (2008) needed to match each bird in (a) to its corresponding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

Having collected empirical data, these studies investigate properties such as the distance of individuals to their nearest neighbour, or the direction from an individual 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 individuals (Figure 2.5). To collect this data the authors used a combination of stereometric and computer vision techniques. Having extracted the data, 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

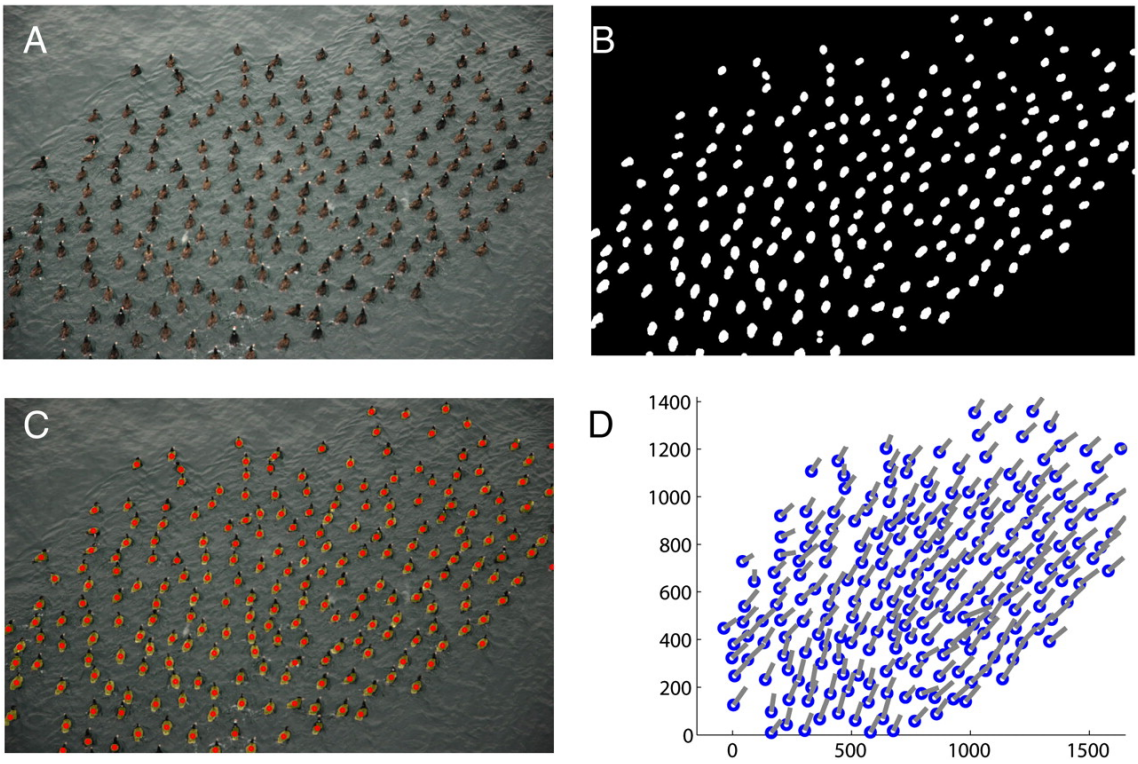



**Figure 2.6:** An image of field data and visualisations of the transformations used to extract positions of individuals in a flock of foraging birds (Lukeman, Li and Edelstein-Keshet 2010).

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), who collected and analysed data of large flocks of 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 a factor of ten the number of individuals which could be reliably tracked though time (Lukeman 2009). Plots of nearest neighbour densities were constructed from the extracted dataset. 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. Model parameters were fitted which could best reproduce angular and radial neighbour distributions. It was concluded that a zonal repulsion-alignment-attraction model with an additional frontal interaction was best able to reproduce the desired 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 individual fish through time. Working in a classical mechanics framework the authors considered the effective forces acting on a focal fish as a function of position and velocity. The authors found that the dominant interaction between fish was the regulation of their speed. No evidence was found of an explicit alignment interaction 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.

With many different models capable of producing realistic-looking flocks, researchers have sought to use ideas from the model-selection literature to quantify model performance. Mann et al. (2013) fit a number of models—including: a topological model; a non-local model in which all individuals interact, and a variety of different spatial models—to data describing the collective motion of glass prawns confined to an annular arena. To quantify the performance of each candidate model the authors computed the marginal likelihood. With this the authors were able to compare model performance and conclude that a spatial model with a persistent "memory" effect provided the best fit to their data. Following the methodology outlined by Mann et al. (2013), Strandburg-Peshkin et al. (2013) used the marginal likelihood to compare the performance of models with metric, topological, Voronoi, and visual interaction rules, to explain the movements of shoals of fish. In doing so the authors were able to conclude that Voronoi and visual models were better at explaining their data than metric and topological models.

The marginal likelihood provides an appealing measure with which to compare model fits: penalising models for complexity, and hence favouring simpler models over complex models. Although the marginal likelihood is often used to compare model performance, under certain circumstances the marginal likelihood is understood to be highly sensitive to prior dispersion, even when the posterior is robust against prior dispersion (Fong and Holmes 2020). As such, estimating the marginal likelihood with information criteria (Section 3.3.1) remains a popular alternative approach to model comparison.

Analysis of empirical data has so far focused on epiphenomena such as nearest neighbour distances or angular neighbour densities. Research has then focused on fitting models which are best able to replicate these properties, rather than movements of individuals themselves. With technological and methodological 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 more numerous empirical studies, parameters were inferred on their ability to explain the *movements* 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 a 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.

Although not utilised frequently in the literature, such simulation studies represent a useful aid in developing the statistical machinery to fit models of collective behaviour to real data.

## Conclusions

Although our understanding of collective behaviour has developed considerably from early speculations of telepathy, there still remain *many* unknowns. We saw that the notion of biological fitness goes a long way to explaining the *why* of collective behaviour. However, we also reflected on how much is unknown about the *how* of collective behaviour.

Much of the literature was seen to have utilised mathematical models in an attempt to understand the mechanics underlying the formation and maintenance of flocks. Two popular modelling paradigms were introduced: Eulerian, and Lagrangian. Considering the strengths and weaknesses of these approaches, we concluded that the Lagrangian framework represents a more intuitive and appealing paradigm for our study of collective behaviour.

Having considered a number of Lagrangian models, we saw that a variety of different models were able to produce visually similar flocking events. We argued that comparison between model and data is essential to assess the performance and realism of theoretical models. A review of the literature taking measure of real flocking events was detailed. Previous work has been limited by the availability of data of flocking events. When data of real events *has* been utilised, the focus has been on fitting models to best reproduce properties such as nearest-neighbour densities. We argue that the fitting process should instead centre on explaining the *movements* of the flock, rather than some epiphenomena of their movements.

In this thesis we seek to address the shortcomings of previous work. In particular, we seek to fit theoretical models of collective behaviour to observations of real events. In contrast with previous work, this fitting process will seek to explain the *movements* of flocks, rather than some statistical properties of the flocks. Parameter uncertainty will be quantified in all subsequent analyses. In addition to this, multiple competing models will be fit to the same observations, allowing us to perform a quantitative comparison of predictive performance, and assess aprioristic modelling assumptions.

# 3

# *Bayesian statistics*

In this thesis we utilise methods from Bayesian inference to fit mathematical models of collective behaviour to real and simulated data. Bayesian inference represents a fully probabilistic approach to parameter inference. This allows a practitioner to quantify their uncertainties about inferred parameter values. 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 outline some important concepts of Bayesian inference. Algorithms to infer model parameters from data will be presented, and the interpretation of such output shall be discussed. We will also make brief study of some ideas from the literature of model selection. The study of model selection allows a practitioner to make quantitative comparison between competing models.

## 3.1 Bayesian inference

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

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

The likelihood represents the probability density of observing data $x$, given the model parameters $\theta$. The prior distribution $\pi(\theta)$ is used to quantify our prior knowledge about the parameters. Bayes' Theorem provides a methodology to realise posterior distributions from a combination of our prior beliefs and the likelihood of the data:

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

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 the posterior distribution as proportional to the product of the likelihood and our prior beliefs:

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

For reasons of numerical stability it is common practice to target the logarithm of the posterior distribution, rather than posterior itself. With this we may express the log-posterior as equal to the *sum* of our log-prior and log-likelihood, plus some additive constant.

## 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 (MCMC) algorithms represent a class of algorithm which provide methods to sample from the posterior density $\pi(\theta \,|\, x)$, whilst avoiding the evaluation of the troublesome normalising constant.

Unfortunately, although MCMC algorithms circumvent evaluation of the normalising constant, the consecutive samples realised by MCMC are *not* independent of one another. Instead, as the samples are constructed to form a Markov chain, each sample is related to the samples which occur before it.

### 3.2.1 *Gibbs sampling*

The Gibbs sampler, named after the physicist Josiah Willard Gibbs, was first outlined by S. Geman and D. Geman (1984). This algorithm allows the Bayesian practitioner to sample from a multivariate density via full conditional distributions of parameters. The Gibbs algorithm is useful when these conditional densities can be expressed in a standard form

which is easy to sample form. The Gibbs sampler is a special case of the Metropolis–Hastings algorithm (introduced in Section 3.2.2), and is outlined in Algorithm 1.

**Algorithm 1:** Pseudo-code demonstrating $n$ iterations of a Gibbs sampler used to target $\pi(\theta \mid x)$, for $\theta = (\theta_1, \ldots, \theta_d)^T$.

1 Initialise chain with $\theta^{(0)}$.
2 **for** $i = 1$ **to** $n$ **do**
3  **for** $j = 1$ **to** $d$ **do**
4  $\theta_j^{(i)} \sim \pi\left(\theta_j^{(i)} \mid \{\theta_k^{(i)}\}_{k<j}, \{\theta_k^{(i-1)}\}_{k>j}, x\right)$
5  **end for**
6 **end for**

The Gibbs sampler is a popular MCMC algorithm used in Bayesian inference problems, partly owing to its ease of implementation. However, Gibbs limits us to posteriors for which we can express the full conditional distributions of all the parameters of interest. Other algorithms, such as the more general Metropolis–Hastings algorithm, allow us to target the posterior *without* necessitating that the full conditional distributions be known.

### 3.2.2 *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 (typically the posterior).

The algorithm (described formally in Algorithm 2) begins by initialising a Markov chain with parameters $\theta^{(0)}$. After this the algorithm proposes new parameter values $\theta^\star$ from a proposal distribution $q(\theta^\star \mid \theta^{(0)})$. The proposed values are accepted with probability $\alpha(\theta^\star \mid \theta^{(0)})$. The acceptance probability depends on a ratio of the posterior density evaluated at the current state, $\theta^{(0)}$, and the posterior density evaluated at the proposed state, $\theta^\star$ (line 4 of Algorithm 2). In this ratio the normalising constants from Equation (3.2) cancel, and we see that the target distribution need only be known up to a constant of proportionality—effectively circumventing the need to ever compute the normalising constant. If the proposed state is accepted then the next state of the Markov chain is set to the proposed value; otherwise, the next state is set to the current value (lines 5–12 in Algorithm 2). The process of proposing and accepting or rejecting proposals continues for a predefined number of iterations (in Algorithm 2, $n$ iterations), and ideally until the sampler "forgets" about its starting location, and makes a satisfactory number of draws from the

**Algorithm 2:** Pseudo-code demonstrating an implementation of the Metropolis–Hastings algorithm to target $\pi(\theta \mid x)$ with $n$ iterations.

Initialise chain with $\theta^{(0)}$
**for** $i = 1$ **to** $n$ **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**

posterior (more on this in Section 3.2.4). To allow the chain to "forget" about its starting position, and to converge to the posterior distribution, it is common to allow some burn-in period (sometimes called the warm-up period). This period is intended to allow the chain to converge. Samples drawn during this period are discarded.

### *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 $\pi(\theta \mid x)$. However, it is not obvious how to choose a proposal distribution which will realise these desires.

A special case of Metropolis–Hastings arises when the proposal distribution is symmetric. Symmetric proposals have the property that:

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

This symmetry results in a simplification of the acceptance ratio:

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

The random walk sampler is a popular implementation of Metropolis–Hastings which makes use of symmetric proposals. Here, proposals are generated as

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

where the $\omega$ are sampled as

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

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 a chain moves around the sample space, and how long it takes for the chain to converge to the target distribution (see Figure 3.1 for a visualisation of well-mixing and poorly-mixing chains). Crucially then, the parameter $\Sigma$ can be used to control the mixing of chains. So, naturally, we desire to select a $\Sigma$ which produces well-behaved chains. 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 an optimum acceptance probability, a common technique is to choose

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

where $\widehat{\mathrm{Var}}(\theta \mid x)$ represents an estimate of the posterior variance made by a sample run.

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).

### 3.2.3 *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 a very 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:

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

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:

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

As a consequence of using information about the *gradient* of the posterior distribution, HMC requires that the posterior distribution be differentiable almost everywhere. Because of this, HMC has the shortcoming that it *cannot* be used to target distributions with discrete valued parameters; instead requiring a continuous-parameter valued posterior.

### *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 3 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 the number of steps $L$, the step size $\epsilon$, and the mass matrix $M$. Unfortunately, making a poor choice for any of these parameters can result in a significant decrease in the performance of HMC (Hoffman and Gelman 2014). Fortunately, $\epsilon$ can be tuned during the algorithm's implementation, using ideas from 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 ideas from stochastic optimisation in numerical analysis, the authors were able to tune $\epsilon$ and $M$ during the implementation of their algorithm. NUTS performs *at least* as efficiently as, and often *more* efficiently than, an equivalent well-tuned implementation of a standard HMC method. Altogether then, NUTS represents an efficient implementation of a HMC-like algorithm, where the practitioner is left free the obligation of choosing 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 programme, specifying how to compute the

**Algorithm 3:** Pseudo-code demonstrating an implementation of Hamiltonian Monte Carlo to target $\pi(\theta \mid x)$ with $n$ iterations. The Leapfrog method is used to approximate solutions to Hamilton's equations with discretisation $\epsilon$, and trajectory length $L$.

Initialise chain with $\theta^{(0)}$
**for** $i = 1$ **to** $n$ **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

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

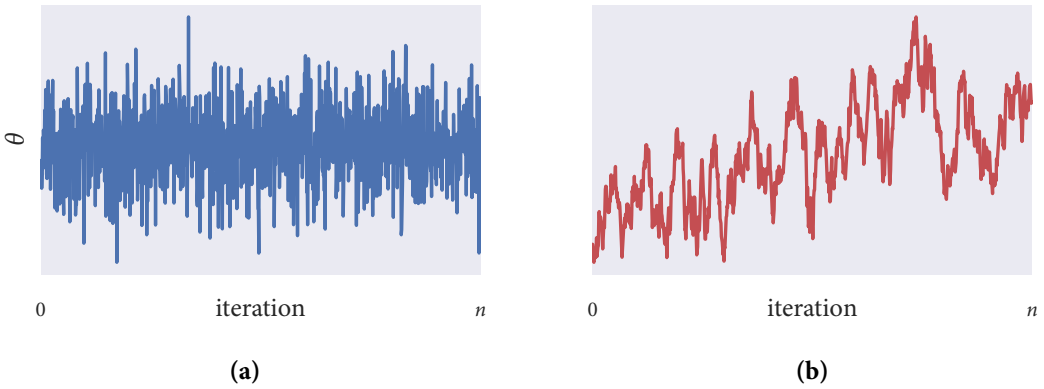
  Draw $u \sim \text{Uniform}(0, 1)$
  **if** $u \leq \alpha(\theta^\star, p^\star \mid \theta^{(i-1)}, p^{(i-1)})$ **then**
    *# Accept proposal*
    $\theta^{(i)} \leftarrow \theta^\star$
  **else**
    *# Reject proposal*
    $\theta^{(i)} \leftarrow \theta^{(i-1)}$
  **end if**
**end for**

**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** $\theta$, $p$

posterior density for the targeted model (Stan Development Team 2015). With this, Stan implements the NUTS algorithm, and returns the user samples drawn from the posterior.

### 3.2.4 *Convergence diagnostics*

Having implemented our sampler to make draws from the posterior distribution, we are now in a position to inspect our output. However, before we can begin to consider assessing measures such as the predictive performance of our fitted models, we must first assess our sampler's output to check for convergence.




**(a)** **(b)**

**Figure 3.1:** Example Markov chains produced by a sampler. (a) A chain which looks to be mixing well, and shows signs of convergence: the chain oscillating in a regular fashion around a fixed location. The chain displays no obvious irregularities, and successive samples do not appear highly correlated. (b) A poorly mixing chain. Realisations are observed to be highly correlated with successive samples. This chain is inefficiently exploring the sample space, and could benefit from a proposal mechanism allowing larger innovations.

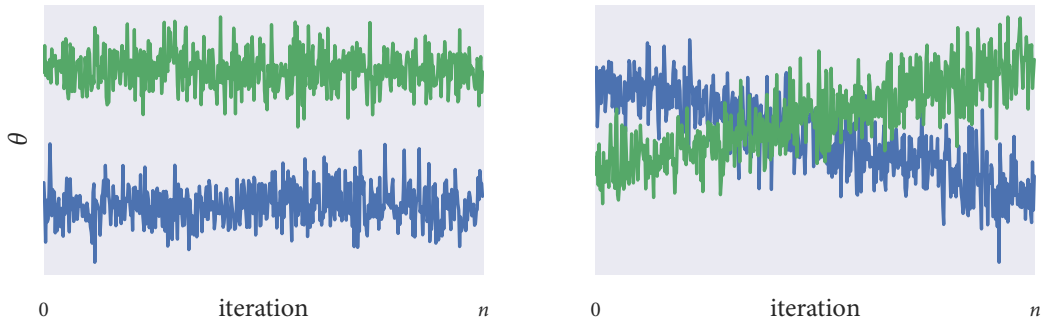
### *Graphical diagnostics*

An informal, yet simple and effective method to assess convergence is to inspect trace plots of the Markov chains produced by our samplers. When we do so we check for any irregularities in the chains. In Figure 3.1 we visualise Markov chains generated by two samplers which were implemented for $n$ iterations, targeting some parameter $\theta$. Figure 3.1(a) shows a chain which is mixing well and looks to have converged to the posterior distribution, as it moves around some fixed region with constant variance. However, the chain displayed in Figure 3.1(b) does *not* look to have converged; consecutive samples look to be strongly autocorrelated, and the iterations do not appear to be oscillating around any common distribution.

For a more robust inspection of convergence it is advisable to run multiple chains initialised at different locations (Figure 3.2). Stan makes the running of multiple chains initialised at different locations a trivial alteration for the user, and allows the user to run each separate chain in parallel.

### *Split-$\widehat{R}$*

Split-$\widehat{R}$ represents a popular convergence diagnostic (Gelman, Carlin et al. 2013). This measure uses information about the within-chain variance *and* between-chain variance to quantify how well chains have mixed.




**(a)** **(b)**

**Figure 3.2:** The importance of running multiple chains. (a) If only a single chain had been inspected, here we may have concluded that the corresponding sampler had converged. However, in running multiple chains it is clear that the chains have *not* converged to a common distribution. (b) Likewise, viewed separately, the trajectories here do not appear stationary, but taken together they do appear to cover a common distribution.

Consider that we have $m$ chains of length $n$, which target some estimand of interest $\theta$. We label the draws as $\theta^{(i,j)}$ for $i = 1, \ldots, n$; $j = 1, \ldots, m$. The between-chain variance $B$ is then computed as:

$$B = \frac{n}{m-1} \sum_{j=1}^{m} (\bar{\theta}^{(\cdot,j)} - \bar{\theta}^{(\cdot,\cdot)})^2 \quad \text{where} \quad \bar{\theta}^{(\cdot,j)} = \frac{1}{n} \sum_{i=1}^{n} \theta^{(i,j)}, \quad \bar{\theta}^{(\cdot,\cdot)} = \frac{1}{m} \sum_{j=1}^{m} \bar{\theta}^{(\cdot,j)}.$$

The within-chain variance $W$ is captured in a similar manner:

$$W = \frac{1}{m} \sum_{j=1}^{m} s_j^2, \quad \text{and} \quad s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\theta^{(i,j)} - \bar{\theta}^{(\cdot,j)})^2,$$

The between-chain and within-chain variance can be combined in a weighted sum to estimate the marginal posterior variance of the estimand:

$$\widehat{\mathrm{Var}}^{+}(\theta \,|\, x) = \frac{n-1}{n} W + \frac{1}{n} B. \tag{3.9}$$

The convergence diagnostic split-$\widehat{R}$ is then computed as:

$$\widehat{R} = \sqrt{\frac{\widehat{\mathrm{Var}}^{+}(\theta \,|\, x)}{W}}. \tag{3.10}$$

From Equation (3.9) observe that as $n \to \infty$ the estimated marginal posterior variance tends to $W$. With this, and Equation (3.10), we see that $\widehat{R} \to 1$ as $n \to \infty$. As such, we look to

compute $\widehat{R} \approx 1$ to indicate that a sampler has converged.

### *Effective sample size (ESS)*

The effective sample size (ESS) is introduced as a notion to assess the "size" of a sample when the samples are correlated. The idea is that 10,000 *autocorrelated* samples from a distribution encode less information than 10,000 *independent* samples from that same distribution. The effective sample size is computed as the number of samples drawn, minus some penalty whose size depends on some measure of autocorrelation.

For a single chain, a common definition of the effective sample size is:

$$\text{ESS} = \frac{n}{1 + 2\sum_{k=1}^{\infty} \rho_k}, \tag{3.11}$$

where $n$ is the number of samples and $\rho_k$ is the autocorrelation at lag $k$. Although a popular definition, in practice this measure is often too optimistic when chains haven't converged (Gelman, Carlin et al. 2013). The practitioner must also decide when to truncate the infinite sum which appears in the denominator.

A more sophisticated estimate of the effective sample size is presented by Gelman, Carlin et al. (2013). In this definition the authors circumvent the infinite sum in the denominator of Equation (3.11). The inclusion of between-chain variance also makes this measure more robust.

In estimating ESS, the authors first estimate the sum of the correlations $\rho$. To compute the correlations it is first necessary to compute the variogram at each lag $t$:

$$V_t = \frac{1}{m(n-t)} \sum_{j=1}^{m} \sum_{i=t+1}^{n} (\theta^{(i,j)} - \theta^{(i-t,j)})^2.$$

Using the variogram computed at each lag $t$, along with the estimate of the marginal posterior variance given in Equation (3.9), the correlations can be estimated as:

$$\widehat{\rho}_t = 1 - \frac{V_t}{2\widehat{\text{Var}}^{+}}.$$

The infinite sum of Equation (3.11) is replaced by a partial sum. Here, correlations are added until the sum of two successive lags $\widehat{\rho}_{2t'} + \widehat{\rho}_{2t'+1}$ is negative. This gives the estimate:

$$\text{ESS} = \frac{mn}{1 + 2\sum_{t=1}^{T} \widehat{\rho}_t},$$

where $T$ is the first odd positive integer for which $\widehat{\rho}_{T+1} + \widehat{\rho}_{T+2}$ is negative.

## 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 dataset 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. Gelman, Carlin et al. (2013) find WAIC more appealing than AIC and DIC as WAIC has the more desirable property of averaging over the entire posterior distribution.

However, to use WAIC a practitioner must partition their data into independent (or close to independent) groups. It has been remarked that this can be very difficult to achieve for time-series data, or data with a spatial component. Since models of collective behaviour have both a time component *and* a spatial component, WAIC does not seem a viable information criterion for our use case.

Instead of using WAIC we will opt to use AIC. AIC remains a popular choice of information criterion as it is easy to compute from MCMC samples. The Akaike information criterion penalises model complexity and so is designed to provide a measure which trades-off between model-fit and model-complexity. AIC is defined as:

$$\mathrm{AIC} = 2p_{\mathrm{AIC}} - \log\big(\widehat{L}\big), \tag{3.12}$$

where $p_{\mathrm{AIC}}$ is the number of model parameters, and $\widehat{L}$ is the maximised likelihood.

## Conclusions

The Bayesian paradigm represents a realisation of the scientific process of updating beliefs in light of observation. With this, Bayesian inference presents itself as a very natural approach to statistics.

As reflected in Section 2.4, an important aspect of model fitting is capturing *uncertainty* in estimated parameters. This is particularly prudent when the data used to form opinion is prone to exhibiting high levels of noise. As Bayesian inference represents a fully-probabilistic approach to parameter estimation, its application to the study of collective behaviour—a subject of datasets susceptible to large amounts of noise—seems pragmatic.

The algorithms outlined in Sections 3.2.2 and 3.2.3 provide the Bayesian practitioner the methodology to make realisations from the posterior distribution. These algorithms will prove instrumental to this thesis.

The study of collective behaviour is awash with models formed on aprioristic assumption, with little-to-no verification against observation. To compare these models we will use ideas from the model selection literature; as we seek to trade-off a model's ability to describe data, with its complexity.

# 4

# *Model development*

In this chapter we shall introduce one of the most popular and well-studied agent-based models in the literature: the Vicsek model. The Vicsek model represents a simple alignment model in which agents interact with neighbours within some fixed distance (Vicsek, Czirók et al. 1995). Despite its simplicity, this model can produce sophisticated dynamics reminiscent of real flocking events (Ginelli 2016). A phase transition from order to disorder is observed as the amount of noise in the model is regulated (Vicsek and Zafeiris 2012).

The Vicsek model, like many other agent-based models (ABMs) (Aoki 1982; Huth and Wissel 1992; Couzin, Krause, James et al. 2002), implements a so-called "zonal" interaction rule. These rules represent discontinuous interactions. With this, the onset of interaction between individuals is very sensitive to small perturbations in distances. We consider continuous interaction rules as providing a biologically-motivated alternative to zonal interaction rules. Such rules ensure interactions are more robust to small perturbations in distances, but without the penalty of introducing additional model complexity. We believe that the introduction of continuous interaction rules represents a step towards biological realism.

## 4.1 The Vicsek model

The Vicsek model simulates the movements of $N$ individuals over $T$ time steps (Figure 4.1). Individuals travel with constant speed $v$, within a square-cell with periodic boundary conditions and side length $L$. To initialise a simulation agents are allocated a random position

within the cell, and a random direction of motion. From time $t$ to time $t+1$ ($\Delta t = 1$) agent $i$ updates its position as:

$$\boldsymbol{x}_{i,t+1} = \boldsymbol{x}_{i,t} + \boldsymbol{v}_{i,t}\Delta t, \tag{4.1}$$

where the velocity $\boldsymbol{v}_{i,t}$ is constructed to have speed $v$ and direction of motion $\theta_{i,t+1}$. Updating $\theta_{i,t}$ to $\theta_{i,t+1}$ models how agents update their directions of motion in light of their neighbours' movements. The ability of individuals to observe and react to the movements of their neighbours is assumed to be imperfect, and so a noise term, in the form of a random directional perturbation, is introduced. In the Vicsek model noise is considered to be uniformly distributed as $\mathcal{U}(-\eta/2, \eta/2)$, for some constant $0 < \eta \leq 2\pi$. The directional update of agent $i$ can then be expressed as a realisation from

$$\theta_{i,t+1} \,|\, \langle\theta\rangle_{i,t}, \eta \sim \mathcal{U}(\langle\theta\rangle_{i,t} - \eta/2, \langle\theta\rangle_{i,t} + \eta/2), \tag{4.2}$$

where $\langle\theta\rangle_{i,t}$ represents the average direction of motion of agent $i$'s neighbours at time $t$. A weighted circular mean, computed using the definition of the atan2 function given in Equation (A.1), is used to compute $\langle\theta\rangle_{i,t}$ as:

$$\langle\theta\rangle_{i,t} = \operatorname{atan2}\left(\sum_{j=1}^{N} \omega_{ij,t} \sin\theta_{j,t}, \sum_{j=1}^{N} \omega_{ij,t} \cos\theta_{j,t}\right). \tag{4.3}$$

With this, the weighting $\omega_{ij,t}$ represents the strength of the interaction between agent $i$ and agent $j$ at time $t$. In the Vicsek model agent $i$ interacts with neighbours which are within distance $r \in \mathbb{R}^+$ of its current position. This interaction can be implemented with the weighting rule

$$\omega_{ij,t} = \begin{cases} 1 & \text{if } d_{ij,t} \leq r, \\ 0 & \text{otherwise,} \end{cases} \tag{4.4}$$

where $d_{ij,t}$ is the Euclidean-distance between the positions of agent $i$ and agent $j$ at time $t$:

$$d_{ij,t} = \sqrt{(x_{j,t} - x_{i,t})^2 + (y_{j,t} - y_{i,t})^2}.$$

The interaction rule implemented in the Vicsek model represents a discontinuous interaction (see Figure 4.3(a)), as the interaction kernel is very sensitive to small perturbations in distances. More formally, this discontinuity can be shown by considering the value of $\omega_{ij,t}$ as $d_{ij,t}$ approaches $r$ from above and below. As $d_{ij,t}$ approaches $r$ from above we realise the limit $\lim_{d_{ij,t}\to r^+} \omega_{ij,t} = 0$. Conversely, as $d_{ij,t}$ approaches $r$ from below we observe

$\lim_{d_{ij,t}\to r^-} \omega_{ij,t} = 1$. As the weighting tends to different values as the limit of $d_{ij,t} = r$ is approached, we see that there is a discontinuity at $d_{ij,t} = r$. This weighting rule and its resulting discontinuity are visualised in Figure 4.3(a).

The Vicsek model was inspired by models of ferromagnetism. In these models particles align spin states with neighbouring particles. Although a discontinuous interaction rule may be appropriate for such models, it is not clear whether the hard cut-off imposed by the interaction radius $r$ is appropriate for biological systems. Later (Section 4.2.2), we shall introduce models implementing continuous interaction rules as a biologically-motivated alternative.

Vicsek, Czirók et al. (1995) used this model to examine polarisation of simulated flocks as the magnitude of the noise, $|\eta|$, was varied. The polarisation of a flock at time $t$—sometimes referred to as its alignment—was quantified by the absolute value of the normalised velocity of the flock:

$$v_{a,t} = \frac{1}{Nv}\left|\sum_{i=1}^{N} \boldsymbol{v}_{i,t}\right|. \tag{4.5}$$

A flock moving incohesively at time $t$, with individual members directed randomly, has polarisation $v_{a,t} \approx 0$. Conversely, a highly polarised flock in which all members are moving in the same direction at time $t$ has an alignment $v_{a,t} \approx 1$. As the amount of noise in a simulation is increased flocks are observed to transition from an ordered $v_{a,t} \approx 1$ state to a disordered $v_{a,t} \approx 0$ state. For a fixed flock density, smaller flocks are observed to be more resistant to noise than larger flocks.

### 4.1.1 *Stochasticity*

Inspired by models of interacting particles, the uniformly distributed noise implemented in simulations of the Vicsek model is analogous to temperature in a physical system. However, it is not clear whether this distribution is a reasonable choice for biologically-motivated systems. In some of the literature authors instead assume that noise is normally distributed (Couzin, Krause, James et al. 2002). In this case an agent's directional update typically takes the form:

$$\theta_{i,t+1} \,|\, \langle\theta\rangle_{i,t}, \sigma_Y \sim \mathcal{N}(\langle\theta\rangle_{i,t}, \sigma_Y).$$

Given that the $\theta_{i,t}$ represent circular quantities, it is not clear whether a choice of normally distributed noise is appropriate. Consider that the normal distribution has infinite support, yet $\theta_{i,t} \in [-\pi, \pi)$ for $i = 1, \ldots, N$ and $t = 1, \ldots, T$. The von Mises distribution

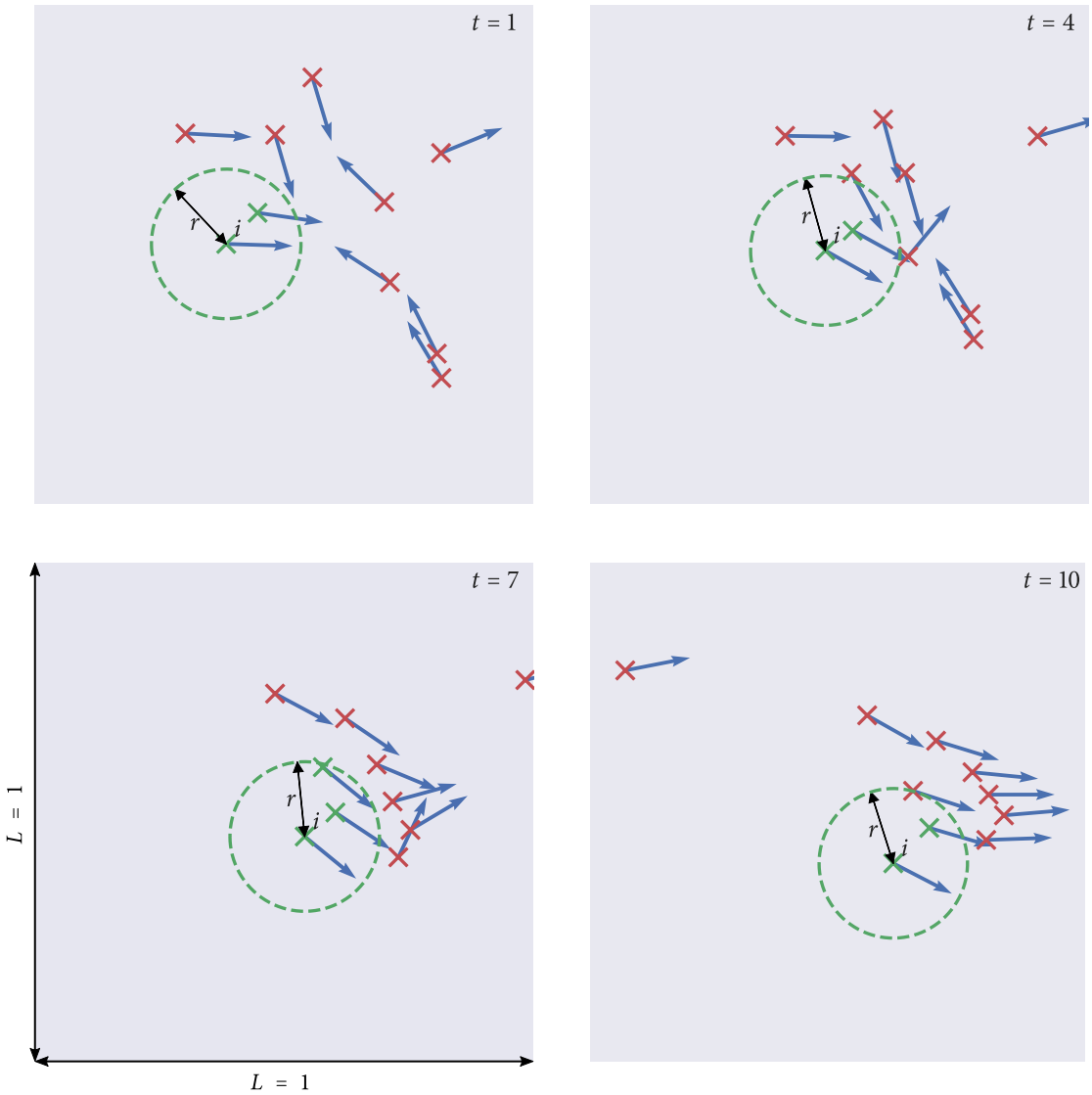



**Figure 4.1:** Visualisations from a simulation of the Vicsek model. At time $t = 1$, $N = 10$ agents are assigned random positions within a square-cell of side length $L = 1$. Initially, the directions of motion of individuals are realised from a $\mathcal{U}(-\pi, \pi)$ distribution. Between time steps agents move with speed $v = 0.03$, and update directions according to Equation (4.2), with $\eta = \pi/16$. The interaction zone of agent $i$ is illustrated throughout the simulation by a circle of radius $r$ centred at $\boldsymbol{x}_{i,t}$. The positions of neighbours within agent $i$'s interaction zone are visualised with a green cross. Individuals which lie outside of agent $i$'s interaction zone have their positions denoted by a red cross.

(Equation (A.3)) is a continuous probability distribution on the circle, and so potentially represents a more appropriate noise structure. However, we observe that when the magnitude of the noise is small, the von Mises distribution can be approximated by a normal

distribution, as shown in Appendix A.4, and so offers no more flexibility than normally distributed noise.

The Student's $t$-distribution is similar to the normal distribution—symmetric, bell-shaped, and with infinite support—but has heavier tails. How *much* extra weight is given to the tails is controlled by the degrees of freedom parameter $\nu$. As the degrees of freedom parameter $\nu$ tends to infinity, a Student's $t$-distribution with scale $\sigma_Y$ and location $\mu$, tends to a normal distribution with corresponding mean and standard deviation. With this the Student's $t$-distribution represents a more flexible noise structure than the normal distribution.

Throughout this work we will assume that the noise experienced by individuals is distributed according to some generalised Student's $t$-distribution with $\nu$ degrees of freedom and scale $\sigma_Y$. In this situation, the directional update of agent $i$ can be expressed:

$$\theta_{i,t+1} \,|\, \langle\theta\rangle_{i,t}, \nu, \sigma_Y \sim t_\nu(\langle\theta\rangle_{i,t}, \sigma_Y). \tag{4.6}$$

Although the Student's $t$-distribution represents a more flexible noise structure, it also comes at the cost of model complexity: with the introduction of the degrees of freedom $\nu$ as an additional model parameter.

### 4.1.2 *Boundary conditions*

Recall that simulations of the Vicsek model take place in a square-cell with periodic boundary conditions and side-length $L$. In this way the density of a cell remains constant throughout a simulation. However, implementing periodic boundary conditions when we attempt to mimic real flocking events is clearly inappropriate. Instead, we shall consider simulations to take place in the unrestricted continuous domain represented by $\mathbb{R}^2$. On top of realism, performing simulations in this unrestricted domain allows more informative visualisations of the resulting data, as in Figure 4.2. Here we are able to visualise the positions of agents throughout a simulation in a single graphic. Trajectory plots of simulations taking place in periodic cells can be difficult to interpret, as it is not easy to visually track individuals as they cross the periodic boundary.

## 4.2 DEVELOPMENT

Although we shall consider variations on the standard Vicsek model, we shall leave most aspects of this model unchanged. In the following models, agents shall always update

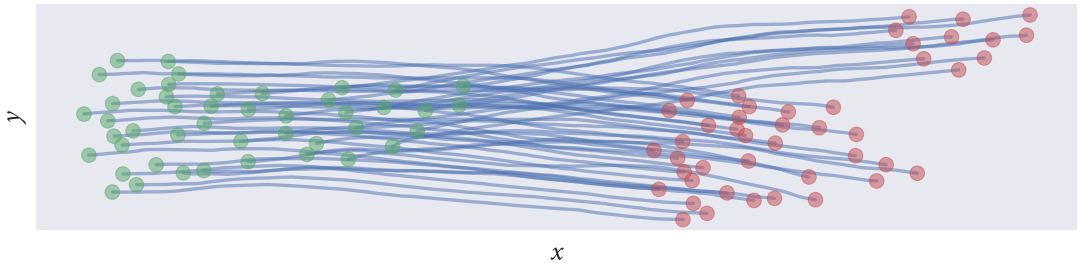



**Figure 4.2:** A trajectory plot representing the movements of $N = 45$ agents moving through the unrestricted continuous domain $\mathbb{R}^2$. Agents are initially positioned at the locations represented by the green markers. The model is simulated for $T = 200$ time steps. The red markers represent the positions of agents at the end of the simulation. Blue lines represent the trajectories of motion of individual agents throughout the simulation.

their positions as in Equation (4.1). Similarly, agents will always update their directions of motion according to Equations (4.3) and (4.6). In fact, all our variations on the standard Vicsek model shall be made by manipulating the computation of the weighting $\omega_{ij,t}$ (Equation (4.4)). Recall that $\omega_{ij,t}$ represents the strength of the interaction between agent $i$ and agent $j$ at time $t$. Altering the functional form of $\omega_{ij,t}$ allows us to investigate how the behaviour of a neighbour influences a flock member.

### 4.2.1 *The Null model*

In Chapters 5 and 6 we fit a number of different models to data. We then assess the predictive performance of these models. Model performance is assessed by the ability to explain observation (quantified by the likelihood), with simple models favoured over complex models. In making such comparisons, it can be useful to have a baseline with which to compare all subsequent models. It is with this intention which we introduce the Null model.

In the Null model there are *no* interactions between flock members. All directional changes are attributed to random noise. The Null model implements directional updates as in Equation (4.6), with location $\langle\theta\rangle_{i,t}$ computed from Equation (4.3) with weighting $\omega_{ij,t} = \delta_{ij}$, where $\delta_{ij}$ is the Kronecker delta:

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

With this agents effectively ignore their neighbours and update their directions of motion as $\theta_{i,t+1} \,|\, \theta_{i,t}, \nu, \sigma_Y \sim t_\nu(\theta_{i,t}, \sigma_Y)$.

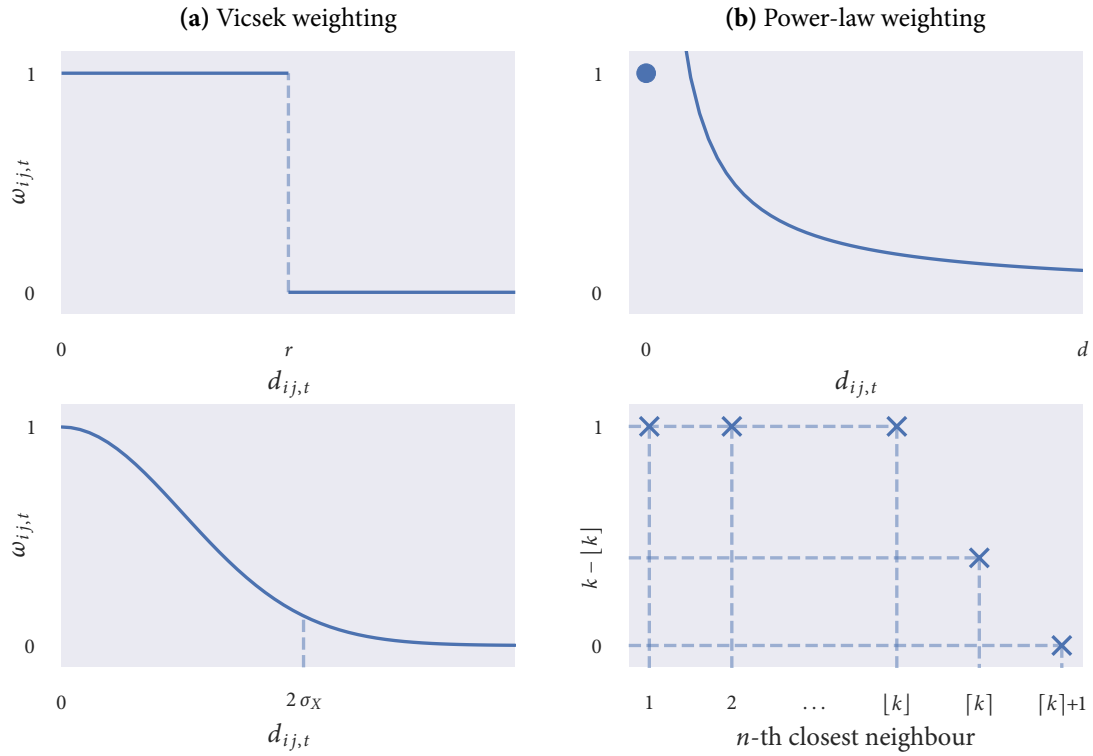



**Figure 4.3:** Visualisations of the weighting rules investigated in this thesis. (a) The interaction implemented in the Vicsek model, parameterised by the interaction radius $r$, exhibits a discontinuity (Equation (4.4)). (b) In the power-law weighted model the strength of the interaction between individuals decays as a power-law relationship with their distance apart, parameterised by decay rate $\alpha$ (Equation (4.7)). (c) The strength of interaction between individuals is controlled by a Gaussian kernel with standard deviation $\sigma_X$ (Equation (4.8)). (d) A realisation of a topological interaction which allows contributions from partial neighbours (Equation (4.9)).

### 4.2.2 *Continuous interaction models*

We have observed that the interaction rule implemented in the Vicsek model exhibits a discontinuity (Section 4.1 and Figure 4.3(a)). This discontinuity raises two concerns; (*i*) the biological-realism of an interaction rule implementing a hard cut-off; (*ii*) the ability to fit such a model to data (recall from Section 3.2.3 that gradient-based inference schemes cannot be applied to discontinuous target distributions).

These concerns are addressed by considering variations of the Vicsek model in which *continuous* interaction rules are implemented. These rules arguably represent more biologically realistic behaviours. In addition to this, the MCMC algorithms used in the model-fitting process perform better for continuously differentiable target distributions.

Metric models, of which Vicsek is an example, implement weighting rules which are explicitly distance-dependent. Here we shall motivate two novel metric interaction rules. We propose that these represent intuitive and biologically realistic rules.

### *Power-law interaction*

A power-law relationship between two quantities occurs when a relative change in one quantity results in a proportional change in the other. Power-law relationships are found frequently in physical and biological systems (Newman 2005), and so represent a natural choice of interaction rule to investigate.

For the power-law weighted model we assume that the interaction strength between individuals decays as some power-law with their distance apart:

$$\omega_{ij,t} = \begin{cases} 1 & \text{if } d_{ij,t} = 0, \\ d_{ij,t}^{-\alpha} & \text{otherwise.} \end{cases} \tag{4.7}$$

The interaction of the Vicsek model is parameterised by the radius $r \in \mathbb{R}$. Here the interaction is parameterised by some $\alpha \in \mathbb{R}^+$. The parameter $\alpha$ controls the rate at which the interaction strength between individuals decays.

In this model, the *closer* a neighbour to agent $i$, the *more influence* it has over agent $i$'s behaviour. Conversely, the *further away* a neighbour is from the position of agent $i$, the *less influence* it has on agent $i$'s movements. The rate at which influence decays is controlled by the value of $\alpha$. Small values of $\alpha$ represent a more gradual decay of influence with distance, whereas large values of $\alpha$ correspond to a weighting which decays more rapidly. As $\alpha \to 0$ this model becomes equivalent to Vicsek with $r \to \infty$, with each individual interacting with all flock members with equal strength. This weighting rule is visualised in Figure 4.3(b).

This interaction rule, although continuous for $d_{ij,t} > 0$, still exhibits a discontinuity at $d_{ij,t} = 0$. However, to the flocking events which we fit this model we observe $d_{ij,t} \gg 0$ $\forall\, i \neq j$, and so this discontinuity does not pose an issue during the model fitting.

### *Gaussian interaction*

Similar to the power-law weighted model, in the Gaussian weighted model the influence which neighbour $j$ exerts over agent $i$ is inversely proportional to their distance apart; the *further* away a neighbour the *less* influence it exerts. The Gaussian weighted interaction is

continuous for all $d_{ij,t}$, and can be expressed as:

$$\omega_{ij,t} = \frac{1}{\sqrt{2\pi\sigma_X^2}} e^{-\frac{1}{2}\left(\frac{d_{ij,t}}{\sigma_X}\right)^2}. \tag{4.8}$$

This rule is illustrated in Figure 4.3(c).

From Equation (4.8) the parameter $\sigma_X$ can be seen to control the length-scale over which interaction takes place. Small values of $\sigma_X$ represent a weighting function which is very peaked around agent $i$'s position. As $\sigma_X \to \infty$ the weighting approaches a schema in which all neighbours in the flock are given the same weight. This limiting behaviour is equivalent to that of Vicsek as $r \to \infty$, and the power-law weighted model as $\alpha \to 0$.

Although the Gaussian model implements a similar interaction kernel to that of the power-law weighted model, there remains differences in their functional forms. Notably, the power-law interaction represents a convex weighting function, whereas the Gaussian interaction is a concave weighting (Figure 4.3). The Gaussian model also has some inherent length scale, expressed by the parameter $\sigma_X$, whereas the power-law weighted interaction is scale-free.

### 4.2.3 *Topological models*

Topological interactions represent a different type of behaviour to metric rules. In topological models agents interact with their $k \in \mathbb{N}$ nearest neighbours, *regardless* of their distance apart. This interaction rule was motivated by simulations which suggested that flocks implementing topological rules were more resistant to perturbations, such as those provoked by predators, than those implementing metric rules (Ballerini et al. 2008; Ginelli and Chaté 2010; Camperi et al. 2012). As such, it was argued that topological interaction rules offer an evolutionary advantage to flocks over metric interaction rules. It was also posed that the cognitive limits of individuals puts some upper bound on the number of neighbours which agents can interact with at any one time (Nieder 2005; Giardina 2008).

Whilst retaining the form of the topological rule, we look to relax the constraint that agents may only interact with some *integer* number of closest neighbours. Instead, we will allow an agent to interact with some $k \in \mathbb{R}^+$ nearest neighbours. This allows partial neighbours, which may make some lesser contribution to an agent's behaviour. An additional benefit of this more flexible interaction structure is given by the observation that real-valued parameters are typically easier to infer from data than discrete parameters (recall from Section 3.2.3 that continuous valued parameter distributions are necessary for gradient-based

inference schemes).

For some $k \in \mathbb{R}^+$ we say that each agent gives weighting 1 to its $\lfloor k \rfloor$ nearest neighbours, weighting $k - \lfloor k \rfloor$ to its $\lceil k \rceil$-th nearest neighbour, and weighting 0 to all other agents, where $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ denote the floor and ceiling function respectively. More concisely, we can express this as the weighting

$$\omega_{ij,t} = \begin{cases} 1 & \text{if } d_{ij,t} \leq \min_j(d_{ij,t}, \lfloor k \rfloor), \\ k - \lfloor k \rfloor & \text{if } d_{ij,t} = \min_j(d_{ij,t}, \lceil k \rceil), \\ 0 & \text{otherwise,} \end{cases} \tag{4.9}$$

where $\min_j(d_{ij,t}, k)$ is the distance to agent $i$'s $k$-th closest neighbour at time $t$. This weighting rule is illustrated in Figure 4.3(d).

### 4.2.4 *Intra-flock variation*

In all the model variations considered thus far, there has been no allowance for behavioural or biological variation within flocks. That is, presented with the same circumstances, all of our agents would behave identically. This is despite our understanding of the importance of biological variation in nature.

Within a single flock, we may reasonably expect there to be variation in the age, cognitive ability, social standing and physical attributes of the individuals. Naturally then, it seems reasonable to suggest that if the attributes of individuals may vary, then their behavioural responses to a particular set of circumstances may vary too. This is in keeping with the observation that the orientation abilities of migrating raptors is age dependent (Thorup et al. 2003).

Models allowing intra-flock variation have been considered before. Particularly impactful was the work of Couzin, Krause, Franks et al. (2005), which simulated flocks of individuals partitioned into a group of leaders and a group of followers. The two groups interacted according to different behavioural rules. Followers sought only to amend their directions in response to the movements of their neighbours. Whereas leaders balanced their desire to interact with neighbours and their desire to move in a particular direction (for example, toward a migratory of foraging goal).

Instead of dividing agents into distinct behavioural categories, we permit a range of behavioural responses by allowing interaction and noise parameters to vary between individuals. Some agents may be more or less susceptible to noise than others. Consider then

that agent $i$ experiences noise distributed according to a generalised Students $t$-distribution with $\nu$ degrees of freedom and scale $\sigma_{Y_i}$. Agent $i$'s directional update is then given as

$$\theta_{i,t+1} \,|\, \langle\theta\rangle_{i,t}, \nu, \sigma_{Y_i} \sim t_\nu(\langle\theta\rangle_{i,t}, \sigma_{Y_i}).$$

In the interest of simplicity we shall maintain a global-flock value for the degrees of freedom parameter, $\nu$, such that $\nu = \nu_i$. Instead allowing variation in susceptibility of noise to be expressed through the scale $\sigma_{Y_i}$.

In addition to allowing noise parameters to vary within a flock, we also seek to allow interaction parameters to vary within a flock. In the case of Vicsek, agent $i$ would then interact with neighbours positioned within distance $r_i$. This corresponds to the weighting rule:

$$\omega_{ij,t} = \begin{cases} 1 & \text{if } d_{ij,t} \leq r_i, \\ 0 & \text{otherwise.} \end{cases} \tag{4.10}$$

A similar allowance can be made for the power-law weighted ($\alpha \mapsto \alpha_i$), Gaussian weighted ($\sigma_X \mapsto \sigma_{X_i}$) and topological ($k \mapsto k_i$) models considered earlier.

## *Imposing hierarchy*

Rather than considering the individual interaction and noise parameters as completely unrelated, we can consider them as being realisations from some population-level distribution. Hierarchical models allow us to incorporate this extra level of structure into our models.

Hyperparameters are introduced as parameters of the population-level distribution. Hyperpriors are then specified to reflect the practitioner's prior beliefs about likely values of the hyperparameters.

The noise and interaction parameters in our models are all strictly positive, and not bounded above. As such, we shall assume that the population distributions for these parameters follow some gamma distribution. Consider, for example, that the noise parameters $\sigma_{Y_i}$ are distributed according to a gamma distribution with shape $\alpha_Y$ and rate $\beta_Y$. More concisely, we consider our noise parameters as being distributed according to

$$\sigma_{Y_i} \,|\, \alpha_Y, \beta_Y \sim \mathrm{Ga}(\alpha_Y, \beta_Y). \tag{4.11}$$

Our hyperparameters here are $\alpha_Y$ and $\beta_Y$. The practitioner is then required to specify hyperpriors: their prior beliefs about likely values for $\alpha_Y$ and $\beta_Y$. It can be hard to get an intuitive feel for plausible values for the shape and rate parameters, and so we shall instead re-

parameterise our population-level distribution in terms of the mean and variance. Consider that the mean and variance of the gamma distribution can be computed as $m_Y = \alpha_Y/\beta_Y$ and $v_Y = \alpha_Y/\beta_Y^2$ respectively. From these two relations it can be seen that $\alpha_Y = m_Y^2/v_Y$ and $\beta_Y = m_Y/v_Y$. Equation (4.11) can then be re-expressed as:

$$\sigma_{Y_i} \mid m_Y, v_Y \sim \text{Ga}\left(\frac{m_Y^2}{v_Y}, \frac{m_Y}{v_Y}\right). \tag{4.12}$$

This parameterisation allows the practitioner to specify their hyperpriors over the mean and variance of the distribution, instead of the shape and rate. The hope is that this will give the practitioner a more intuitive feel for likely values of the hyperparameters.

Population-level distributions are specified in a similar way for the interaction parameters of the Vicsek weighting, power-law weighting, Gaussian weighting and topological models:

$$r_i \mid m_r, v_r \sim \text{Ga}\left(\frac{m_r^2}{v_r}, \frac{m_r}{v_r}\right) \tag{4.13}$$

$$\alpha_i \mid m_\alpha, v_\alpha \sim \text{Ga}\left(\frac{m_\alpha^2}{v_\alpha}, \frac{m_\alpha}{v_\alpha}\right) \tag{4.14}$$

$$\sigma_{X_i} \mid m_X, v_X \sim \text{Ga}\left(\frac{m_X^2}{v_X}, \frac{m_X}{v_X}\right) \tag{4.15}$$

$$k_i \mid m_k, v_k \sim \text{Ga}\left(\frac{m_k^2}{v_k}, \frac{m_k}{v_k}\right) \tag{4.16}$$

## Conclusions

In this chapter we introduced a model developed by Vicsek, Czirók et al. (1995). This work received much attention in the following literature, and many studies were made of its simulations and results. We also considered some of the biological shortcomings of this model, and proposed alterations to address them. Given its relative simplicity, and widespread study, the Vicsek model represents a sensible starting point for model fitting.

To allow meaningful comparison between predictive performance of variations on the Vicsek model we introduced a Null model. This model does *not* simulate an interaction rule, and so directional changes are introduced entirely by noise. Having questioned the biological-realism of Vicsek's discontinuous interaction rule, and considered the practical implications of fitting such a model, we introduced models implementing continuous interaction rules. We argued that continuous interaction rules make behaviours more robust

to small changes in distances. A power-law and Gaussian weighted model were introduced in which interaction strength decays continuously as the distance between individuals increases. A continuous version of a topological rule was also introduced, allowing individuals to make partial contributions to an agent's behaviours.

Variations on these models were considered in which biological and behavioural variation between individuals is allowed; again motivated by considerations of biological-realism. Population-level variation was modelled by a hierarchical structure, with hyperparameters introduced to parameterise the population-level distributions. The population level distributions were reparameterised in terms of the mean and variance, with the intention of easing prior specification.

# 5

# *Simulation studies*

The feasibility of performing parameter inference on real data can be assessed by simulation study. In simulation studies, candidate models are forward simulated for *known* parameter values. This simulated data is then used to assess whether the known parameter values can be captured by inference.

Simulation studies provide good opportunity to assess (and frankly, debug) our implementation of MCMC algorithms constructed to sample realisations from the posterior distribution. If a simulation study is successful, then the statistical machinery built to perform the inference on simulated data can be repurposed to perform inference on *real* data. If parameter inference is *not* possible on simulated data, we know that the same inference will not be possible on real data either.

## 5.1 Global models

Global models refer to models in which all agents take the *same* parameter values. These are in contrast to hierarchical models, in which all agents take *different* parameter values. We shall first seek to perform simulation studies on these global models, as they represent a simpler inference problem.

Recall from Bayes' Theorem (Equation (3.2)) that to realise the posterior distribution we need to combine the likelihood of the data with our prior beliefs. Let us begin by considering the likelihood of observing a single agent updating its direction, in the presence of neighbours, from $\theta_{i,t}$ to $\theta_{i,t+1}$. From Equation (4.6) we have that agent $i$ updates its

direction as a draw from the generalised Students $t$-distribution with location $\langle\theta\rangle_{i,t}$, scale $\sigma_Y$ and degrees of freedom $\nu$. As such, the likelihood of observing this directional update can be quantified by the probability density of the generalised Student's $t$ distribution with $\nu$ degrees of freedom, location $\langle\theta\rangle_{i,t}$, and scale $\sigma_Y$, evaluated at $\theta_{i,t+1}$:

$$L(\sigma_Y, \nu, \langle\theta\rangle_{i,t} \,|\, \theta_{i,t+1}) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}\sigma_Y}\left(1 + \frac{1}{\nu}\left(\frac{\theta_{i,t+1} - \langle\theta\rangle_{i,t}}{\sigma_Y}\right)^2\right)^{-\frac{\nu+1}{2}},$$

where $\Gamma$ is the gamma function. Building on this, let us consider the likelihood of observing agent $i$'s directional updates from time $t$ to time $t+2$. As the noise experienced by an agent is independent between time steps, we may express this likelihood as the product:

$$L(\sigma_Y, \nu, \langle\theta\rangle_{i,t}, \langle\theta\rangle_{i,t+1} \,|\, \theta_{i,t+1}, \theta_{i,t+2}) = L(\sigma_Y, \nu, \langle\theta\rangle_{i,t+1} \,|\, \theta_{i,t+2}) \times L(\sigma_Y, \nu, \langle\theta\rangle_{i,t} \,|\, \theta_{i,t+1}).$$

In general, we wish to express the likelihood of observing agent $i$'s directional updates over any number of times steps. Suppose then that we observe the directions of agent $i$ from time $t = 1$ to time $t = T$. Again, as consecutive realisations from the noise distribution are independent, we can compute the likelihood as a product

$$\begin{aligned} L(\sigma_Y, \nu, \langle\theta\rangle_{i,1:T-1} \,|\, \theta_{i,2:T}) &= \prod_{t=1}^{T-1} L(\sigma_Y, \nu, \langle\theta\rangle_{i,t} \,|\, \theta_{i,t+1}) \\ &= \prod_{t=1}^{T-1} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}\sigma_Y}\left(1 + \frac{1}{\nu}\left(\frac{\theta_{i,t+1} - \langle\theta\rangle_{i,t}}{\sigma_Y}\right)^2\right)^{-\frac{\nu+1}{2}}, \end{aligned}$$

where $\langle\theta\rangle_{i,1:T-1}$ is shorthand for $\langle\theta\rangle_{i,1}, \langle\theta\rangle_{i,2}, \ldots, \langle\theta\rangle_{i,T-1}$, and similarly for $\theta_{i,2:T}$. Finally, although we have expressed the likelihood of observing a single agent's directional changes over $T$ observations, we really wish to express the likelihood of observing *an entire flock's* directional changes over $T$ observations. Consider observing a flock of $N$ individuals over $T$ time steps. The likelihood of observing their directional updates can be quantified as:

$$\begin{aligned} L(\sigma_Y, \nu, \langle\theta\rangle_{1:N,1:T-1} \,|\, \theta_{1:N,2:T}) &= \prod_{i=1}^{N}\prod_{t=1}^{T-1} L(\sigma_Y, \nu, \langle\theta\rangle_{i,t} \,|\, \theta_{i,t+1}) \\ &= \prod_{i=1}^{N}\prod_{t=1}^{T-1} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}\sigma_Y}\left(1 + \frac{1}{\nu}\left(\frac{\theta_{i,t+1} - \langle\theta\rangle_{i,t}}{\sigma_Y}\right)^2\right)^{-\frac{\nu+1}{2}}. \end{aligned} \tag{5.1}$$

The likelihood function for all our models takes the same form. What differs between

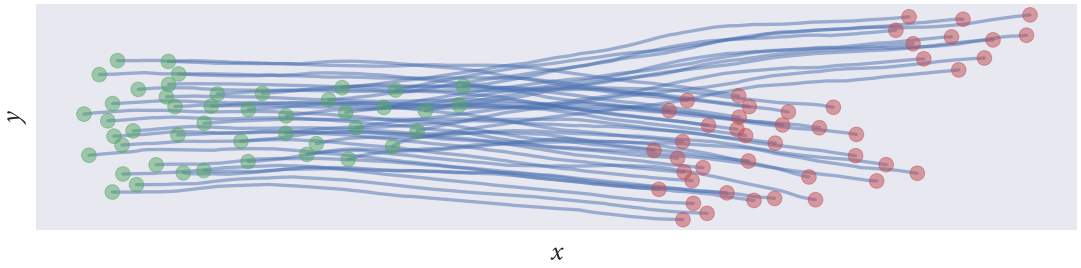



**Figure 5.1:** The trajectories of $N = 45$ agents simulated over $T = 200$ time steps according to the Vicsek model. The initial conditions of the simulation reflect that of a real flocking event, introduced in the proceeding Chapter 6. The parameters of the model are given by $r = 50$, $\sigma_Y = 0.03$ and $v = 7$. Initial conditions and parameter values were chosen with the intention to generate a simulated flock similar to that of a real flocking event. At $t = 0$ agents are positioned at the locations represented by the green markers. Agents travel along the blue lines, with the red markers representing their positions at the end of the simulation.

models is the specification of the weighting function $\omega_{ij,t}$ and the corresponding computation of $\langle\theta\rangle_{i,t}$.

### 5.1.1 *Vicsek model*

We simulate the Vicsek model to realise a flock of $N = 45$ agents moving for $T = 200$ time steps. We desire to work with simulated data which is similar to real data (see, for example, Figures 6.2 to 6.4). As such, the initial conditions for this simulation are taken from an observation of a real flocking event, to be introduced in the proceeding chapter: Chapter 6. Parameters $r = 50$, $\sigma_Y = 0.03$ and $v = 7$ are set for this simulation. Figure 5.1 shows the data generated by this simulation. We seek to assess whether we can recover the true parameter values from the simulated data.

The likelihood of observing the data can be quantified by Equation (5.1). However, to target the posterior we also need to specify our prior beliefs about the model parameters. To allow the data to drive the inference, we shall specify weakly informative priors. Once realisations from the posterior have been made, we overlay our priors to assess how our beliefs have updated in light of the data.

Our prior beliefs about likely values for the degrees of freedom parameter $v$ shall be reflected by a $\text{Ga}(2, 0.1)$ distribution, as popularised by Juárez and Steel (2010). As $r$ and $\sigma_Y$ both represent strictly positive measures, we shall use a gamma distribution to quantify our beliefs about them. These prior beliefs, as detailed in Equation (5.2) and visualised in

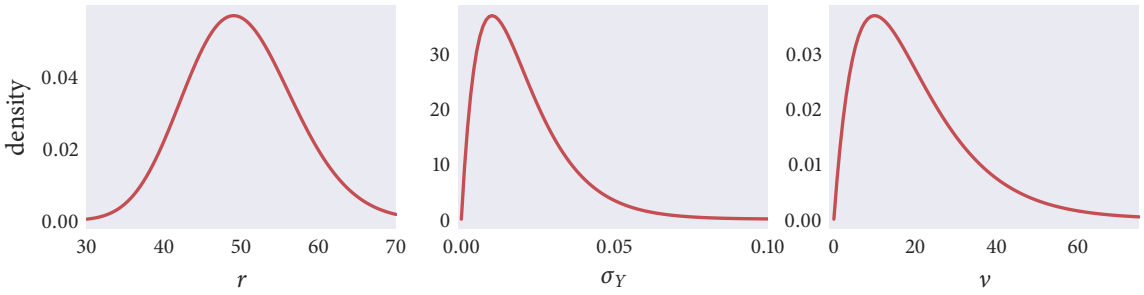



**Figure 5.2:** Probability density functions representing our prior beliefs about plausible parameter values of the Vicsek model. These distributions represent weakly-informative priors, allowing a large range of possible parameter values. These beliefs are expressed mathematically in Equation (5.2).

Figure 5.2, represent vague beliefs, allowing a large range of possible parameter values.

$$\begin{aligned} r &\sim \text{Ga}(50, 1) \\ \sigma_Y &\sim \text{Ga}(2, 100) \\ \nu &\sim \text{Ga}(2, 0.1) \end{aligned} \tag{5.2}$$

Having specified the likelihood function, and quantified our prior beliefs about the model parameters, we now have all the ingredients necessary to target the posterior. Unfortunately, the discontinuity in the weighting function of the Vicsek model results in a discontinuous posterior. As Stan (the probabilistic programming language introduced in Section 3.2.3) requires a continuous-parameter valued posterior that is differentiable almost everywhere, we are unable to use Stan and instead target the posterior distribution with a random walk Metropolis–Hastings sampler.

We implement the random walk sampler for 1,100,000 iterations. The first 100,000 iterations of the scheme are discarded to allow the sampler to converge. The chains were initialised as draws from our prior beliefs. The output of this scheme is summarised in Table 5.1.

Figure 5.3 shows the chains generated by our inference scheme, having discarded the initial burn-in period. From these trajectories we see evidence that our sampler has converged: the chains oscillating around fixed locations with some common variance. The samples generated by these chains are visualised in Figure 5.4. The true parameter values used to generate the simulated data are shown by vertical green lines. See that in each posterior the true value is successfully captured by the posterior density, and lies close to the posterior mode. The posterior densities about the parameters $\nu$ and $\sigma_Y$ look Gaussian,

however, the posterior distribution over $r$ is non-Gaussian. The non-Gaussian form of the posterior about $r$ is a product of the discontinuous interaction rule which it represents. Our prior beliefs are overlain in red and appear flat in comparison to our posteriors, reflecting that our beliefs have updated considerably in light of the data.

Considering the output of this simulation study, we conclude that we can accurately infer the true parameter values of data simulated from the Vicsek model. This gives us confidence in moving forward to fit this model to real data.

### 5.1.2 *Continuous models*

We have demonstrated that we can fit the Vicsek model to simulated data. We now seek to fit the power-law weighted (Equation (4.7)) and Gaussian weighted (Equation (4.8)) models to simulated data.

As these models implement continuous interaction rules, we may attempt parameter inference using the Stan programming language. Equation (5.1) can be used to quantify the likelihood of observing simulated data. To target the posterior we combine this likelihood with our prior beliefs.

The power-law and Gaussian weighted models represent similar interaction rules. As with the parameters $v$, $\sigma_Y$ and $r$, we shall use gamma distributions to quantify our prior beliefs about $\alpha$ and $\sigma_X$. To ensure that our prior beliefs between these models are consistent, we shall determine our beliefs by considering plausible values of the weighting $\omega_{ij,t}$. Using information about $d_{ij,t}$, along with Equations (4.7) and (4.8), we can use our beliefs about $\omega_{ij,t}$ to derive beliefs about $\alpha$ and $\sigma_X$.

To realise our prior beliefs we shall make two probabilistic statements about $\omega_{ij,t}$. These can be used to determine prior distributions which best reflect these statements. Firstly, we consider the weighting which agent $i$ gives to its closest neighbour at time $t$. We believe with probability 0.025 that this weighting will be less than or equal to 0.25. Similarly, we believe with probability 0.975 that the weighting which agent $i$ gives to its fifth closest neighbour will be less than or equal to 0.90. More concisely, we can express these statements as:

$$\begin{aligned} P(\omega_{ij,t}(\min_j(d_{ij,t}, 1)) \leq 0.25) &= 0.025, \\ P(\omega_{ij,t}(\min_j(d_{ij,t}, 5)) \leq 0.90) &= 0.975, \end{aligned} \tag{5.3}$$

where $\omega_{ij,t}(\min_j(d_{ij,t}, k))$ is the influence of agent $i$'s $k$-th closest neighbour at time $t$. To determine prior beliefs which best reflect these statements, we seek shape and scale

| Parameter | mean | sd | 5% | 95% | ESS |
|---|---|---|---|---|---|
| $r$ | 50.02 | 0.04 | 49.99 | 50.11 | 12 290 |
| $\sigma_Y$ | 0.03 | 0.00 | 0.03 | 0.03 | 44 560 |
| $\nu$ | 7.01 | 0.51 | 6.31 | 7.99 | 47 420 |

**Table 5.1:** Summarising the posterior realisations for the parameters inferred in fitting the Vicsek model to simulated data. Realisations were made by implementing a random walk Metropolis–Hastings sampler. Columns report the posterior mean and standard deviation for each parameter. In addition to this, the fifth and ninety-fifth percentiles of our posteriors are quantified. Finally, the effective sample size (ESS) of each chain is computed (and rounded to the nearest multiple of ten). The large values of the ESS observed give reliability to our computed posterior summaries.

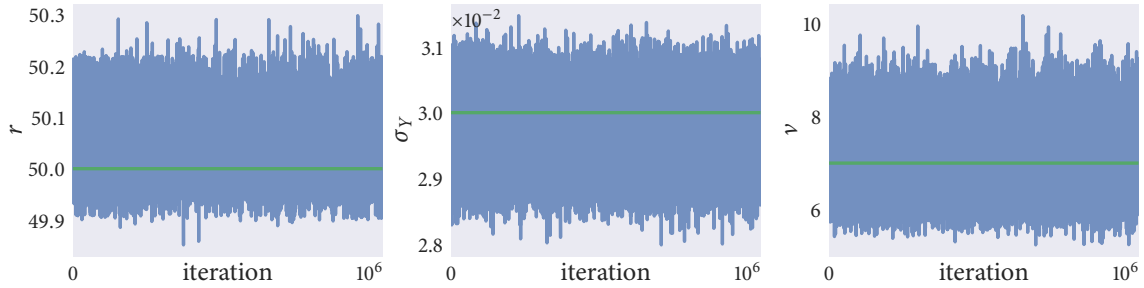



**Figure 5.3:** Sequences generated by a random-walk Metropolis–Hastings algorithm fitting the Vicsek model to simulated data. The initial 100,000 iterations are discarded to allow a burn-in period. Horizontal green lines represent the true parameter values used to generate the simulated data. The chains look well-behaved, showing no irregularities, and suggest that our sampler has converged.

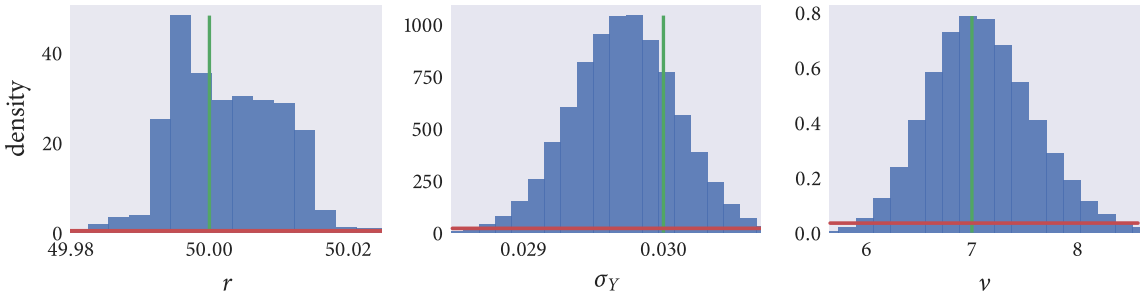



**Figure 5.4:** Distributions of the posterior samples drawn in fitting the Vicsek model to simulated data. The green lines represent the known parameter values which we seek to recover. Prior beliefs are overlain in red. Our posterior densities are seen to accurately capture the true parameter values. The data can be seen to have been very informative, as our posterior densities have updated considerably from our prior beliefs.

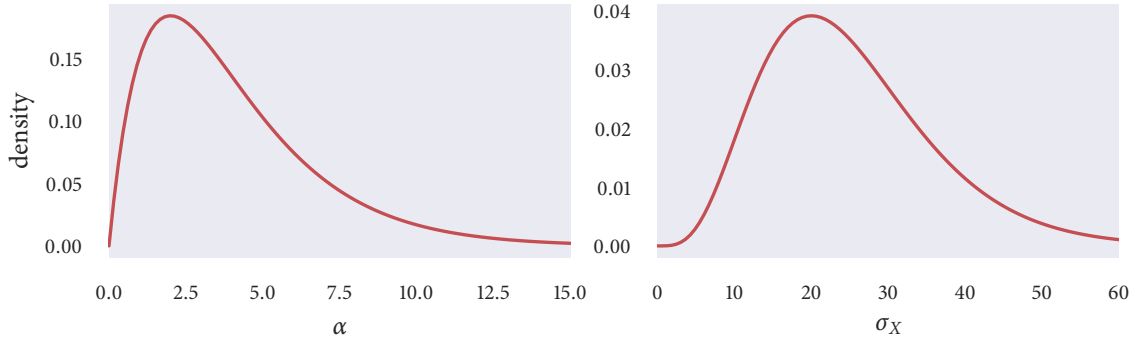



**Figure 5.5:** Probability density functions representing our prior beliefs about $\alpha$ and $\sigma_X$, the interaction parameters of the power-law and Gaussian weighted models respectively. These prior beliefs are given analytically in Equation (5.4). Our prior beliefs about these interaction parameters represent equivalent beliefs about the weighting function $\omega_{ij,t}$.

parameters for our gamma priors which minimise:

$$[P(\omega_{ij,t}(\min_j(d_{ij,t}, 1)) \leq 0.25) - 0.025]^2 + [P(\omega_{ij,t}(\min_j(d_{ij,t}, 5)) \leq 0.90) - 0.975]^2.$$

A simple numerical optimisation routine can be used to minimise this function. With this we realise prior beliefs:

$$\begin{aligned} \alpha &\sim \text{Ga}(2, 2) \\ \sigma_X &\sim \text{Ga}(5, 5), \end{aligned} \tag{5.4}$$

which we visualise in Figure 5.5.

### *Power-law weighted*

The initial conditions for our simulated data are taken from an observation of a real flocking event, the same event used to initialise the simulation in Section 5.1.1. A realistic-looking flock is generated with parameters $\alpha = 1.5$, $\sigma_Y = 0.03$ and $\nu = 7$ (Equation (4.7)). Figure 5.6 shows the trajectories realised by this simulation. We then seek to recover the known parameter values by inference.

Stan is used to perform parameter inference on this model. Four independent chains are initialised as realisations from our prior beliefs. Each chain is then simulated for 10,000 iterations, with the first 5000 iterations discarded to allow a warm-up period. Table 5.2 summarises the posterior draws generated by this run. The tabulated split-$\widehat{R}$ values (Equation (3.10)) indicate that our sampler has converged for all parameter values.

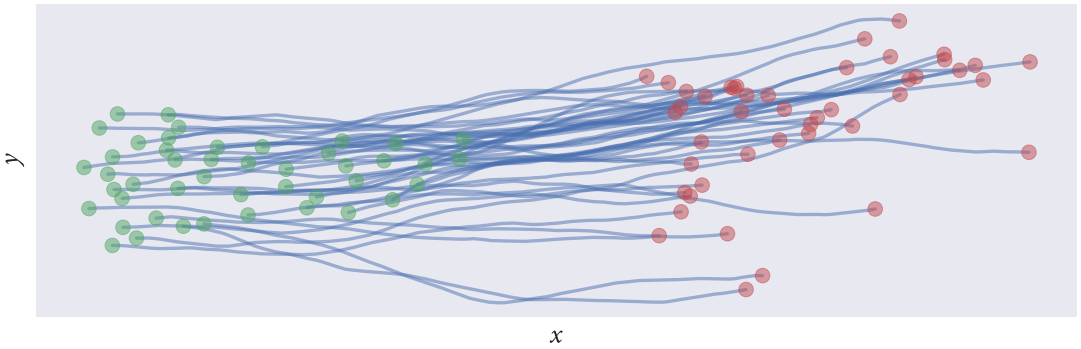



**Figure 5.6:** Simulated data generated by a run of the power-law weighted model with parameter values $\alpha = 1.5$, $\sigma_Y = 0.03$ and $\nu = 7$. The data represents the trajectories of motion of $N = 45$ agents moving for $T = 200$ frames. Initial conditions for this simulation were realised from an observation of a real flocking event.

| Parameter | mean | sd | 5% | 95% | ESS | $\widehat{R}$ |
|---|---|---|---|---|---|---|
| $\alpha$ | 1.50 | 0.02 | 1.48 | 1.54 | 13 130 | 1.0 |
| $\sigma_Y$ | 0.03 | 0.00 | 0.03 | 0.03 | 11 400 | 1.0 |
| $\nu$ | 7.09 | 0.50 | 6.28 | 7.91 | 11 580 | 1.0 |

**Table 5.2:** Summarising the posterior realisations made by Stan fitting simulated data to the power-law weighted model. Columns show the posterior mean and standard deviation of each model parameter, along with the fifth and ninety-fifth percentiles of our beliefs. The number of effective samples drawn from the posterior is quantified by the ESS. The values of $\widehat{R}$ computed indicate that our chains have all converged.

The chains generated by this sampler are shown in Figure 5.7. Each colour represents an independent chain. Although the four chains were initialised at different locations, they all converge to the same region.

Histograms of our posterior draws are presented in Figure 5.8. The vertical green lines in these plots represent the true parameter values. In each case, our posteriors can be seen to capture the true values. In fact, the true parameter values are seen to be well represented by the posterior mode. In contrast with the posteriors derived in fitting the Vicsek model to simulated data (Section 5.1.1), the posterior distribution of the interaction parameter (here $\alpha$), *is* Gaussian. Recall that our posterior density about $r$ (Figure 5.4) was non-Gaussian, owing to the discontinuous weighting rule it represents. Our prior beliefs are overlain in red. These priors appear flat in comparison to our posteriors, indicating that we have learnt a lot from the data.

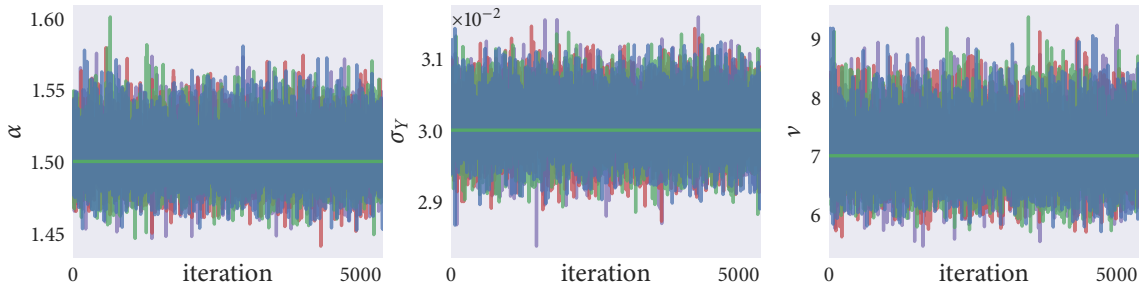



**Figure 5.7:** Trajectories of the four independent chains for each parameter inferred in fitting the power-law weighted model to simulated data. Each independent chain is overlain in a different colour. Although each chain was initialised at a different starting point, they can be seen to converge to the same distribution. Chains were simulated for 10,000 iterations, with the first 5000 iterations discarded to allow convergence. The chains appear well-behaved, showing no irregularities and are seen to oscillate around the true parameter values, represented by the green horizontal lines.

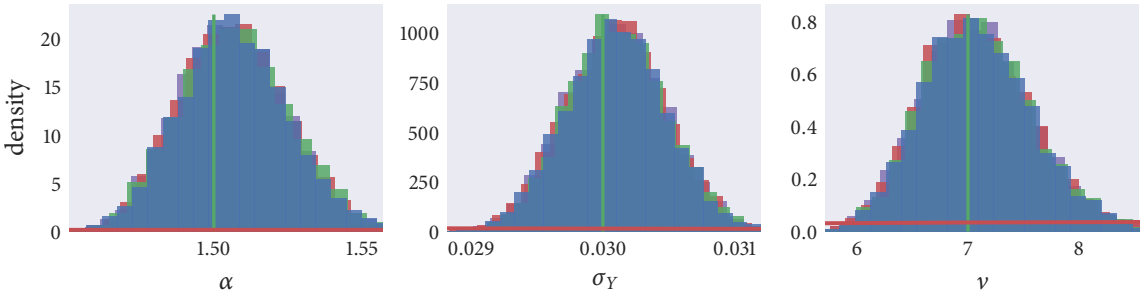



**Figure 5.8:** Histogram plots of our samples drawn from the posterior distribution in fitting simulated data to the power-law weighted model. The true parameter values are represented by the green vertical lines. See that the true value for each parameter is well represented by the posterior mode. Prior beliefs are overlain in red, and appear flat in comparison to our posterior densities. This indicates that the observed data was informative, and that our prior beliefs did not adversely effect our posterior densities.

### *Gaussian weighted*

As with the previous simulation studies, we forward simulate this model with $N = 45$ agents and $T = 200$ time steps, using the same initial state used in Sections 5.1.1 and 5.1.2. Figure 5.9 illustrates the data generated by this simulation. Here, parameters $\sigma_X = 20$, $\sigma_Y = 0.025$ and $\nu = 7$ were used to produce realistic-looking flocking behaviour.

We shall use Stan to perform parameter inference. Four independent sequences are initialised at draws from our prior beliefs and simulated for 10,000 iterations. The first 5000 iterations are discarded to allow for a warm-up period. Table 5.3 summarises the output from this run. The reported split-$\widehat{R}$ values indicate that the sequences have converged.

The output visualised in Figures 5.10 and 5.11 shows a well-behaved sampler, capable of

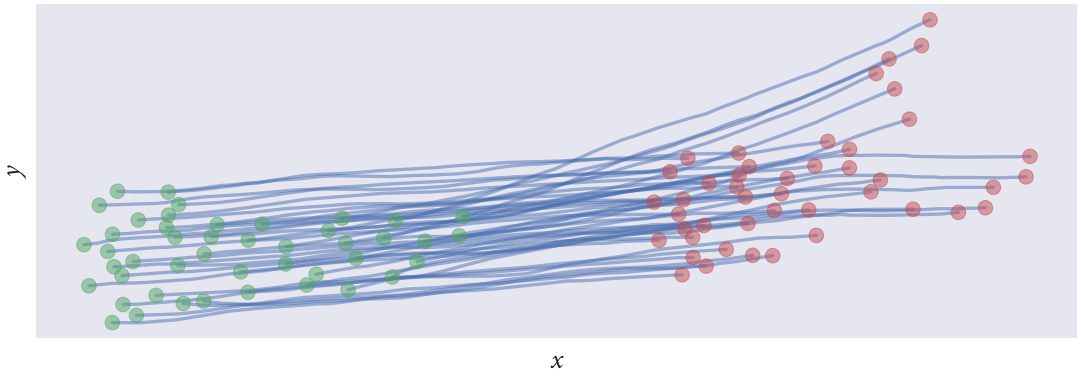



**Figure 5.9:** Data generated from a simulation of the Gaussian weighted model with parameters $\sigma_X = 20$, $\sigma_Y = 0.025$ and $v = 7$. A total of $N = 45$ agents were simulated for $T = 200$ time steps. Initial conditions were taken from an observation of a real flocking event, with the intention to generate data similar to that of an observed flock.

| Parameter | mean | sd | 5% | 95% | ESS | $\widehat{R}$ |
|---|---|---|---|---|---|---|
| $\sigma_X$ | 19.94 | 0.20 | 19.60 | 20.27 | 13 310 | 1.0 |
| $\sigma_Y$ | 0.025 | 0.00 | 0.024 | 0.025 | 11 700 | 1.0 |
| $v$ | 7.25 | 0.53 | 6.41 | 8.12 | 11 360 | 1.0 |

**Table 5.3:** Summaries of the posterior densities realised in fitting the Gaussian weighted model to simulated data. Each row represents a parameter of the Gaussian model. Each column presents a different summary of our posterior density. The computed $\widehat{R}$ indicate that the simulated chains converged, and the realised ESS values show a satisfactory number of draws made from the posterior.

capturing the true parameter values of the Gaussian weighted model from simulated data.

### 5.1.3 *Topological model*

In Section 5.1.2 we demonstrated that it is possible to accurately capture the parameters of the power-law and Gaussian weighted models from simulated data. However, we are interested in comparing the predictive-performance of metric and topological models with real data. As such, we need a simulation study on our topological model.

Equations (4.9) and (5.1) allow us to quantify the goodness of fit of the topological model with parameters $k$, $\sigma_Y$ and $v$ to observations of a flocking event. To target the posterior we must combine this likelihood with prior beliefs about the model parameters. We have no reason to believe that the noise experienced by individuals under a topological regime would be any different to that experienced under a metric regime. As such, our prior beliefs

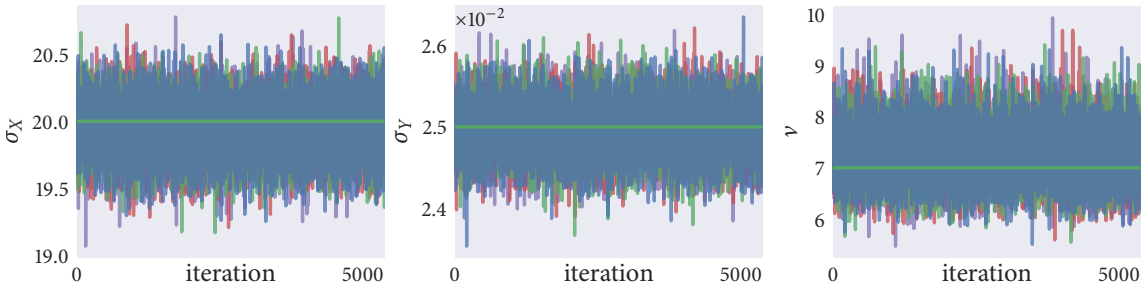



**Figure 5.10:** Visualising the Markov chains generated fitting the Gaussian weighted model to simulated data. Each colour chain represents an independent chain. Although the chains are initialised as different draws from our priors, they all converge to the same common distribution. The chains look well-behaved; they show no irregularities, and are observed to oscillate around the true parameter values with constant variance.

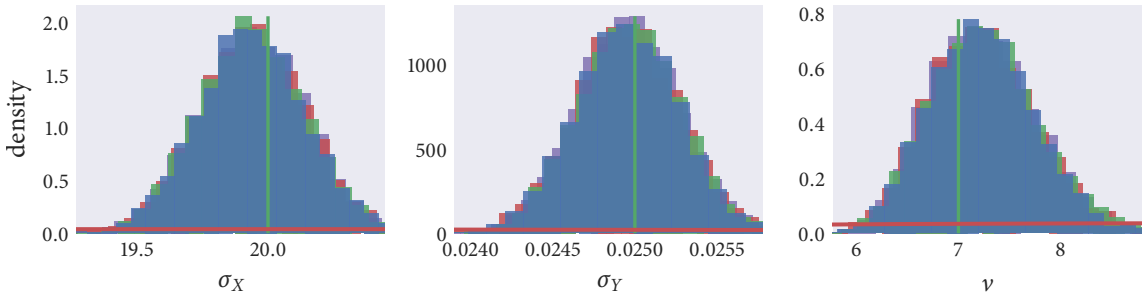



**Figure 5.11:** Draws realised by Stan in fitting the Gaussian weighted model to simulated data. The true values (green vertical lines) are well represented by the posterior mode. Our posteriors appear much more peaked than our priors (overlain in red), suggesting that our choice of priors had little influence over the realised posteriors.

about the noise parameters $\sigma_Y$ and $\nu$ shall remain the same as in the metric case: as expressed in Equation (5.2).

We are then left to quantify our prior beliefs about $k$, the number of nearest neighbours which an agent interacts with. Previous work suggested that agents interact with their six to seven nearest neighbours (Ballerini et al. 2008). However, this work investigated three-dimensional flocking events of flying birds. As we are focusing on data restricted to a two-dimensional plane—flocking sheep (Chapter 6) and swimming birds (Section 7.2)—we believe that the number of nearest neighbours will be lower than the six to seven found previously. We find that a $\text{Ga}(6, 1/2)$ distribution captures a wide range of plausible values for $k$ (Figure 5.12).

The topological model is forward simulated to generate data for the simulation study. Forty-five agents are simulated for two-hundred time steps. Each agent interacts with its closest $k = 3$ nearest neighbours. Noise was generated from a generalised Student's $t$-


**Figure 5.12:** Prior beliefs about the number of nearest neighbours $k$ which an agent interacts with in a 2-dimensional environment, expressed by a $\text{Ga}(6, 1/2)$ distribution. Our prior beliefs here are in part informed by the work of Ballerini et al. (2008).

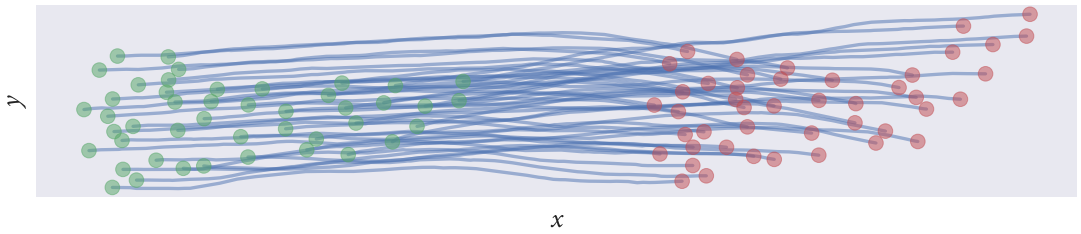



**Figure 5.13:** Data used for our simulation study of the topological model. Each agent was simulated to interact with its nearest $k = 3$ neighbours, and was subject to noise generated from a generalised Student's $t$-distribution with location $\mu = 0$, scale $\sigma_Y = 0.03$ and degrees of freedom $\nu = 7$.

distribution with scale $\sigma_Y = 0.03$ and degrees of freedom $\nu = 7$. The trajectories generated by this simulation are illustrated in Figure 5.13, and resemble that of a real flocking event.

Stan was used to make realisations from the posterior distribution. Four independent chains were initialised at draws from our prior beliefs. Each chain was simulated for 10,000 iterations. The initial 2500 iterations were discarded to allow the chains to converge. The sequences generated by this sampler are shown in Figure 5.14. Corresponding posteriors are plotted in Figure 5.15. Our prior beliefs about the model parameters are overlain in red, and appear flat in comparison to our posteriors: indicating that we have learnt a lot in observing the data. Summaries of this output are tabulated in Table 5.4. Considering the output of this scheme we see that we can accurately recover the true parameter values of data simulated from the topological model.

| Parameter | mean | sd | 5% | 95% | ESS | $\widehat{R}$ |
|---|---|---|---|---|---|---|
| $k$ | 2.98 | 0.02 | 2.94 | 3.02 | 16 710 | 1.0 |
| $\sigma_Y$ | 0.03 | 0.0 | 0.03 | 0.03 | 13 240 | 1.0 |
| $\nu$ | 7.04 | 0.50 | 6.20 | 7.81 | 13 355 | 1.0 |

**Table 5.4:** Tabulating summaries of the posterior draws made in fitting the topological model to simulated data. Posterior samples were generated by Stan's implementation of the No-U-Turn-Sampler. Each row represents a parameter of the topological model. The values of ESS and $\widehat{R}$ computed show that the chains converged and realised a large number of posterior samples.

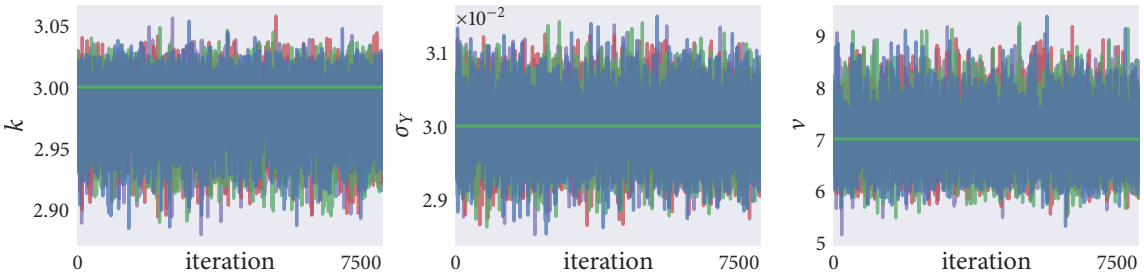



**Figure 5.14:** Markov chains generated in inferring the parameters of the topological model from simulated data. Stan was used to simulate four independent chains constructed to target the posterior distribution. These chains were initialised at different initial conditions. The resulting chains are seen to converge to the same common distribution, and oscillate regularly around the true parameter values.

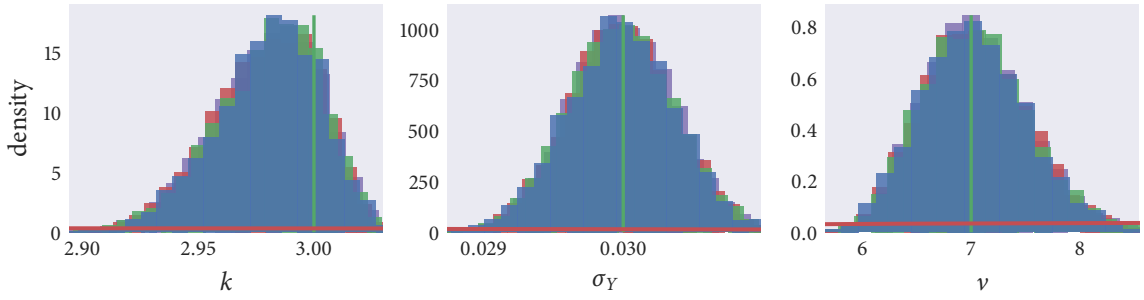



**Figure 5.15:** Histogram plots of our posterior densities having fitted the topological model to simulated data. Our posteriors are seen to have updated considerably from our prior beliefs (overlain in red), and accurately capture the true values (green vertical lines).

## 5.2 Hierarchical models

Hierarchical models represent a more demanding inference problem owing to the increased number of model parameters to infer. The simulation studies performed in Section 5.1 required three parameters to be inferred per model: one interaction parameter and two noise parameters (scale and degrees of freedom). In the hierarchical models introduced in Section 4.2.4 each agent has its *own* interaction parameter, as well as its *own* scale parameter for the noise distribution. For simplicity, the agents share a global degrees of freedom parameter for the noise distribution. To fit a hierarchical model to data describing the movements of $N$ individuals we are then required to infer $2N + 1$ parameters and 4 hyperparameters.

Introducing this extra level of structure into our model needs to be accounted for in the likelihood function. The likelihood function for our differing hierarchical models takes the same general form. For illustration purposes, here we shall consider the likelihood of data generated from a hierarchical Gaussian weighted model. The likelihood function for the remaining hierarchical models can be derived in a similar fashion.

Suppose we simulate the movements of $i = 1, \ldots, N$ individuals over $t = 1, \ldots, T$ time steps, according to a hierarchical Gaussian weighted model. Here, agent $i$ has interaction parameter $\sigma_{X_i}$ and noise-scale parameter $\sigma_{Y_i}$. The interaction parameters are distributed according to a population-level distribution, here taking the form of a gamma distribution with mean $m_X$ and variance $v_X$ (Equation (4.15)). Noise-scale parameters $\sigma_{Y_i}$ are distributed in a similar manner (Equation (4.12)). The likelihood of observing this flock's directional updates, along with the parameters $\sigma_{X_i}$ and $\sigma_{Y_i}$ can be expressed as the product:

$$\begin{aligned} L(\sigma_X, \sigma_Y, v, \langle\theta\rangle_{1:N,1:T-1} \,|\, \theta_{1:N,2:T}, m_X, v_X, m_Y, v_Y) \\ = L(\sigma_Y, v, \langle\theta\rangle_{1:N,1:T-1} \,|\, \theta_{1:N,2:T}) \times L(m_X, v_X \,|\, \sigma_X) \times L(m_Y, v_Y \,|\, \sigma_Y). \end{aligned} \tag{5.5}$$

Taking the form of the likelihood from Equation (5.1) and considering the hierarchy im-

posed in Equations (4.12) and (4.15), we may write down the likelihood in full as:

$$
\begin{aligned}
&L(\sigma_X, \sigma_Y, \nu, \langle\theta\rangle_{1:N,1:T-1} \,|\, \theta_{1:N,2:T}, m_X, \nu_X, m_Y, \nu_Y) \\
&\qquad = \prod_{i=1}^{N}\prod_{t=1}^{T-1} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}\sigma_Y}\left(1 + \frac{1}{\nu}\left(\frac{\theta_{i,t+1} - \langle\theta\rangle_{i,t}}{\sigma_Y}\right)^2\right)^{-\frac{\nu+1}{2}} \\
&\qquad \times \prod_{i=1}^{N} \frac{(m_X/\nu_X)^{m_X^2/\nu_X}}{\Gamma(m_X^2/\nu_X)} \sigma_{X_i}^{m_X^2/\nu_X - 1} \exp\left(-\frac{m_X}{\nu_X}\sigma_{X_i}\right) \\
&\qquad \times \prod_{i=1}^{N} \frac{(m_Y/\nu_Y)^{m_Y^2/\nu_Y}}{\Gamma(m_Y^2/\nu_Y)} \sigma_{Y_i}^{m_Y^2/\nu_Y - 1} \exp\left(-\frac{m_Y}{\nu_Y}\sigma_{Y_i}\right).
\end{aligned}
\tag{5.6}
$$

### 5.2.1 *Continuous models*

In Section 5.1.2 we used Stan to perform parameter inference on data simulated from continuous models, and a random-walk Metropolis–Hastings algorithm to perform inference on the Vicsek model. Similarly, here we shall use Stan to perform parameter inference on data simulated from our *hierarchical* continuous models. However, owing to the increase in dimensionality of the problem, performing inference on the hierarchical Vicsek model with a Metropolis–Hastings algorithm (necessary because of the discontinuous posterior) was found to be computationally infeasible, and is omitted from this study.

#### *Power-law weighted*

Figure 5.16 shows data simulated from the hierarchical power-law weighted model. The interaction parameters of the 45 simulated agents are drawn from a gamma distribution with mean $m_\alpha = 1.2$ and variance $\nu_\alpha = 0.02$. The noise-scale parameters of the agents are drawn from a gamma distribution with mean $m_Y = 0.04$ and variance $\nu_Y = 0.001$. As such, $m_\alpha$, $\nu_\alpha$, $m_Y$ and $\nu_Y$ represent the hyperparameters of this model.

Vague hyperpriors are chosen to represent our prior beliefs about the hyperparameters. These, along with the likelihood function, are combined with Bayes' Theorem to target the posterior distribution. We use Stan's NUTS algorithm to simulate four independent chains to target the posterior. Each chain is initialised as a draw from our prior beliefs and simulated for 5000 iterations. The initial 2500 iterations are discarded to allow the chains to converge from their initial conditions toward the posterior.

Our posterior realisations of the hyperparameters are visualised in Figure 5.17. Each coloured histogram overlain represents the realisations generated by an independent chain. Our hyperpriors are overlain in red. See that in all cases our hyperpriors appear flat in

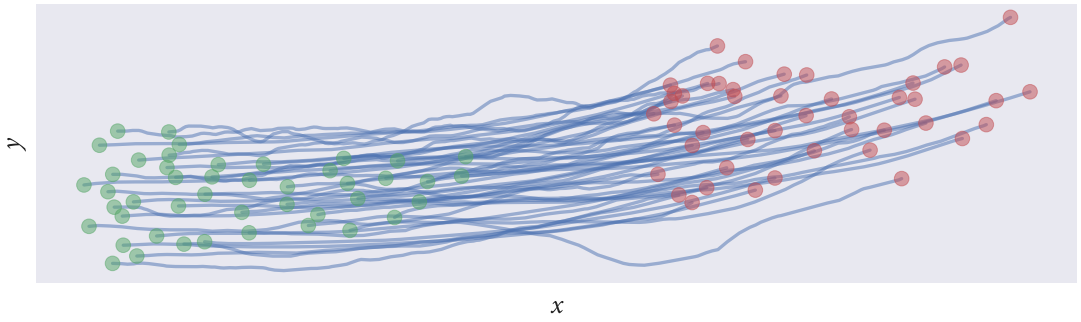



**Figure 5.16:** Forward simulation of the hierarchical power-law weighted model. The movements of $N = 45$ agents are simulated for $T = 200$ time steps. For realism, the initial positions and directions of motions of the agents are taken from an observation of a real flocking event. The interaction parameters $\alpha_i$ are drawn from a gamma distribution with mean $m_\alpha = 1.2$ and variance $v_\alpha = 0.02$. Similarly, the noise-scale parameters $\sigma_{Y_i}$ are realisations from a gamma distribution with mean $m_Y = 0.04$ and variance $v_Y = 0.001$.

comparison to our posteriors. This reflects having updated our beliefs considerably in light of the data. In this plot vertical green lines represent *population*-level summary statistics (the true values of the hyperparameters), and the vertical black lines represent the corresponding *sample*-level summary statistics. For example, consider that although the $\alpha_i$ were drawn from a distribution with population mean $m_\alpha = 1.20$ (vertical green line), their sample mean was $\overline{\alpha} \approx 1.18$ (vertical black line). We see that all the hyperparameters are well-captured by our posterior distributions, and conclude that they can be accurately inferred from simulated data.

Posterior densities about the $N$ interaction and $N$ noise-scale parameters are summarised in Figure 5.18. Here, each boxplot represents a posterior density. As is convention, box boundaries represent the 25th and 75th percentiles of the posterior. Whiskers represent the median $\pm 1.5 \times$ IQR. See that in all but one case ($\alpha_6$), the true parameter values lie within our posteriors. We observe that the $\sigma_{Y_i}$'s are generally well represented by the posterior median. It appears that smaller noise-scale parameters can be captured with less uncertainty than larger noise-scale parameters. Although forty-four of the forty-five interaction parameters are captured by our posteriors, the interaction parameters are not recovered with as much accuracy as the noise-scale parameters. However, we still see that our posteriors generally capture the interaction parameters.

Although not included here for brevity, tabulating split-$\widehat{R}$ for every parameter reveals that all chains converged with $\widehat{R} = 1.00$. Computing the effective sample size for each parameter showed that the *worst* performing chain still realised approximately 10,000 ef-

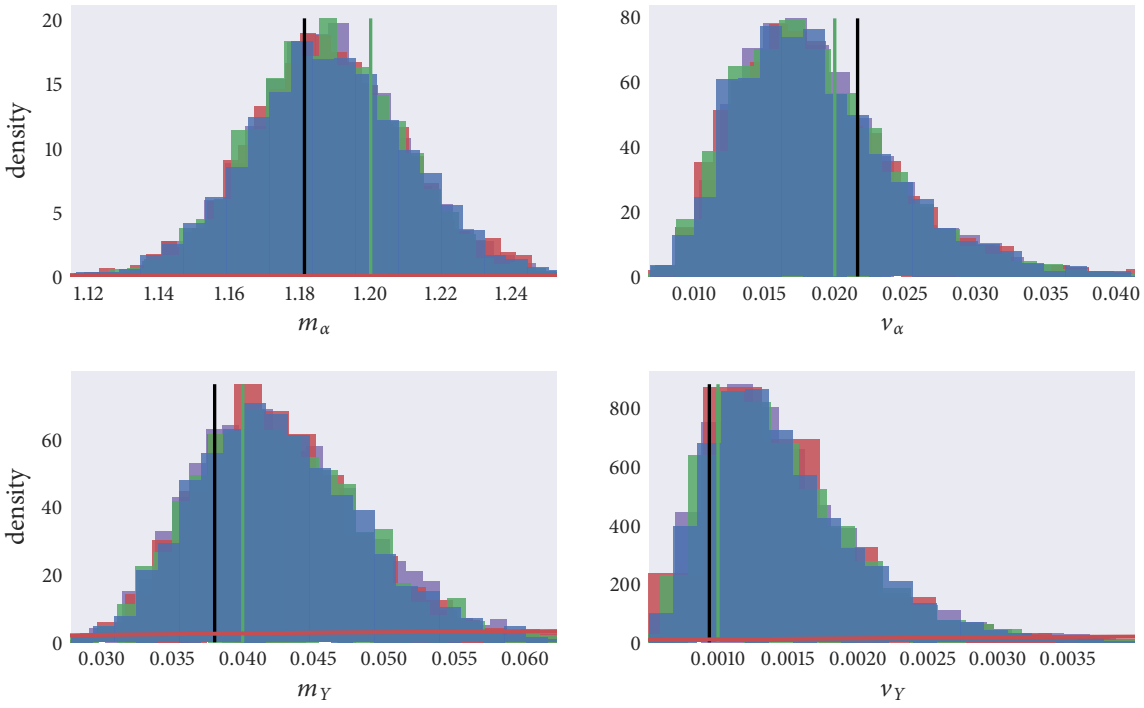



**Figure 5.17:** Histogram plots representing our posterior densities about the hyperparameters of the hierarchical power-law weighted model fitted to simulated data. Vertical green lines represent the true values of the hyperparameters (the *population* means and variances of the parameters). Vertical black lines represent *sample* means and variances of the parameters. Hyperpriors are overlain in red.

fective samples from the posterior. This gives confidence that our posteriors have been well realised.

Considering the posterior samples generated by our Stan model, we have demonstrated that we can accurately recover the hyperparameters and parameters of data simulated from a hierarchical power-law weighted model.

## *Gaussian weighted*

In a similar manner to the simulation study just performed on the hierarchical power-law weighted model, we now seek to perform a simulation study on the hierarchical Gaussian weighted model. As before, our model is implemented to simulate the movements of 45 agents over 200 observations. Here, the interaction parameters $\sigma_{X_i}$ are drawn randomly from a gamma distribution with mean $m_X = 25$ and variance $v_X = 1$. The forty-five $\sigma_{X_i}$'s sampled for this simulation have mean $\overline{\sigma_X} \approx 25.2$ and variance $\mathrm{Var}(\sigma_X) \approx 1.1$. The trajectories realised by this simulation are illustrated in Figure 5.19.

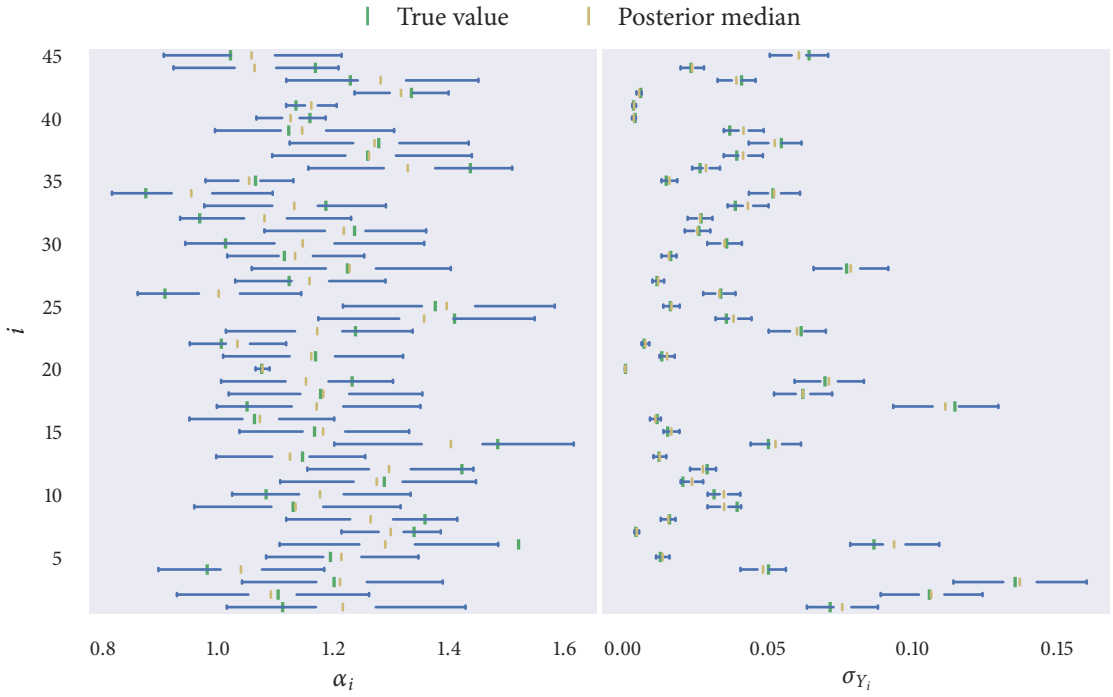



**Figure 5.18:** Boxplots summarising our posteriors about the interaction and noise-scale parameters of the hierarchical power-law weighted model fitted to simulated data. As is convention, boxes show the 25th, 50th (the median) and 75th percentiles of the posterior. Whiskers represent the median $\pm 1.5 \times$ IQR.

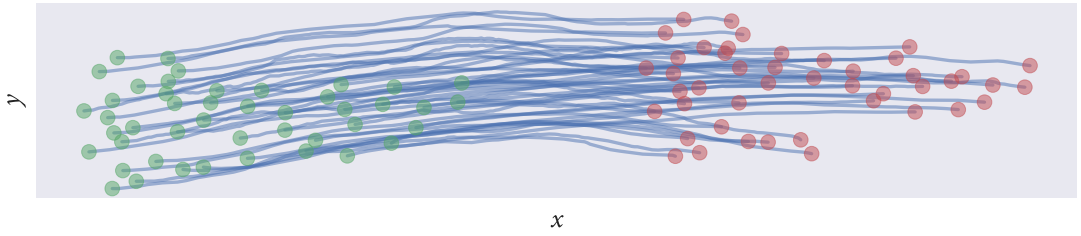



**Figure 5.19:** Data generated by a simulation of the hierarchical Gaussian weighted model. The interaction parameters $\sigma_{X_i}$ are sampled from a gamma distribution with mean $m_X = 25$ and variance $v_X = 1$. Similarly, the noise-scale parameters are drawn from a gamma distribution with mean $m_Y = 0.04$ and variance $v_Y = 0.001$.

We use the Stan modelling language to codify Equation (5.5) and express our vague prior beliefs about the hyperparameters of the model. Four independent chains are again simulated for 5000 iterations, with the first 2500 iterations discarded to allow a warm up period.

Our posterior realisations about the hyperparameters are visualised in Figure 5.20. The four independent chains simulated are seen to have converged to a common distribution.

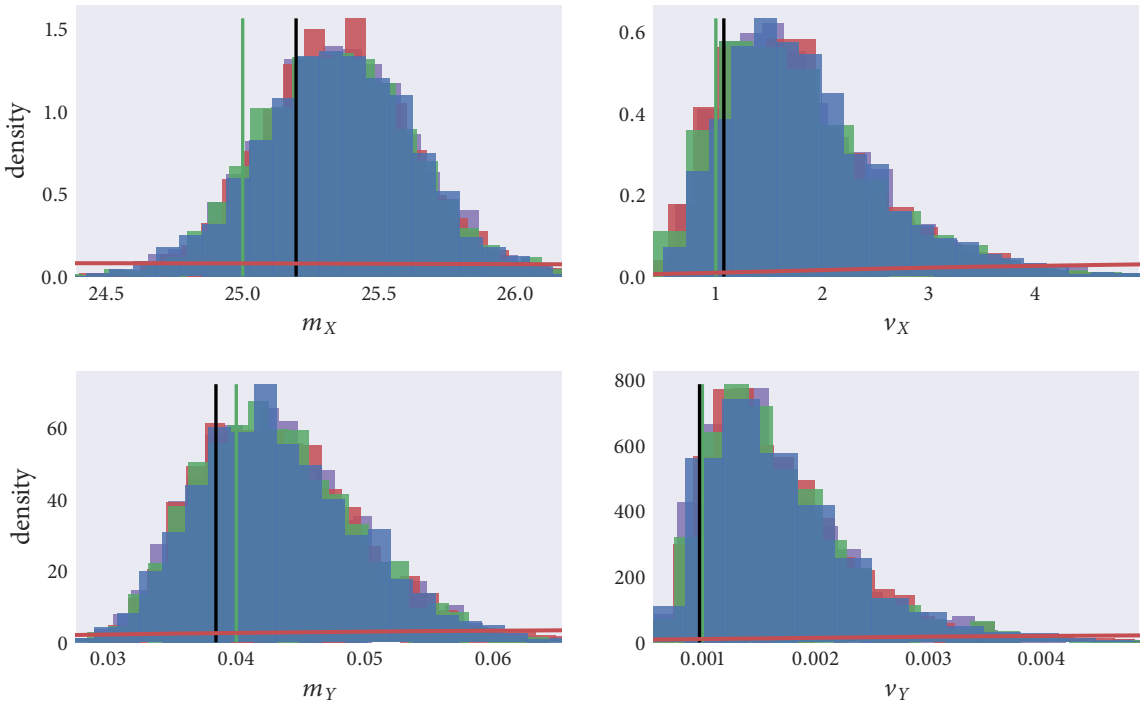



**Figure 5.20:** Realisations from the posterior distribution of the hierarchical Gaussian weighted model. Vertical green lines show the true values of the hyperparameters. Vertical black lines show the sample summary statistics of the model parameters. Hyperpriors are overlain in red.

The true values of the hyperparameters (vertical green lines) lie well within our posteriors, and lie close to the posterior mode.

Figure 5.21 is constructed to provide a convenient summary of our posterior distributions about the $2 \times N = 90$ model parameters used for this simulation. The true parameter values are shown by green markers, and posterior medians are represented by yellow markers. All the true parameter values are seen to lie within our posteriors, indicating that they can be captured by parameter inference. Our posteriors about the interaction parameters of the agents overlap considerably. This indicates that the parameters drawn for simulation represent very similar interaction weightings. The noise-scale parameters, on the other hand, are captured with much less uncertainty and overlap.

### 5.2.2 *Topological model*

Having satisfied ourselves that we can fit continuous hierarchical models to simulated data, we now turn our attention to our hierarchical topological model. The data simulated for this study is shown by the trajectories of Figure 5.22. The trajectories generated for this

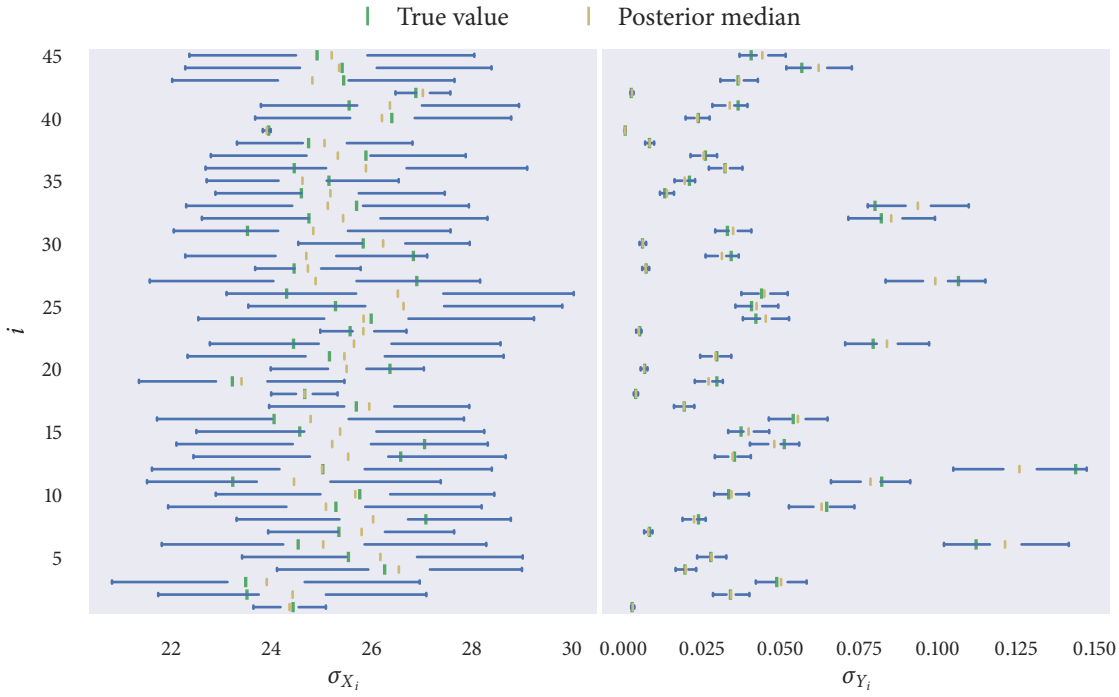



**Figure 5.21:** Boxplots summarising the posterior densities about the interaction parameters and the noise-scale parameters used to forward simulate the hierarchical Gaussian weighted model. Each boxplot illustrates the posterior median, lower and upper quartiles, and the posterior median ±1.5 × IQR.

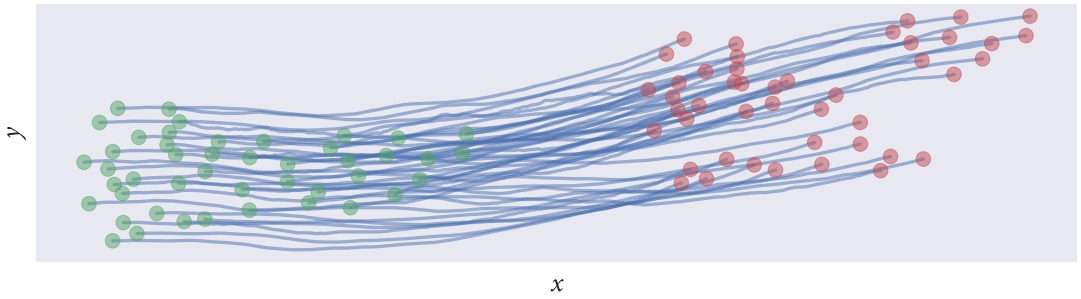



**Figure 5.22:** An illustration of the trajectories realised by a forward simulation of the hierarchical topological model. In this model agent $i$ interacts with its $k_i$ nearest neighbours, and experiences noise distributed according to a Student's $t$-distribution with scale $\sigma_{Y_i}$ and degrees of freedom $\nu$. Interaction and noise scale parameters are drawn from population level distributions, parameterised by the hyperparameters $m_k = 1$, $\nu_k = 2$, $m_Y = 0.4$ and $\nu_Y = 0.001$.

simulation bear a strong resemblance to that of a real flocking event (see, eg. Figure 6.2).

Interaction and noise-scale parameters were drawn from population-level distributions. The interaction parameters $k$ were sampled from a Gamma distribution with mean $m_k = 4$

and variance $\nu_k = 2$. These hyperparameters are represented in Figure 5.23 by vertical green lines. The sampled interaction parameters had mean $\overline{k} \approx 4.15$, and variance $\mathrm{Var}(k) \approx 1.25$. These values are illustrated in Figure 5.23 with vertical black lines. In a similar manner, the noise-scale parameters are drawn from a Gamma distribution with hyperparameters $m_Y = 0.4$ and $\nu_Y = 0.001$, which resulted in parameters $\sigma_Y$ with sample mean $\overline{\sigma_Y} \approx 0.035$ and sample variance $\mathrm{Var}(\sigma_Y) \approx 0.001$.

The posterior samples realised by Stan's implementation of the No-U-Turn-Sampler are shown in Figure 5.23. These histograms represent the posterior realisations made by four independent Markov chains, each simulated for 5000 iterations and initialised at different locations. Sample and population level summary statistics of the parameters are shown to be captured well by our posterior densities, indicating that we can successfully recover the hyperparameters of simulated data by inference. Our posteriors are much more peaked than our hyperpriors, suggesting that our choice of hyperprior had little influence over our posteriors, as was intended.

Having seen that we can successfully recover the hyperparameters of the hierarchical topological model, we now turn our attention to the *parameters* of this model. The posterior densities of the remaining $2N$ parameters are summarised in Figure 5.24. From this plot we see that we are able to capture all 90 of the model parameters in our posterior densities. The true values of the parameters are seen to lie close to the posterior median in the majority of cases.

## Conclusions

Having shown that we can accurately capture the true parameter values of data simulated from the hierarchical topological model, we have now demonstrated that we can perform successful simulation studies on *all* our *continuous* hierarchical models. In Section 5.1 we showed that we can capture the true parameter values of data simulated from our global models (both and continuous *and* discontinuous).

The inference schemes constructed to perform these inferences can now be repurposed with minimal-to-no-modification for use with data of real flocking events. We can be confident in the accuracy and our implementation of these schemes. With this, if we find our implementations unable to fit real data to our models, we can be confident that this is due to a discrepancy between model and data, rather than an error in our implementation.

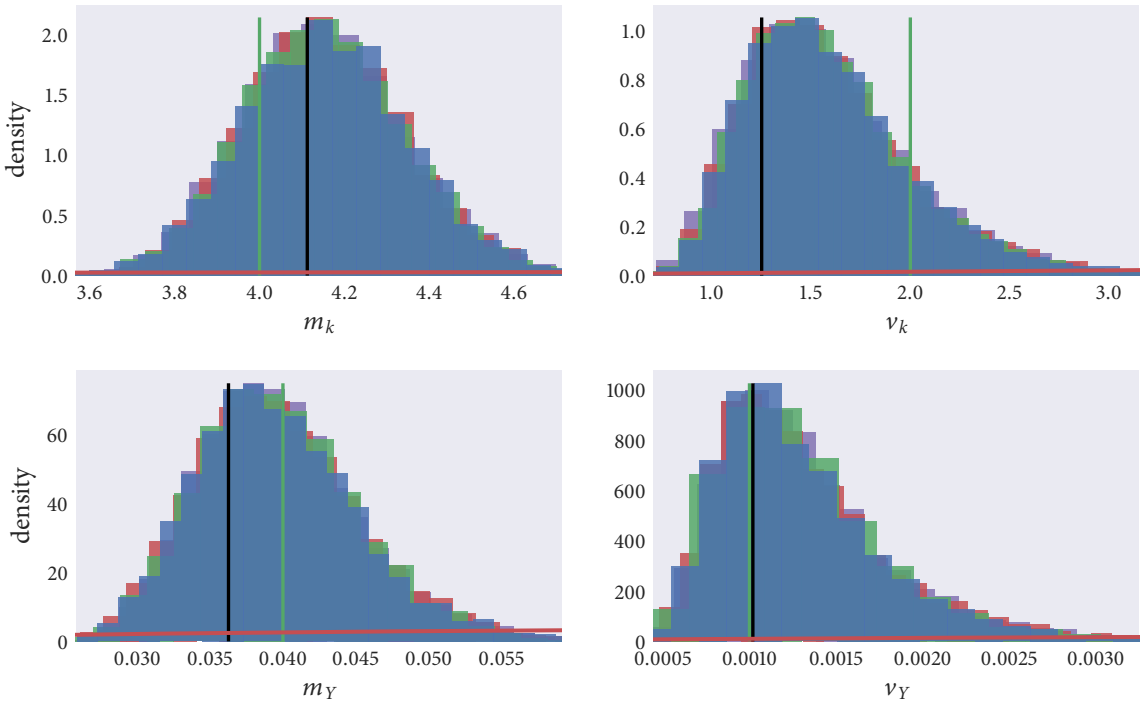



**Figure 5.23:** Histogram plots showing the posterior samples drawn in fitting the hierarchical topological model to simulated data. Samples were drawn by Stan's implementation of the NUTS algorithm. Our hyperpriors (overlain in red) appear flat in comparison to our posteriors: indicating that our beliefs have updated considerably having observed the simulated data. The true values of the hyperparameters are shown by the vertical green lines, and the corresponding sample level summary statistics are shown by black vertical lines.

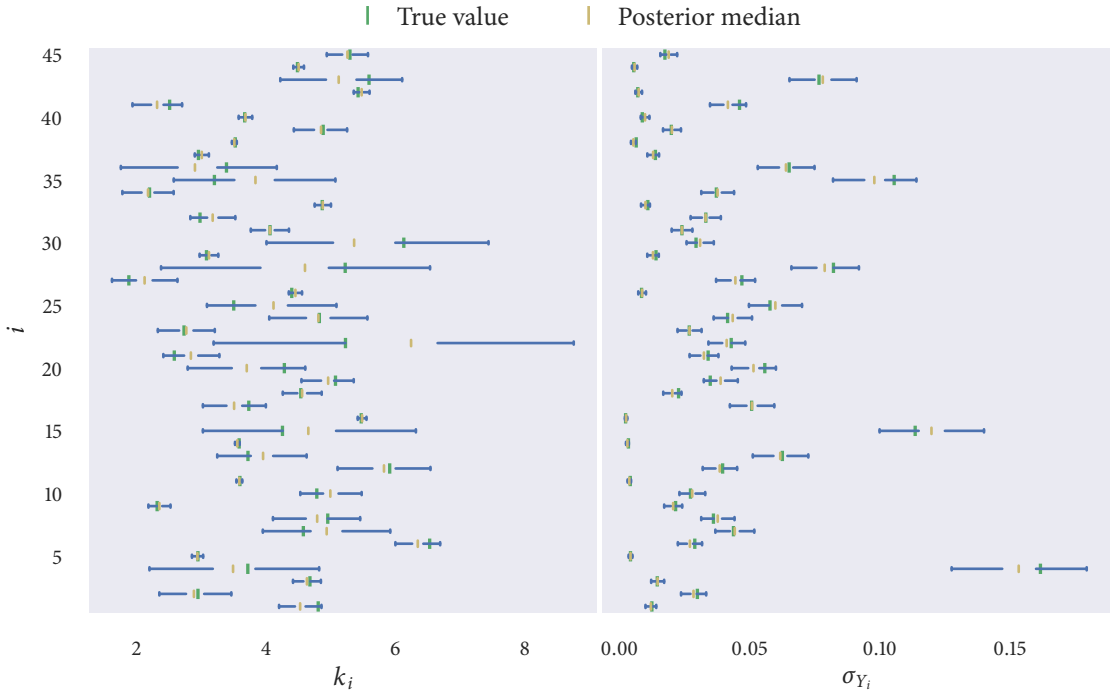



**Figure 5.24:** Summarising our posteriors about the interaction and noise-scale parameters of the $N = 45$ agents simulated according to the hierarchical topological model. Box boundaries represent the upper and lower quartiles of our beliefs, and the whiskers show the posterior median $\pm 1.5 \times$ IQR.

# 6

# *Case study I: flocking sheep*

In Chapter 5 we showed that it is possible to infer the parameters of a number of agent-based models from simulated data. Performing simulation studies before attempting inference on real data is advisable as it allows an opportunity to troubleshoot and assess the accuracy of our inference schemes. The schemes constructed to perform inference on simulated data can now be reused with little-to-no-alteration to perform inference on real data. With the confidence instilled by the success of our simulation studies, if we find our schemes unable to fit a model to *real* data, we can be confident that this is because of a discrepancy between the model and data, rather than an error in our fitting process or our implementation of it.

This chapter shall focus on data of flocking sheep, unseen before in the literature. Aerial footage of flocking sheep was recorded by a commercially available drone. The position of each individual sheep was extracted from the captured video footage using custom tracking software. In total, three flocking sequences were extracted and analysed. The flocking events were captured and their data extracted by Hayley Moore (thesis in preparation). Global models considered in Section 5.1 will be fit to these sequences. The resulting fits will be assessed for validity and ranked for performance by Akaike Information Criteria (Section 3.3.1).

## 6.1 Flocking data

Aerial footage of flocking sheep was captured with a commercially available DJI Phantom 3 drone equipped with a built in high-definition camera, representing a frame resolution

of 4000 × 3000 pixels. This video footage was recorded at a rate of 24 frames per second. In total, three distinct flocking events were captured and analysed. Each flocking event involved the same flock of 45 sheep.

The recorded flocking events took place in large open fields: a natural environment with which the sheep were already familiar. Flocks were recorded in familiar terrain in an attempt to observe the most natural flocking events, and to minimise the effect which an unfamiliar environment may have on the flock dynamics. In some cases the flocking events were initiated by the movements of a quad bike, but were then left to develop naturally and unprompted. Events generally took place away from hard boundaries, such as fences and trees, but the influence of these objects from a distance cannot be ruled-out.

Custom tracking software was constructed to determine the position of each sheep in every video frame. To correct for the movement of the drone due to external influences, such as wind, the resulting frames were transformed with respect to some reference points. Reference points are stationary and identifiable features present in all of the captured frames. Example reference points include field boundaries, fence posts and trees.

Having corrected for extraneous drone-movement, the captured frames were then thresholded to separate the pixels representing sheep from their surrounding environment. Kalman filtering (Kalman 1960) and the Hungarian algorithm (Kuhn 1955) were then used to track the positions of individual sheep between frames. Having linked the positions of sheep between frames, the trajectories of motion of every individual could be reconstructed. Figure 6.1 shows the positions of sheep in a single video frame, overlain are the positions of these sheep in previous frames.

Having realised the trajectories of each individual it is then a simple process to determine the velocity, speed and direction of motion of each individual. Using Equation (4.5) it is possible to compute the mean polarisation of the flocks in each sequence. The attributes of each flocking event are summarised in Table 6.1. We see that sequences are broadly similar: representing highly-polarised flocks moving at similar speeds.

Figures 6.2 to 6.4 show the trajectories of motion of each individual in the three recorded flocking events. The positions of sheep in the first frame of the captured footage are represented by green markers. The sheep then travel along the blue lines, with their last recorded position illustrated by the red markers.

The data realised by the drone captured footage resembles the data used in the simulation studies of Chapter 5. In fact, the first observed frame of sequence 1 was used to seed these forward simulations.

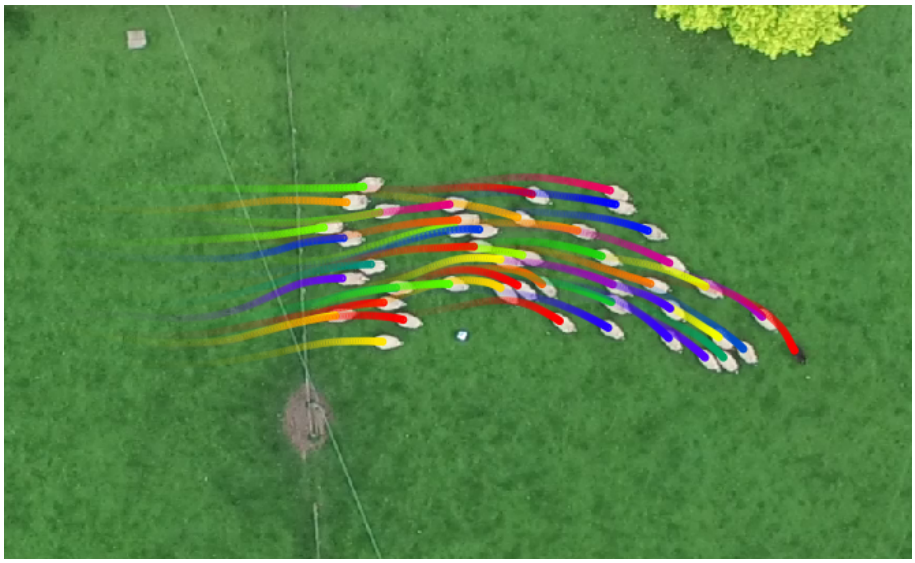


**Figure 6.1:** A visualisation of data realised from a single flocking event. The faint vertical line seen passing through the image shows a boundary fence which was located here previously. The position of each sheep in every video frame was extracted using custom tracking software. Coloured lines represent the positions of sheep in previous frames. Each flocking event tracked the movements of 45 sheep over close to 200 frames. A single flocking sequence then represents approximately $45 \times 200 = 9000$ observations. Footage was recorded at a rate of 24 frames per second, and so each sequence represents 10 seconds of raw footage.

| Sequence | Frames | Sheep | Mean speed (px $s^{-1}$) | Mean alignment |
|---|---|---|---|---|
| 1 | 192 | 45 | 4.20 | 0.99 |
| 2 | 183 | 45 | 3.26 | 0.98 |
| 3 | 194 | 45 | 4.25 | 0.97 |

**Table 6.1:** Summaries of the three flocking sequences analysed in this chapter. Each sequence involved a flock of the same 45 sheep, recorded over around 200 frames. As the data was recorded at a rate of 24 frames per second, each sequence represents around 10 seconds of raw footage. The observed flocks were all highly polarised and moved at similar speeds.

## 6.2 Model fitting

We now proceed to fit the global agent-based models introduced in Chapter 4 to the three flocking events realised. The performance of these fitted models is then assessed by the Akaike Information Criteria. Model adequacy is inspected with standardised residual plots and posterior predictive checking.

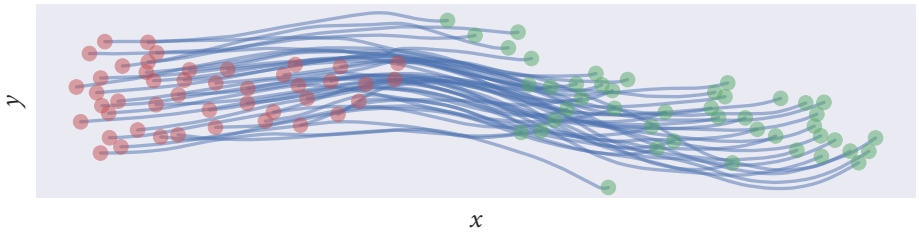



**Figure 6.2:** The trajectories of sheep reconstructed from video footage of sequence 1. The realised flock shows 45 individuals moving cohesively over 192 frames. The observations in the first frame of this sequence were used to seed the forward simulations in Chapter 5.

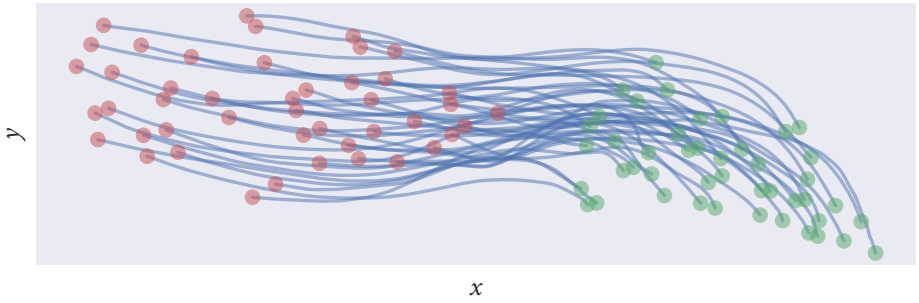



**Figure 6.3:** A trajectory plot of the flock realised in sequence 2. These trajectories correspond to those shown in the drone footage of Figure 6.1.

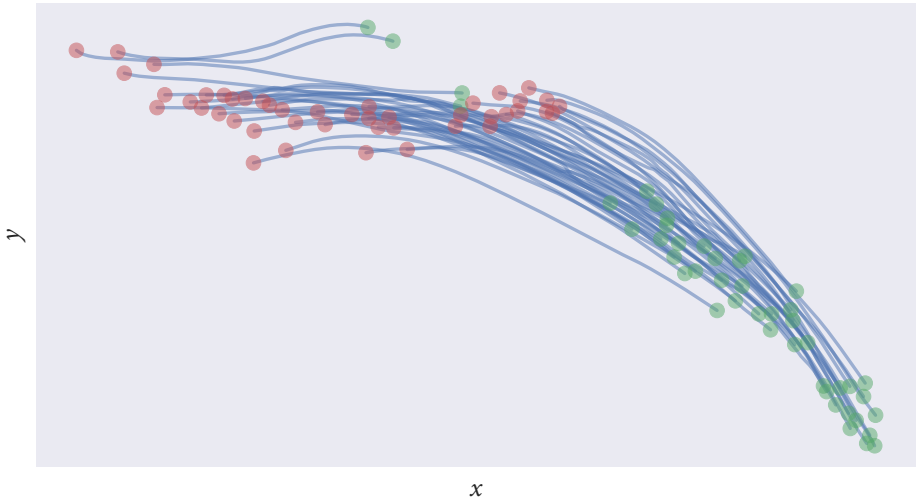



**Figure 6.4:** Illustrating the movement of the flock in sequence 3. Although this flock appears as cohesive as the events in sequences 1 and 2, this flock is also seen to go through greater directional changes.

### 6.2.1 *Sequence 1*

Sequence 1 represents the most cohesive flocking event of the captured sequences. We shall use the inference schemes constructed in Chapter 5 to fit our candidate models to this data. With the models fitted, we can then assess the performance of the model fits, and rank the candidate models by their predictive performance.

#### *Posterior beliefs*

Our posterior beliefs about the parameters of the candidate models fitted to sequence 1 are shown in Figure 6.5. The same vague prior beliefs that were specified in Chapter 5 were reused here. These prior beliefs are plotted in red on top of our posteriors. For all the inferred parameters our prior beliefs appear flat in comparison to our posteriors. This shows that we were able to learn a lot from the data, and that our prior beliefs did not have a strong impact on our posterior realisations.

Stan's No-U-Turn-Sampler was used to draw samples from the posterior distributions of the Null, power-law weighted, Gaussian weighted and topological models. To perform the model fittings we initialised four independent Markov chains at draws from our prior beliefs. Chains were simulated for 10,000 iterations, with the first 5000 iterations discarded to allow convergence. The samples drawn by each independent chain are represented by the different coloured histograms of Figure 6.5. The overlain histograms show that the chains all converged to the same common distribution. Computing values of $\widehat{R}$ also indicate that our chains have converged.

The parameters of the Vicsek model were inferred by implementing a random walk Metropolis–Hastings sampler. This algorithm was simulated for $10^6$ iterations, with the first half of the chains discarded to allow a burn-in period. Our posterior distributions about the interaction radius $r$, shown in Figure 6.5(b), show a very strong negative skew. As our posterior density rapidly drops to zero as $r \to 3.25$, we conclude that at $r \approx 3.25$ an agent suddenly includes the influence of a neighbour which the model finds incompatible. This behaviour is a product of the discontinuity of Vicsek's weighting rule.

The interaction parameters inferred from sequence 1 all represent weak interactions. This finding is consistent with the observation that weak alignment interactions are sufficient to maintain groups which are already highly polarised (Jhawar et al. 2020). The small values inferred for the parameter $\nu$, representing the degrees of freedom of the Student's $t$-distribution, suggests evidence of non-normality of the noise. This result is significant as thus far the assumption of normally distributed noise has been a mainstay of agent-based

modelling.

Although we have shown that our candidate models *can* be fit to sequence 1, this does not say anything about their adequacy or goodness-of-fit. To assess goodness-of-fit we shall inspect residual plots and make posterior predictive checks.

## *Standardised residuals*

Assessing plots of residuals can be used as an informal method to assess modelling assumptions. To compare model and residual, we first rearrange Equation (4.6) to see that:

$$(\theta_{i,t+1} - \langle\theta\rangle_{i,t})/\sigma_Y \,|\, \nu \sim t_\nu, \tag{6.1}$$

where $t_\nu$ is the standardised Student's $t$-distribution with $\nu$ degrees of freedom. To realise a model's residuals we compute $\langle\theta\rangle_{i,t}$ for $i = 1, \ldots, N$ and $t = 1, \ldots, T$ for every posterior sample made by our inference scheme. The residuals can then be computed as $\theta_{i,t+1} - \langle\theta\rangle_{i,t}$ and standardised by dividing by the posterior mean of the scale parameter $\sigma_Y$. The residuals are compared with the model graphically by overlaying a standardised Student's $t$-distribution with the degrees of freedom parameter taken from our posterior mean.

Figure 6.6 shows the standardised residuals of our fitted models. We see that all the models look to give a good fit to the data. It may come as a surprise that the Null model looks to perform so well. However, this is again a product of the observation that our flock was already highly polarised, and that only weak alignment interactions are necessary to maintain a flock which is already cohesive.

Although plots of standardised residuals provide a useful informal measure of model fit, it can be useful to use more formal methods such as information criteria to quantify fit.

## *Model rankings*

Akaike Information Criteria (AIC) favours models which better explain the given data, expressed through a larger likelihood, but penalises model complexity—quantified as the number of model parameters. The Akaike weight of a model can be computed from its AIC value. Typically a model's Akaike weight is interpreted as the probability that the model will make the best predictions on new data, conditional on the set of models considered.

In Table 6.2 we show the value of AIC computed for each model fit. Recall that *smaller* values of AIC indicate a *better* estimate of predictive performance. Table 6.2 shows the ranking of each model fit. We see that the topological model was ranked as the best per-

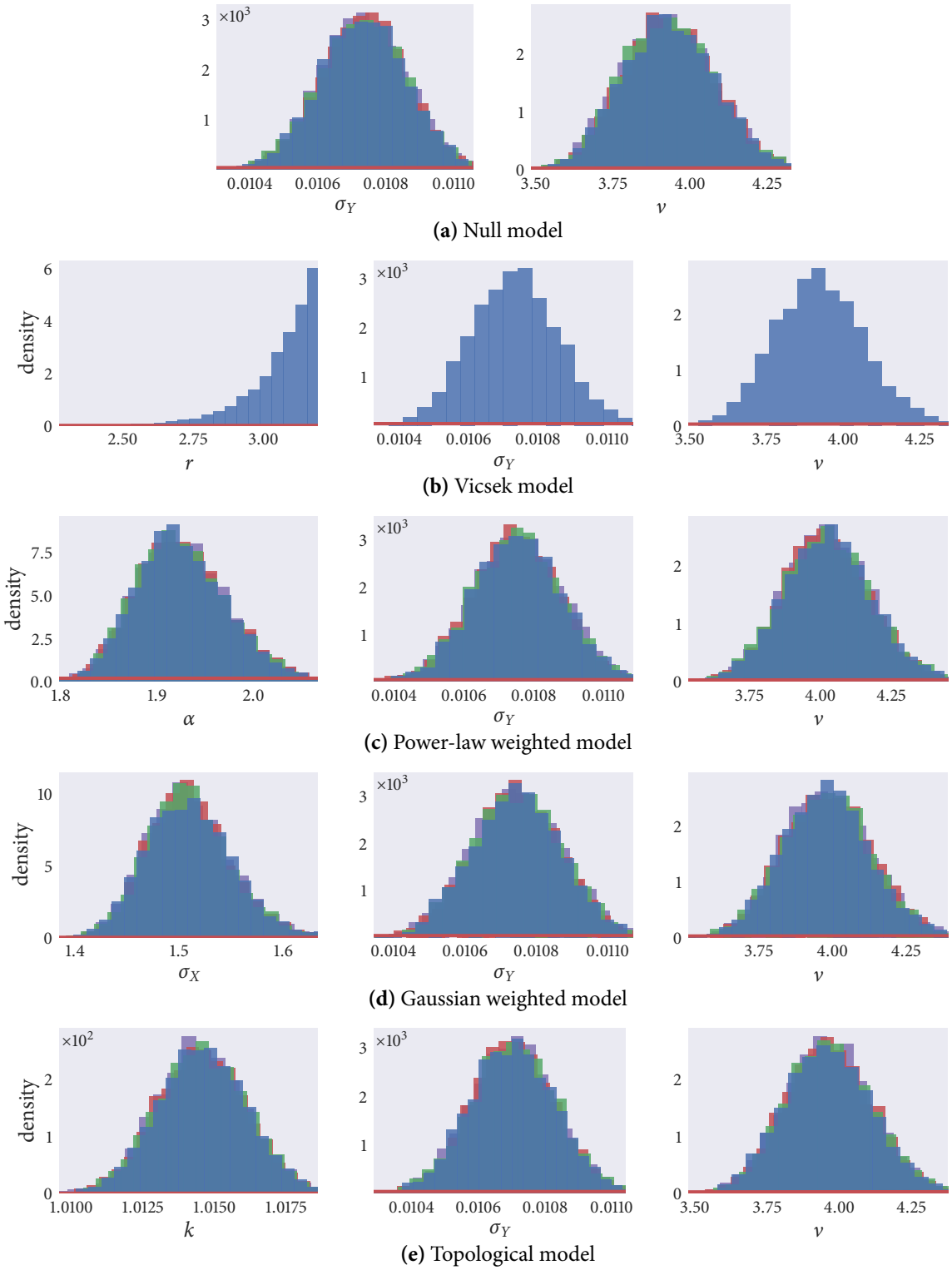



**Figure 6.5:** Histogram plots show the posterior samples realised in fitting candidate models to the data of sequence 1. Our priors beliefs are overlain in red, and in all cases appear flat in comparison to our posterior beliefs; this indicates that the observed data has been very informative.

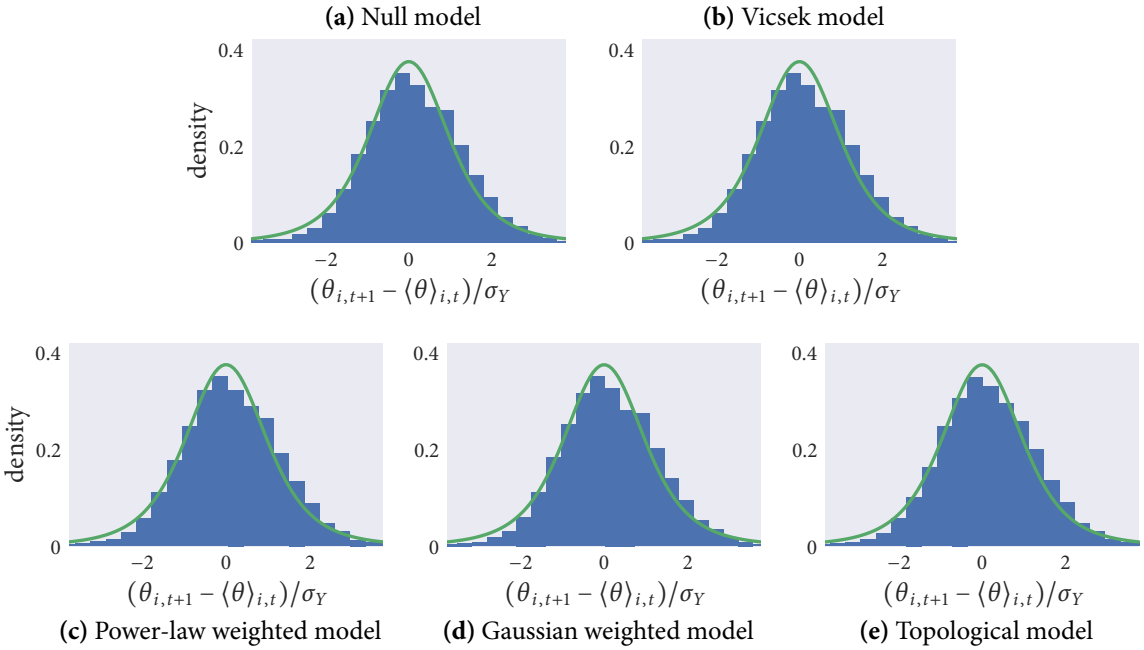



**Figure 6.6:** Standardised residuals of the models fitted to sequence 1. The residuals are standardised by dividing by the posterior mean of the scale parameter $\sigma_Y$. A Student's $t$-distribution with $\nu$ degrees of freedom taken from our posterior mean is overlain in green. We see that models all look to provide a good fit to the data.

| Model | AIC | rank | pAIC | dAIC | weight |
|---|---|---|---|---|---|
| Topological | -49 031 | 1 | 3 | 0.0 | 0.98 |
| Power-law | -49 023 | 2 | 3 | 7.4 | 0.02 |
| Gaussian | -48 985 | 3 | 3 | 46.3 | 0.00 |
| Null | -48 943 | 4 | 2 | 87.7 | 0.00 |
| Vicsek | -48 941 | 5 | 3 | 89.7 | 0.00 |

**Table 6.2:** Tabulated values of AIC and the corresponding Akaike weights (displayed to 2 decimal places) of each model fitted to sequence 1. Models are ranked from 1–5, where 1 is the best fit and 5 is the worst.

forming model, followed by the power-law and Gaussian weighted models. The Null and Vicsek models are ranked as the worst performing. If we were to only compare maximised likelihoods we would rank the Null and Vicsek models the same. However, AIC penalises for model complexity, and so having an extra model parameter without being better at describing the data: the Vicsek model is ranked below the Null model.

### *Posterior predictive checks*

Forward simulating candidate models with parameters drawn from our posterior beliefs provides an informal method to assess how well a model can reproduce realistic behaviours. Here we forward simulate each candidate model one thousand times. To initialise each simulation we randomly draw model parameters from our posterior density. We simulate the movement of 45 sheep, initially positioned and directed as in the first frame of the fitted sequence, for 192 time steps. We compute the alignment of the flocks at each time step.

Figure 6.7 shows the alignment of the simulated flocks. The top panel shows the alignment of two sets of randomly selected simulations. We see that all the models do a reasonable job of capturing the observed flock alignment for around the first 25 time steps. After this, the observed flock experiences a fluctuation in alignment over a time scale of around 100 frames. The candidate models do not appear capable of producing flocks with such long-time fluctuations in alignment. However, this should not come as a particular surprise as these models only predict direction of motion one step ahead. The disparity between observed alignment and simulated alignment may suggest the presence of interactions or external influences not accounted for by our models, or of a longer range dependence on previous directions.

The bottom panel of Figure 6.7 shows the median alignment of simulated flocks as a function of time. The coloured bands around the medians show the upper and lower quartiles of the realised alignments. We see that the topological and Gaussian models, favoured by AIC, produce the most consistently cohesive flocks.

## 6.2.2 *Sequence 2*

This sequence details the movements of our flock of 45 sheep over 183 frames. The recorded flock is again very coherent and highly polarised.

### *Posterior beliefs*

Parameter inference is again performed by a combination of the Metropolis–Hastings and NUTS algorithms. The posterior realisations made by these samplers are shown in Figure 6.8. Having overlain our prior densities in red, we can see that our beliefs have changed considerably in light of the data.

The inferred parameters are similar to those realised from sequence 1; representing weak interactions between individuals and giving evidence of agents experiencing non-normally

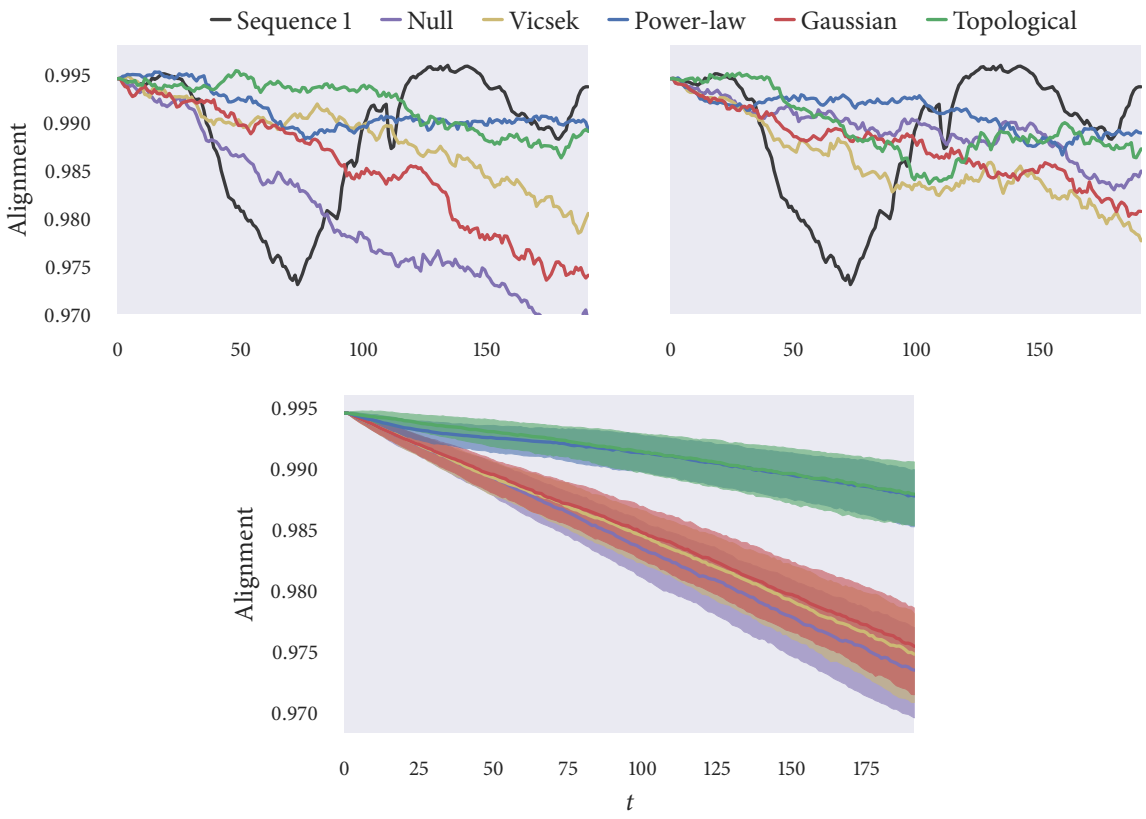


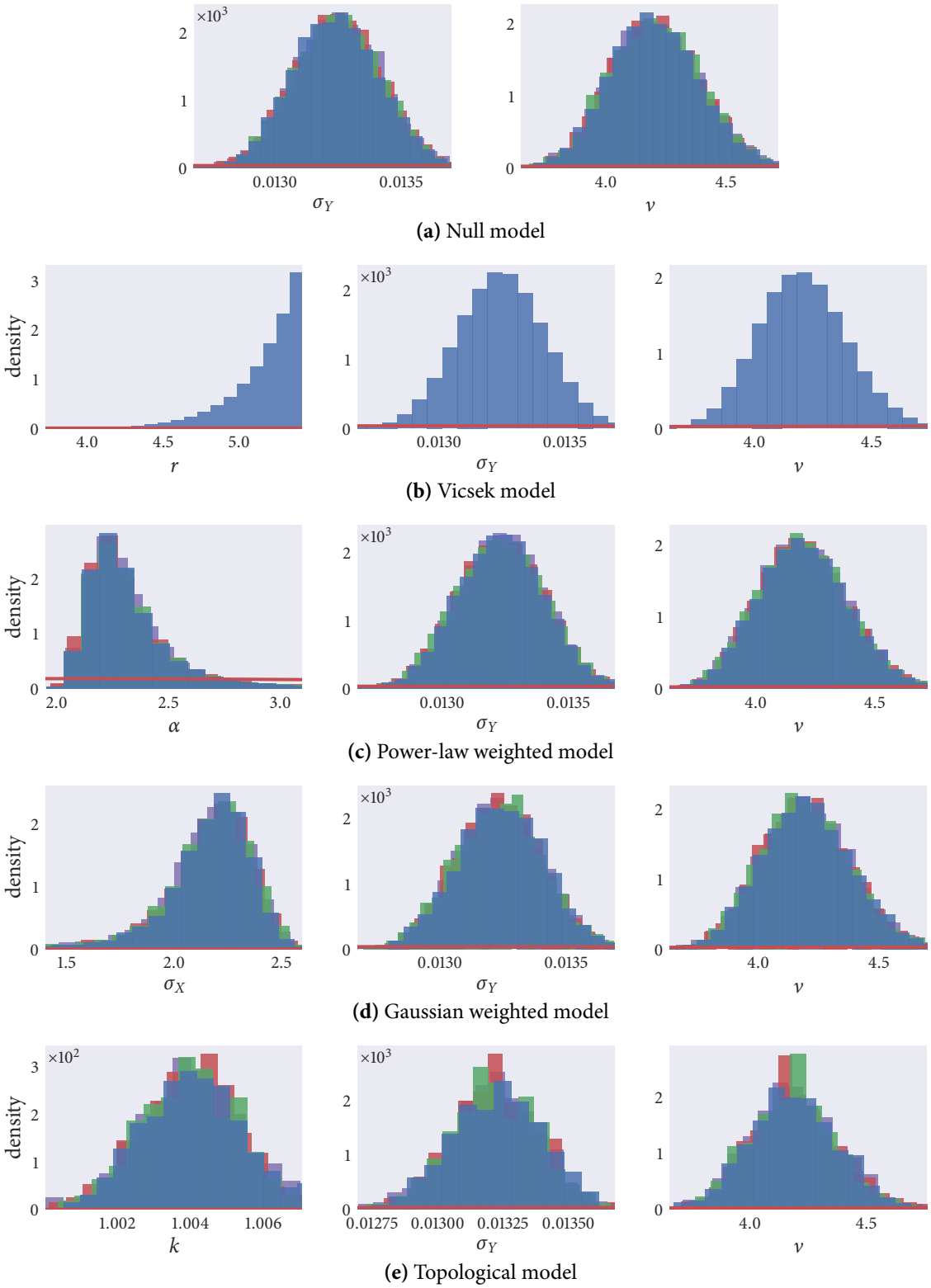

**Figure 6.7:** The alignment of flocks forward simulated with parameters drawn from our posterior densities about model fits to sequence 1. Simulations were initialised as in the first frame of the observed data, and repeated one thousand times. The top panels show two randomly selected sets of forward simulations. The bottom panel shows the median alignment across all simulations. Bands represent the upper and lower quartiles over the one thousand repetitions.

distributed noise. Similar noise-scale and degrees of freedom parameters are inferred for every model fitted to sequence 2.

Our posterior densities about the interaction radius of the Vicsek model again shows a strongly skewed distribution. It is this sudden drop in posterior density which makes the NUTS algorithm ineffective for this problem.

## *Standardised residuals*

We use plots of standardised residuals to compare observation and model prediction. From Equation (6.1) we expect the standardised difference between prediction and observation to follow a Student's $t$-distribution with $\nu$ degrees of freedom. Our observations are the directions of motion $\theta_{i,t+1}$, and our predictions are given by the interaction $\langle\theta\rangle_{i,t}$ computed




**(a)** Null model

**(b)** Vicsek model

**(c)** Power-law weighted model

**(d)** Gaussian weighted model

**(e)** Topological model

**Figure 6.8:** Posterior realisations of the model parameters inferred in fitting candidate models to sequence 2. Posterior samples of the parameters of the Vicsek model were drawn by the implementation of a Metropolis–Hastings sampler. The remaining models were fitted with the implementation of Stan's NUTS algorithm.

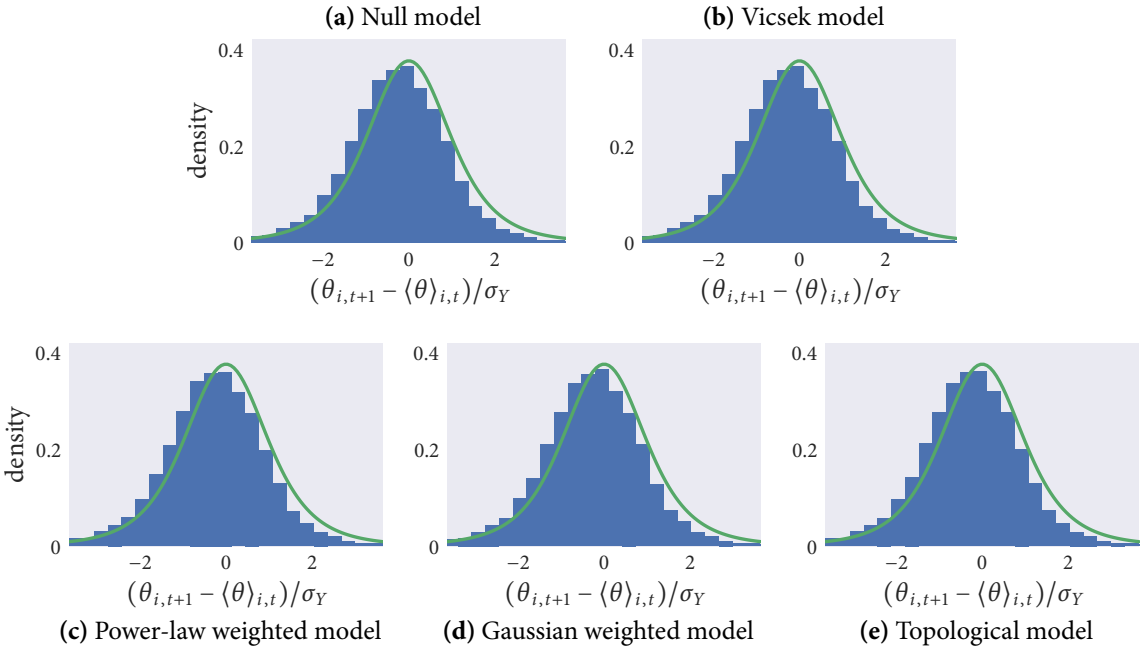



**Figure 6.9:** Standardised residuals of models fitted to sequence 2. The density overlain in green represents a Student's $t$-distribution with $\nu$ degrees of freedom, where $\nu$ is taken as the posterior mean from the corresponding model fit. The distributions of residuals show a mean slightly below zero, suggesting the presence of behaviours not accounted for in our models.

at each posterior sample.

The standardised residuals computed from each model fit are shown in Figure 6.9. The standardised residuals look to follow a generalised Student's $t$-distribution with scale 1, and location slightly below zero. The non-zero mean of the residuals may suggest the presence of additional interactions such as attraction or repulsion, which are not implemented in our models. The non-zero mean could also be explained by the influence of external factors, such as field boundaries, which are not accounted for by our models.

The residual plots for each model fit look very similar. This is again a product of the observation that our flocks are already highly aligned, experience only small fluctuations in directional change, and only require weak alignment interactions to maintain cohesion.

## *Model rankings*

The standardised residuals of each model fit shown in Figure 6.9 look very similar. Recall that model predictions are made at every iteration of our inference schemes. For samples

| Model | AIC | rank | pAIC | dAIC | weight |
|---|---|---|---|---|---|
| Power-law | -43 495 | 1 | 3 | 0.0 | 0.49 |
| Topological | -43 494 | 2 | 3 | 0.2 | 0.44 |
| Gaussian | -43 490 | 3 | 3 | 4.5 | 0.05 |
| Null | -43 488 | 4 | 2 | 6.9 | 0.02 |
| Vicsek | -43 486 | 5 | 3 | 8.9 | 0.01 |

**Table 6.3:** Ranking the models fitted to sequence 2 by their predictive performance, estimated by AIC. The power-law model is ranked as the best performing model, followed closely by the topological model.

drawn using Stan we have 4 independent Markov chains, simulated for 5000 post-warm-up iterations, giving a total of 20,000 iterations. Sequence 2 contains $45 \times 183$ data points. Given the initial directions of individuals, we are then tasked with predicting $45 \times 182 = 8190$ data points. Making our predictions at every iteration, we make a total of $20,000 \times 8190$ predictions per model. These predictions are then used to form our standardised residuals. As such, it can be very difficult to visually ascertain small improvements in model performance from plots of residuals alone. Because of this, more quantitative measures such as AIC can be useful in distinguishing between models.

Computed AIC values and their corresponding model ranks are tabulated in Table 6.3. The computed Akaike weights are not as conclusive as they were with sequence 1; here the power-law and topological models are awarded similar Akaike weights. The similarity in weights computed for these models suggests that in application there may be very little practical distinction between them, or that their differences are so minute it can be difficult to tell them apart.

Below the ranking of the power-law weighted and topological models, the ranking follows the same order as was seen in sequence 1. Again we see that the Vicsek model was not able to explain any more of the data than the Null model. But as AIC favours simplicity, given that both models have the same value of the maximised likelihood function, we see the Vicsek model ranked below the Null model.

### *Posterior predictive checks*

To assess the ability of our fitted models to produce realistic looking behaviours, we forward simulate the candidate models using parameters drawn from our posterior densities. Simulations are initialised as in the first frame of sequence 2, and repeated one thousand times. For each repetition new parameter values are drawn from the posterior.

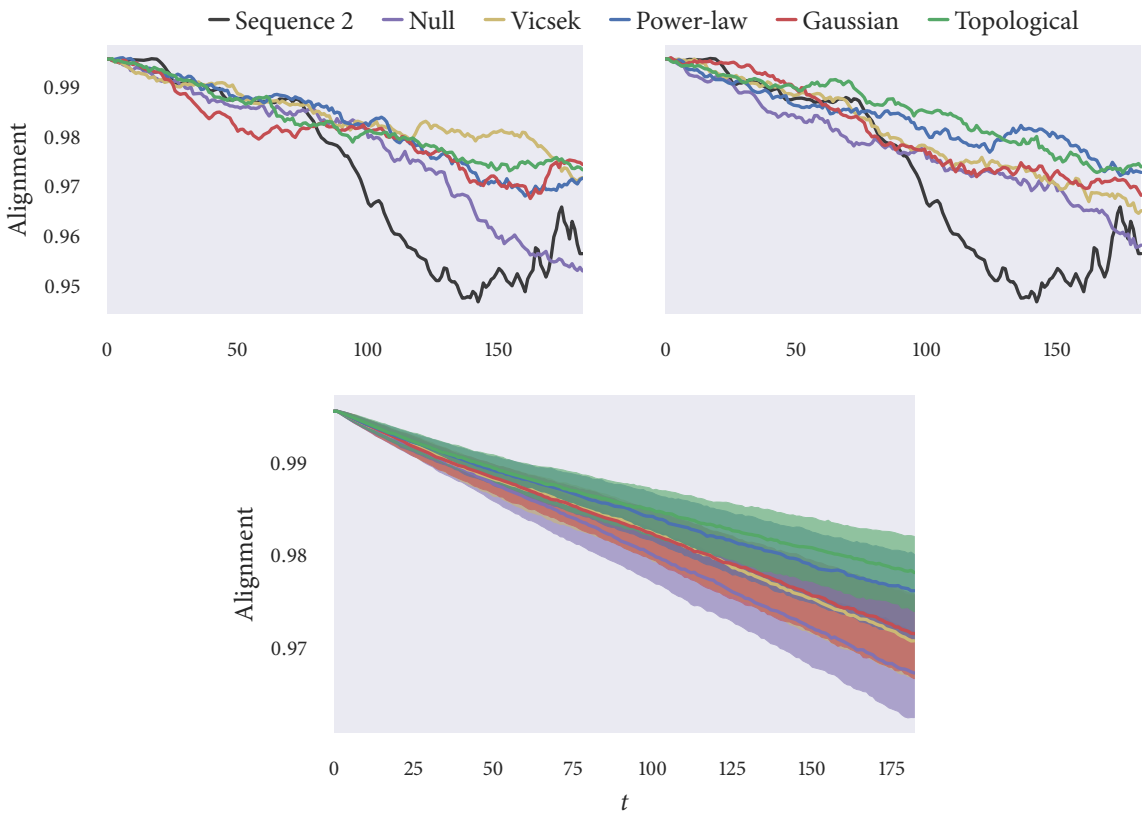



**Figure 6.10:** Flock alignment as a function of time for forward simulated candidate models. Models are initialised from the first frame of sequence 2, with model parameters drawn from our posteriors. Each candidate model is forward simulated one thousand times, with new model parameters drawn for each simulation. The top panel shows two randomly selected sets of forward simulations. The lower panel shows the median alignment of simulations as a function of time, with the bands around the median representing the 25th and 75th percentiles

Figure 6.10 shows alignment as a function of time for the resulting flocks. The top panel of Figure 6.10 shows two sets of randomly selected flocking instances. Here the models look to provide a good prediction of flock alignment for around the first 75 time steps. The bottom panel of Figure 6.10 shows the median flock alignment as a function of time. The bands around the coloured lines represent the 25th and 75th percentiles of the observed alignments. The most cohesive flocks were generated by the topological and power-law models—again corresponding to the models favoured by AIC.

### 6.2.3 *Sequence 3*

The flocking event captured in sequence 3 is much more dynamic than the events captured in sequences 1 and 2: going through a much larger directional change than the other sequences.

In addition to this, the flock of sequence 3 is not as densely packed as the flocks of sequences 1 and 2, as seen in Figure 6.4.

### *Posterior beliefs*

Parameter values are inferred from sequence 3 by sampling. The Null, power-law weighted, Gaussian weighted and topological models are fit using Stan, whereas the Vicsek model is fit with a Metropolis–Hastings algorithm. The samples drawn from the posterior are shown in Figure 6.11. Computing values of $\widehat{R}$ shows that all our samplers converged. Corresponding values of the effective sample size show a large number of samples were realised from the posterior.

The inferred parameters broadly align with those inferred from sequences 1 and 2: with the inferred interaction parameters representing weak interactions. However, the inferred degrees of freedom parameters are smaller than those inferred for sequences 1 and 2, revealing a noise distribution with more weight in its tails.

### *Standardised residuals*

Residuals are computed as the difference between observation and model prediction. Residuals are standardised by dividing by the posterior mean of the inferred noise-scale parameter $\sigma_Y$. Histogram plots of the standardised residuals are shown in Figure 6.12; a Student's $t$-distribution with $\nu$ degrees of freedom, taken from our posterior mean, is overlain in green. From Figure 6.12 we see a poor fit of model to data: although the residuals do look to follow a Student's $t$-distribution they do so with a location less than zero. This suggests that we were unable to capture the directional changes that the flock went through, and suggests the presence of behaviours not included in our model. Such behaviours could be agent-agent interactions such as attraction or repulsion, or could be interactions between agents and their external environment.

### *Model rankings*

Although plots of the standardised residuals show that none of the considered models provide a particularly good fit to data, it can still be informative to consider if any of the models outperformed the others. Table 6.4 shows the AIC values and corresponding Akaike weights computed for each model fit. From this we see that the power-law weighted

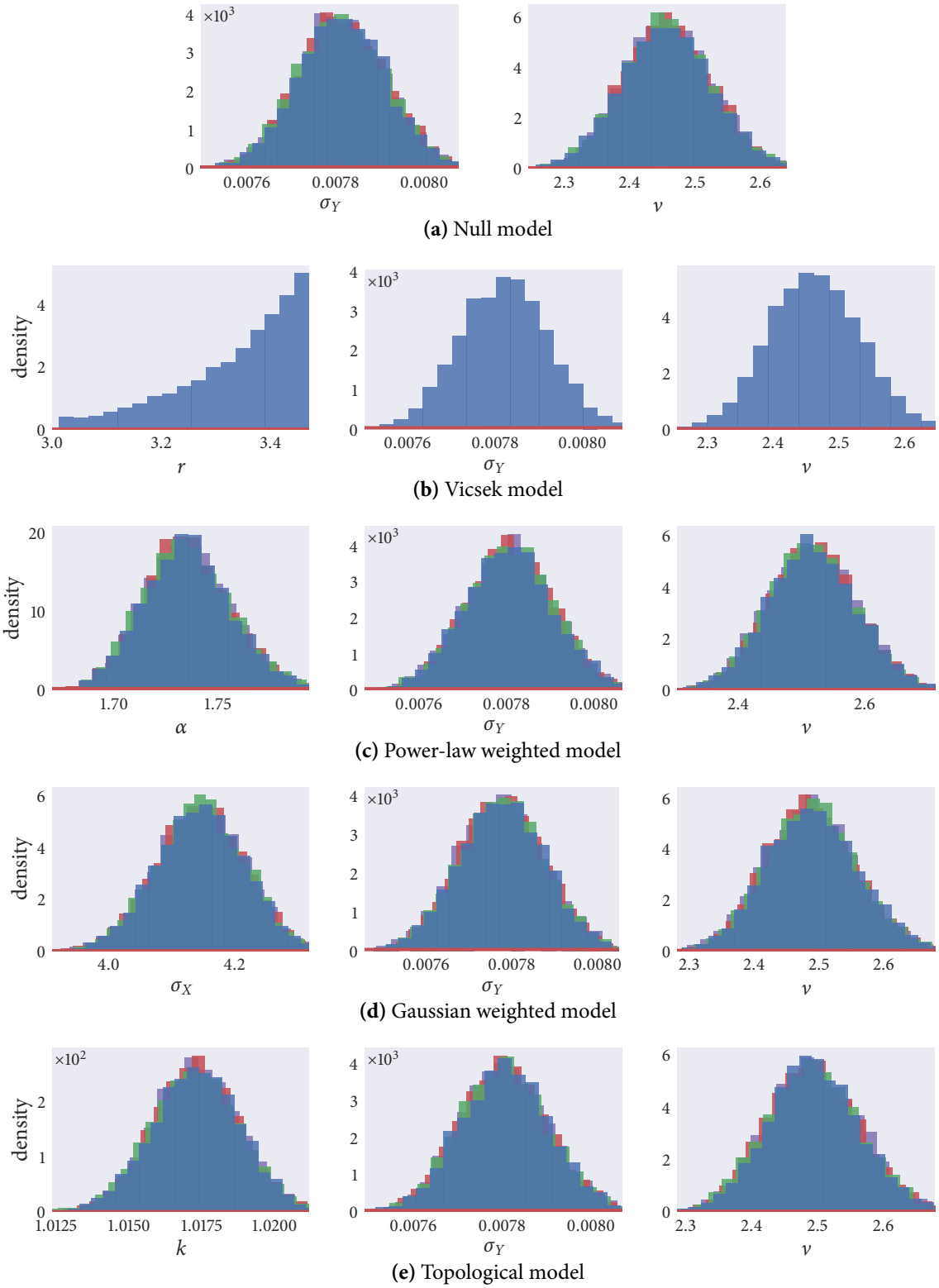



**Figure 6.11:** Posterior beliefs about the model parameters fitted to sequence 3. The Null, power-law weighted, Gaussian weighted and topological models were fit with Stan's NUTS algorithm. The Vicsek model was fit with a Metropolis–Hastings algorithm. Prior beliefs are overlain in red, and appear flat in comparison to our posteriors.

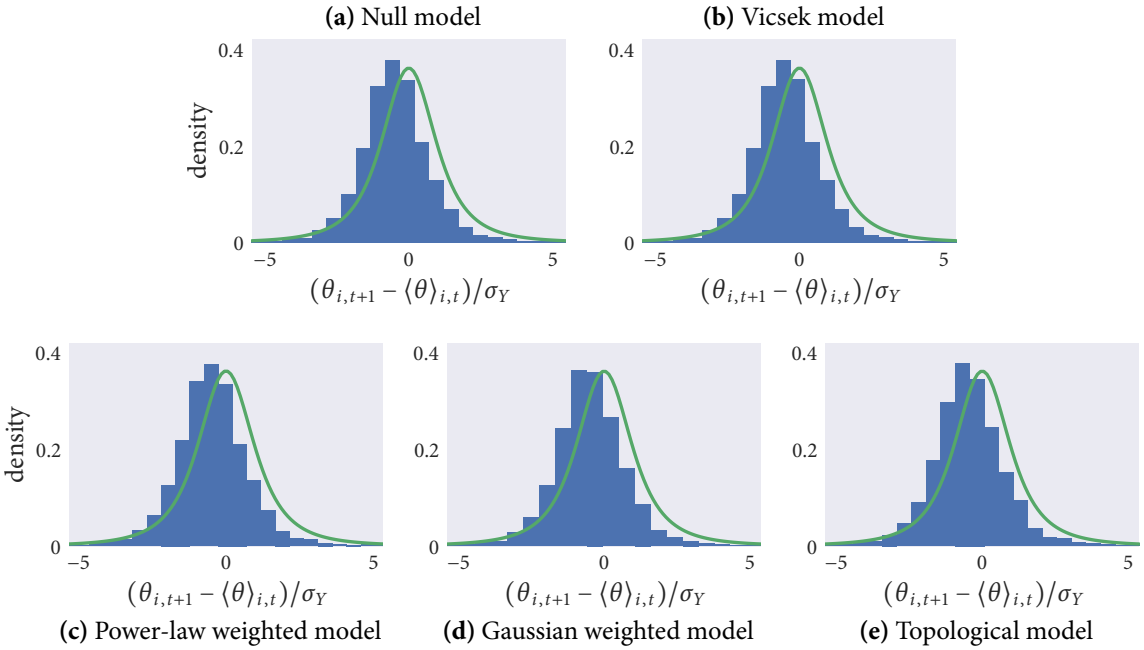



**Figure 6.12:** Standardised residuals from sequence 3 show a poor fit of model to data. Although the residuals do look to follow a Student's $t$-distribution, they do so with a location less than zero. This suggests the presence of behaviours not captured by our models. Missing behaviours could be interactions between sheep and their external environment, or missing sheep-sheep interactions such as repulsion or attraction behaviours.

| Model | AIC | rank | pAIC | dAIC | weight |
|---|---|---|---|---|---|
| Power-law | -52 242 | 1 | 3 | 0.0 | 1.00 |
| Gaussian | -52 222 | 2 | 3 | 20.3 | 0.00 |
| Topological | -52 169 | 3 | 3 | 73.6 | 0.00 |
| Null | -52 037 | 4 | 2 | 205.8 | 0.00 |
| Vicsek | -52 035 | 5 | 3 | 207.8 | 0.00 |

**Table 6.4:** Estimating predictive performance of models fitted to sequence 3 using Akaike Information Criteria. Models are ranked from best performing (1) to worst performing (5). The corresponding Akaike weight for each model is computed.

provided the best fit to the data. Consistent with sequences 1 and 2 we see the Null and Vicsek models as the worst performing candidate models.

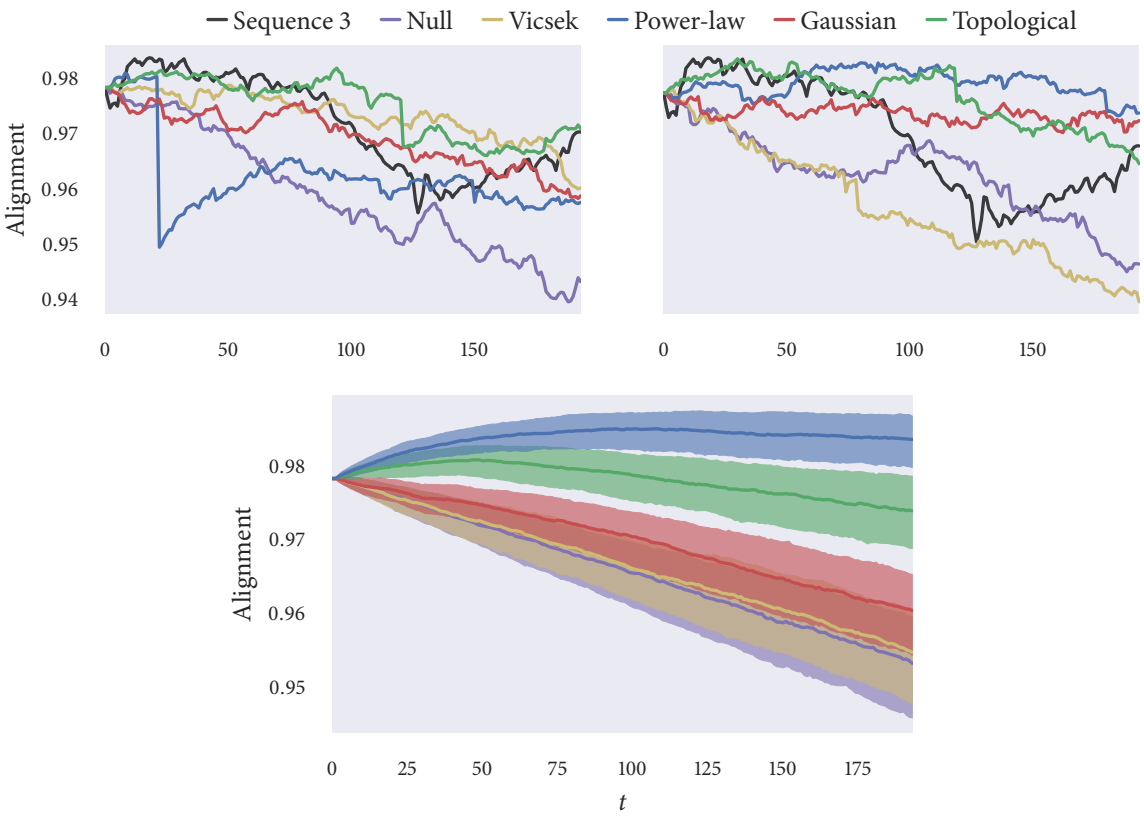



**Figure 6.13:** Alignment as a function of time for forward simulations of sequence 3. Model parameters were chosen as random draws from our posterior beliefs. Each candidate model was forward simulated one thousand times, with new parameter values drawn for each simulation. The top panel shows two sets of randomly selected forward simulations. The lower panel shows the median alignment of flocks as a function of time, the bands around the median represent the upper and lower quartiles of the alignment.

## *Posterior predictive checks*

Forward simulating fitted models provides an opportunity to assess a model's ability to reproduce realistic behaviours. Here we assess the ability of our models to replicate the change in alignment over time of our data. Although our residual plots show that we were unable to explain the directional changes of sequence 3, Figure 6.13 shows that out models do a reasonable job of mimicking the alignment of sequence 3 over time. The bottom panel of Figure 6.13 shows that the power-law weighted model produces the most consistently cohesive flocks, which was also the model favoured by AIC.

## *Conclusions*

We have evidenced that we are able to fit a number of competing agent-based models to real flocking events, working in a Bayesian framework. We demonstrated a number of different ways to assess model fit, and estimate predictive performance. Although here we only focused on a small subset of possible candidate models, our approach could be extended to fit and compare any number of models.

Our inferred posteriors consistently showed that only weak interaction parameters are necessary to maintain flocking events which are already highly-polarised. The small values of the degrees of freedom parameters inferred gave evidence of non-normally distributed noise.

Assessing standardised residual plots showed that the models generally provided a good fit to the data. However, for one of our sequences we observed that our models were unable to capture all the behaviours of the flock. From this we concluded that additional interaction rules, such as attraction or repulsion behaviours, or interactions with the environment, needed to be included.

In ranking our model fits we observe that the power-law weighted and topological models perform the best. That no single model comes out on top suggests that there may be aspects to the data which neither model is able to capture, or that the practical distinction between these models is so small that it is difficult to tell them apart. Though our analysis does not favour a single model as providing the best fit, models implementing continuous interaction rules consistently outrank those with discontinuous rules. We saw that the Vicsek model is no better at describing the data than the Null model, even with the additional complexity of an extra model parameter. Penalised for complexity, AIC consistently ranked the Vicsek model below the Null model; making the Vicsek model the worst performing candidate model.

# 7

# *Missing data*

Collecting data of flocking events is a demanding process (see, for example, Cavagna et al. (2008) and Lukeman, Li and Edelstein-Keshet (2010)). Along with the technical challenges posed by data collection, flocking events are inherently unpredictable, and so it is not possible to know when and where a flocking event may occur next. In this way there can become a frustrating "right-place-right-time" component to data collection.

Typically, recording equipment is set up in a fixed location where the scientist believes a flocking event may occur (Lukeman, Li and Edelstein-Keshet 2010). Stationary recording equipment results in a fixed field of vision in which data may be captured. Unfortunately, this stationary set-up can result in recording incomplete flocking events. This may happen when flock members stray outside the field of vision during a recording event. As the recording equipment is fixed in location the field of vision cannot be adjusted to reinclude those who move out-of-frame.

The flock members which move out of frame *cannot* be ignored during analysis: although we may not observe their movements, they may still be influencing the behaviour of the flock, and so must be accounted for. A simple but undesirable solution is presented by the temptation to discard *every* frame in which *any* individual is out of view. This "solution" has the potential to drastically reduce the amount of data available for analysis. As capturing flocking events can be such an involved and time-consuming endeavour, it seems remiss to discard observations and the information they contain.

In this chapter we will consider how we can handle flocking events which contain missing observations. The movements of missing agents will not be ignored, nor will data be

discarded. Instead, we will work in a Bayesian framework to adjust our posterior densities about model parameters to account for unobserved behaviours. We will see that this can be achieved by integrating over all the possible trajectories of missing agents. We show that this approach gives more accurate results than the naive approach of discarding data, before demonstrating the developed methodology on real observation.

## 7.1 Types of missingness

When we consider flocking events with missing observations, it is important to consider at which point during the sequence a given agent went missing. This is because how we account for the missingness will depend on at which point during the sequence the agent went missing.

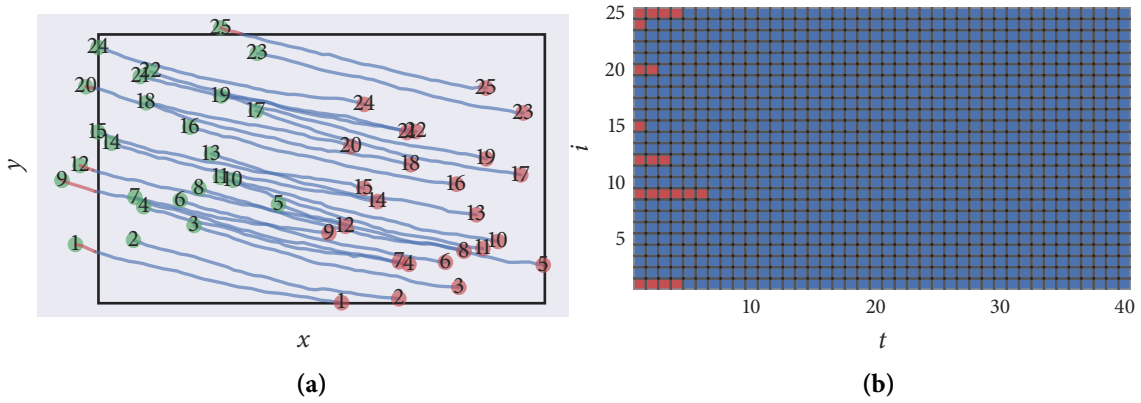
Although there are many different circumstances which can result in an agent leaving our visual field, here we shall inspect the two cases which we consider as the most likely to occur. These two cases arise when agents are out of frame at the *beginning* of a recording event, or when agents are out of frame at the *end* of a recording event.

We shall integrate over the possible missing observations using a Metropolis–Hastings scheme. A proposal mechanism for generating paths missing at the beginning of a sequence is outlined. Proposed paths can then be accepted or rejected using results from Chapter 5. We will then show that we can generate paths missing at the end of a sequence by implementing a Gibbs sampler (Algorithm 1); sampling from our full conditional distributions to realise a distribution of possible paths.

### 7.1.1 *Missing in the beginning of a sequence*

We say that data is missing at the beginning of a flocking event if the observer began recording the sequence before all agents had entered the frame. When we consider this case we assume that all the flock members do eventually enter the visual field, and so the total number of individuals in the flock is known.

We imitate this scenario by forward simulating the Vicsek model and imposing a fixed field of vision on top of the resulting data. We chose to forward simulate 25 agents for 40 time steps. At time $t = 1$ agents were directed and positioned randomly within a square-cell of side length $L = 1$. Agents experienced noise generated from a generalised Student's $t$-distribution with $\nu = 7$ degrees of freedom and scale $\sigma_Y = 0.075$. Each individual interacted with neighbours positioned within distance $r = 0.5$, and moved with speed $v = 0.03$.




(a) (b)

**Figure 7.1:** Illustrating our forward simulation of the Vicsek model and the data considered missing. (a) Simulated trajectories. The rectangle overlain in black represents our imitation of a fixed field of vision: analogous to that of recording equipment used in the field. Agents travel from green marker to red marker, with their ID ($i$) shown at their start and end points. Data points outside of the rectangular region are classified as missing, and are shown by red trajectories. (b) Summarising the missing and observed data points of our simulation. A blue gridpoint at location $(t, i)$ indicates that agent $i$ was within our visual field at time $t$. A red gridpoint at location $(t, i)$ indicates that agent $i$ was outside our visual field at time $t$.

The resulting simulated data is illustrated in Figure 7.1(a). The rectangle overlain in black represents a fixed field of vision: analogous to the fixed visual field generated by recording equipment in the field. The movements of agents outside this fixed area are classified as missing and are represented by the red trajectories. Agents move from the positions denoted by the green markers, to the positions represented by the red makers. The ID number of each agent, $i$, is shown at the beginning and end of its trajectory. It can be seen that some agents were *outside* our visual field at the beginning of the simulation.

Figure 7.1(b) shows a schematic representing which data points of our simulation were considered observed and which data points were classified as missing. A blue gridpoint in location $(t, i)$ tells us that agent $i$ *was observed* at time $t$. Conversely, a red maker at location $(t, i)$ indicates that agent $i$ *was missing* at time $t$. From this schematic we can see that a number of agents were classified as missing at time $t = 1$: the beginning of the simulation.

We may use Equation (5.1) to quantify the likelihood of observing a flock's movements. However, to compute the likelihood we must observe the positions and directions of each individual in every frame. With this, to evaluate the likelihood of a flock with missing observations we must propose candidate values for the missing data points. To do so we shall work backwards from our first observation of each missing agent.

We propose missing directions of motion backwards in time from our first observation

of each missing agent. A symmetric proposal distribution is used to propose a missing direction $\theta_{i,t-1}$ as:

$$\theta^{\star}_{i,t-1} \,|\, \theta_{i,t}, \nu, \sigma_Y \sim t_\nu(\theta_{i,t}, \sigma_Y), \tag{7.1}$$

where $t_\nu$ denotes a generalised Student's $t$-distribution with $\nu$ degrees of freedom. An alternative proposal scheme could be formulated using the interaction term $\langle\theta\rangle_{i,t}$ as:

$$\theta^{\star}_{i,t-1} \,|\, \langle\theta\rangle_{i,t}, \nu, \sigma_Y \sim t_\nu(\langle\theta\rangle_{i,t}, \sigma_Y). \tag{7.2}$$

However, as this proposal necessitates the computation of $\langle\theta\rangle_{i,t}$, it generates candidate values of $\theta^{\star}_{i,t-1}$ at greater computational expense than Equation (7.1). For highly aligned flocks we expect Equations (7.1) and (7.2) to propose similar candidate values, with Equation (7.1) doing so at a lesser computational cost. It is for this reason that we shall use the proposal distribution of Equation (7.1).

The proposed directions are used to reconstruct the trajectories which they represent. This is achieved by rearranging Equation (4.1) such that:

$$\boldsymbol{x}_{i,t} = \boldsymbol{x}_{i,t+1} - \boldsymbol{v}_{i,t}\Delta t. \tag{7.3}$$

Recall from Section 4.1 that $\Delta t = 1$ and that $\boldsymbol{v}_{i,t}$ is constructed to have direction $\theta_{i,t+1}$ and speed $\nu$. Given $\theta_{i,t+1}$, we may compute $\boldsymbol{v}_{i,t} = (\cos\theta_{i,t+1}, \sin\theta_{i,t+1})^T\nu$. With this, and reindexing $t \mapsto t-1$, we may rewrite Equation (7.3) as:

$$\boldsymbol{x}_{i,t-1} = \boldsymbol{x}_{i,t} - (\cos\theta_{i,t}, \sin\theta_{i,t})^T\nu\Delta t. \tag{7.4}$$

Having proposed candidate values for the missing directions of motions with Equation (7.1), we can use Equation (7.4) to compute the proposed paths which these directions correspond to. As these proposed paths represent *missing* data, if any of the proposed positions lie *within* our frame of vision, then the proposals must be rejected.

Having values for the positions and directions of motion of every individual at every time step, we may use Equation (5.1) to quantify the likelihood of these paths, given some model parameters.

To infer the Vicsek model's parameters from our simulated data we shall implement a random walk Metropolis–Hastings sampler (Algorithm 2). At each iteration of the sampler we shall propose model parameters and candidate values for the missing observations. These proposed data points are then accepted or rejected along with the proposed model parameters, with probability given by the acceptance probability. With this, each missing

observation introduces an additional dimension to the posterior distribution, and so increases complexity. As a result, as the amount of missing data increases, the computational demand of the problem increases.

We simulate our random-walk sampler for 100,000 iterations, and thin the resulting output by a factor of 10. The output is thinned to reduce the autocorrelation between successive samples, and to decrease the memory overhead of the computation. We assess convergence by inspecting trace plots of the simulated chains. Figure 7.2 visualises the chains targeting the directions of motion $\theta_{6,1}$, $\theta_{6,2}$, $\theta_{6,3}$ and $\theta_{6,4}$ which were classified as missing. From this plot we see well-behaved chains oscillating regularly around a fixed location, evidencing convergence. The true values of the missing observations are indicated by the horizontal green lines. See that the true values are all captured within the oscillations of our chains, but that the further back in time we work the greater the oscillations become. The magnitude of the oscillations increase as we extrapolate further backwards in time and become more uncertain about the possible movements of missing agents. Figure 7.4 shows the missing paths sampled at four randomly chosen iterations of our scheme, and demonstrates that our scheme is capable of generating realistic trajectories in place of missing observations.

Figure 7.3 shows chains targeting the missing $x$ co-ordinates of agent 6 over the first four frames. The chains targeting the $x$ co-ordinates are related to the chains targeting the missing directions through Equation (7.4). Figure 7.3 gives further evidence that our chains have converged, and that the further back in time we extrapolate the less certain we become about an agent's whereabouts.

As with the hierarchical models considered in Section 5.2, we now have a posterior distribution with a large number of dimensions, and so assessing and presenting every marginal posterior density becomes impractical. Instead, we summarise the output of our chains targeting the missing observations in Figure 7.5. Each panel of Figure 7.5 shows box and whisker plots summarising our posterior samples. The yellow markers show the posterior median, the inner-extent of the whiskers show the upper and lower quartiles, and the outer-extent of the whiskers show the posterior median $\pm 1.5 \times$ IQR.

The left-most panel of Figure 7.5 summarises the posterior densities representing our beliefs about the missing directions of motion. The green markers represent true values. See that as we extrapolate further back in time our posterior variance increases, as we would expect. The $x$ and $y$ co-ordinates corresponding to these directions are shown in the central and right-most panels. As the between-agent variance in positions is much greater than our posterior variance, to compare all the inferred positions in one plot we first subtract the realised samples from the true values. This results in box and whisker plots centered

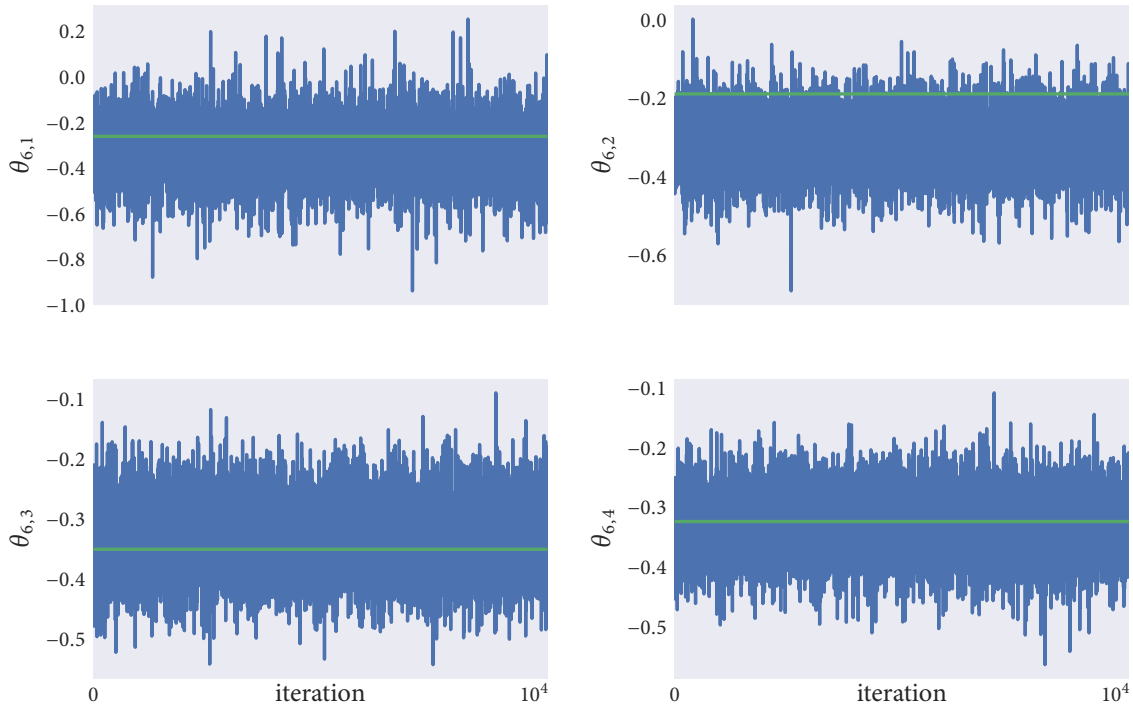



**Figure 7.2:** Chains targeting directions of motion missing at the beginning of a simulation. Each trajectory was simulated for 100,000 iterations. To decrease the autocorrelation between samples and reduce memory overhead the output was thinned by a factor of 10. The true values of the missing directions are shown by the horizontal green lines. The true values are captured within our posterior densities.

approximately around zero (shown by the green markers): indicating that our posterior densities accurately captured the true values. Again, we see that our posterior variance about the inferred positions increases the further back in time we extrapolate.

We have shown that we can capture missing observations by implementing a Metropolis–Hastings sampler. Now that we have this, we wish to inspect how having missing observations has influenced our parameter uncertainty. We also wish to investigate how our approach to missing data compares to the more naive approach of discarding frames which don't capture every individual. To make this comparison we re-perform parameter inference on the simulated data, in the first instance increasing our field of vision to capture all agents, and in the second instance discarding frames where any individual is missing. We expect that the posterior variance about the model parameters with the missingness accounted for should be greater than or equal to the posterior variance had there been no missingness. Similarly, had we discarded the frames with missingness, we would expect a posterior variance greater than or equal to the posterior variance with the missingness integrated over.

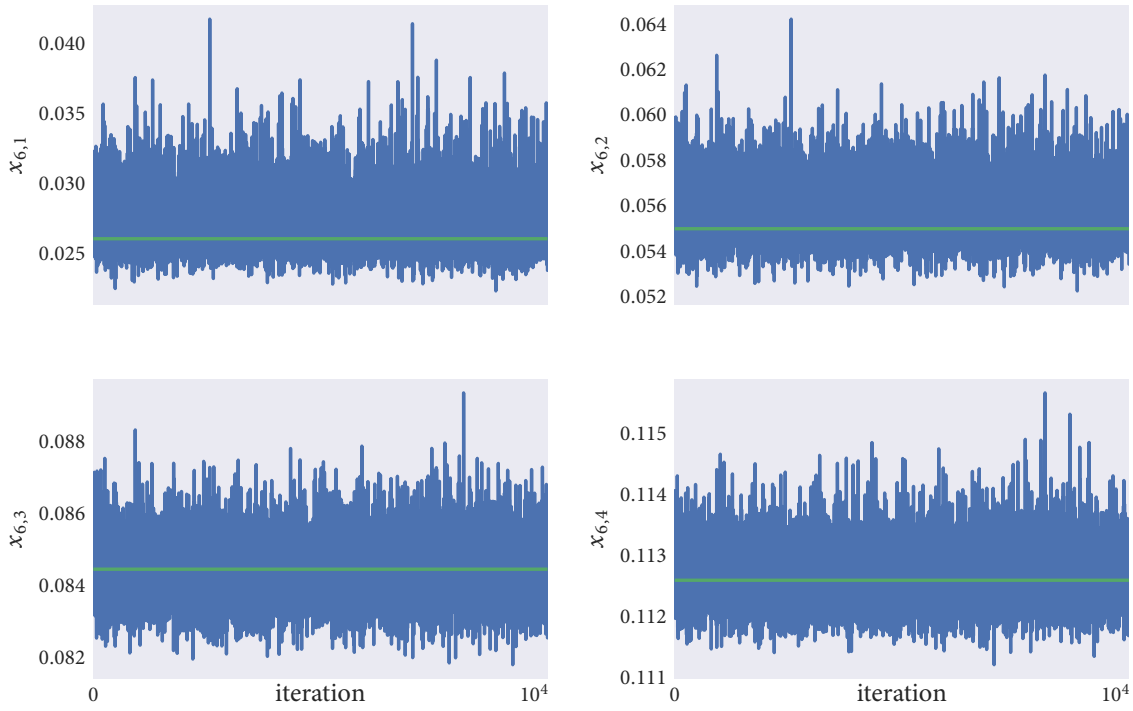



**Figure 7.3:** Trajectories of the $x$ co-ordinates of agents corresponding to the directions of motion seen in Figure 7.2. Sampled directions of motion are related to sampled co-ordinates through Equation (7.4). The true values of the missing co-ordinates are overlain in green.

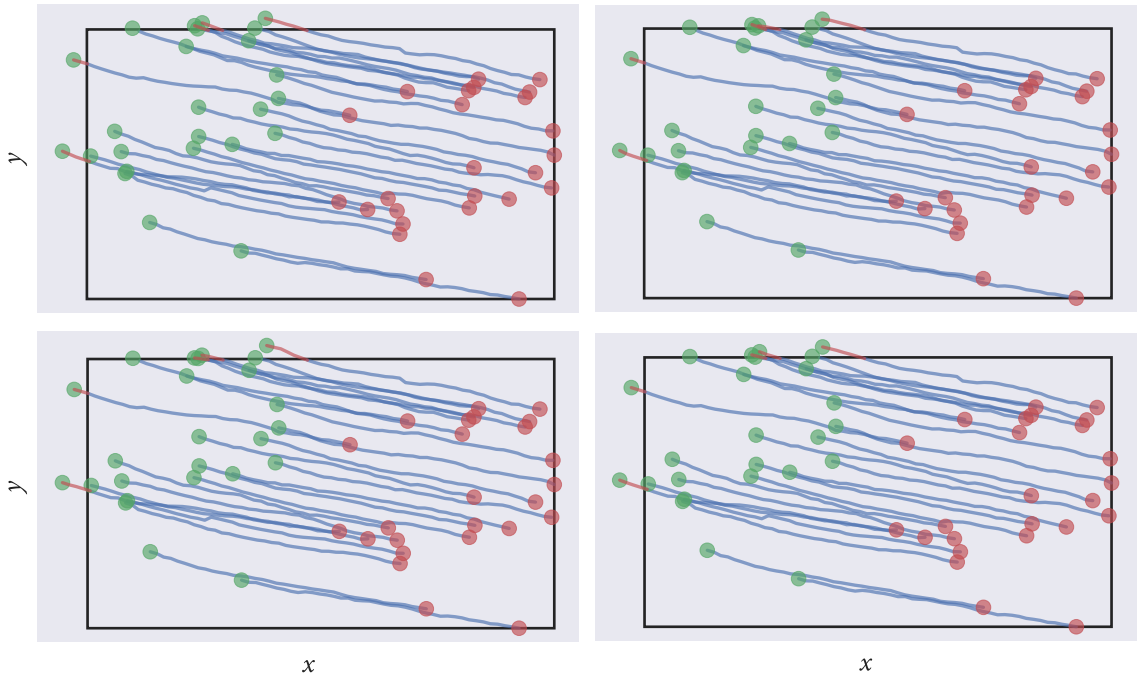



**Figure 7.4:** Visualising completed paths at four randomly chosen iterations of our inference scheme. The plots show plausible trajectories were constructed by our scheme.

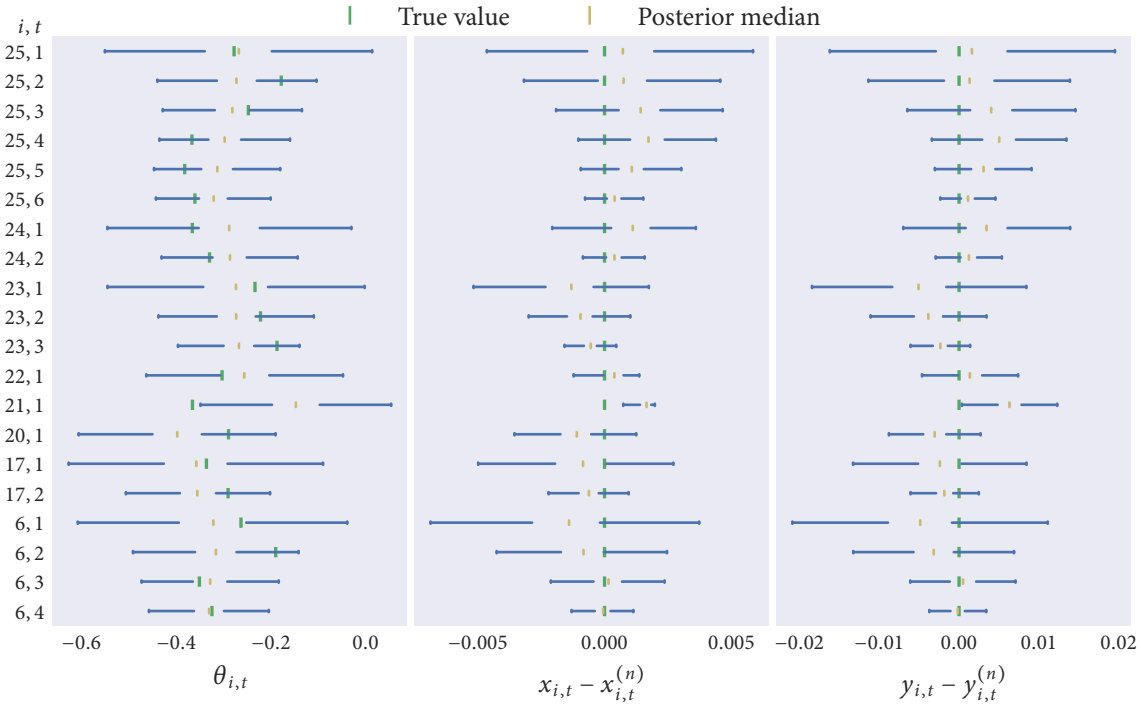



**Figure 7.5:** Summarising the posterior samples targeting the missing directions of motion (left), $x$ co-ordinates (centre), and $y$ co-ordinates (right). True values are shown by the green markers, and posterior medians are shown by the yellow markers. The box and whisker plots show the upper and lower quartiles as well as the median $\pm 1.5 \times$ IQR. Realisations of the missing positions are subtracted from the corresponding true values to make comparison easier.

Figure 7.6 compares posterior densities of the model parameters under three different regimes: the flocking event without any missingness (blue), the event with the missingness integrated out (red), and the event with the missingness discarded (purple). From this plot we see that integrating over the missing observations provides very similar posterior densities to if we had observed all the data. We see that our approach to handling missingness gives smaller posterior variance than the naive approach of discarding observations, indicating the efficacy of our approach. As the computational demand of the problem increases as more data goes missing, effort should still be made by the scientist to minimise the total amount of missingness incurred during data collection.

### 7.1.2 *Missing in the end of a sequence*

We say that data is missing at the end of a flocking event if some individuals leave the frame of vision *before* the recording event was completed. When we consider this case we assume that agents which leave the frame do not later re-enter it. We imitate this situation by

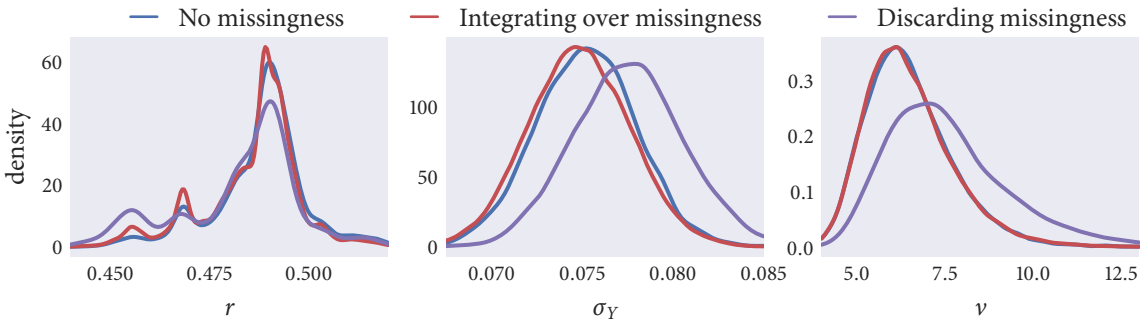
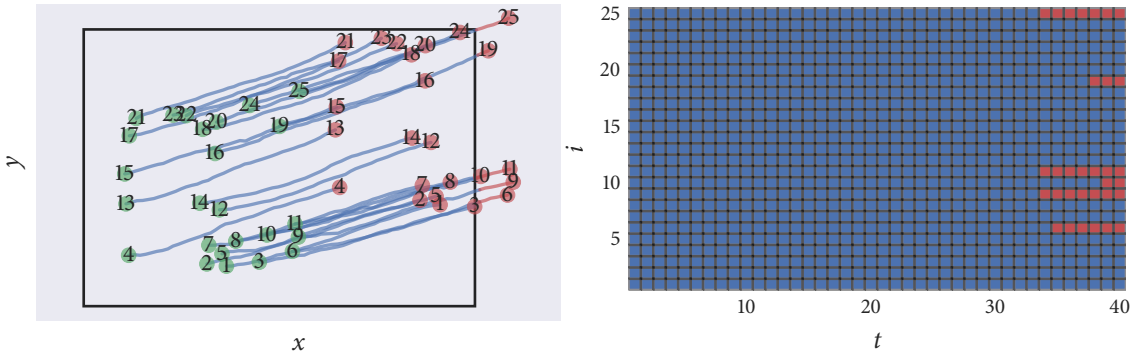



**Figure 7.6:** Comparing posterior densities about the Vicsek model parameters had all the data been observed (blue); with the missingness integrated out (red), and when discarding frames which any individual is missing from (purple). The posterior densities are shown by kernel density estimates of the posterior samples.




**(a)** **(b)**

**Figure 7.7:** Individuals missing at the end of a sequence. (a) Twenty-five agents are simulated for forty frames according to the rules of the Vicsek model. Agents travel from the positions denoted by the green markers to those denoted by the red markers. The ID of each agent is shown at its start and end point. A fixed field of vision is emulated by the rectangle overlain in black. Any movements which occur outside this area are classified as missing; shown by trajectories transitioning from blue to red. (b) A representation of which data points of our simulation were observed and which were classified as missing. A blue tile at location $(t, i)$ indicates that agent $i$ was within our field of vision at time $t$. Conversely, a red tile at location $(t, i)$ indicates that agent $i$ was outside our field of vision—that is, missing—at time $t$.

simulating the movements of 25 individuals over 40 frames, according to the specifications of the Vicsek model (here with parameters $r = 0.45$, $\sigma_Y = 0.07$ and $v = 7$), and overlaying a fixed field of vision. The resulting data is visualised in Figure 7.7.

To fit one of our models to this data we must be able to evaluate the likelihood as specified in Equation (5.1). This likelihood involves a product over $i = 1, \ldots, N$ and $t = 1, \ldots, T$, and so we must have data to represent these observations. As some agents are out of frame at the end of the sequence, we do not have all the observations necessary to evaluate the

likelihood.

We can propose trajectories for the missing agents by forward simulating the model from each agent's last observed position (Equations (4.1) and (4.6)). Doing so we write down the full conditional distribution for missing directions at the end of a sequence as:

$$\theta^{\star}_{i,t+1} \,|\, \langle\theta\rangle_{i,t}, \nu, \sigma_Y \sim t_\nu(\langle\theta\rangle_{i,t}, \sigma_Y) \tag{7.5}$$

Realising paths missing at the end of a sequence is now achieved with a Gibbs step (Algorithm 1). With this there is no accept / reject step for the sampled trajectories, and so new missing paths are evaluated at every iteration of our sampler. As the sampled trajectories represent missing observations, any samples that are generated within our field of vision are discarded and new trajectories are sampled instead. This Gibbs step is embedded within a Metropolis–Hastings sampler which works to infer the model parameters of the Vicsek model.

Figure 7.8 shows the chains targeting the missing directions of motion of agent $i = 6$. These trajectories show evidence of convergence and are seen to capture the true values of the missing data (green). These directions are related to the $y$-component of the missing positions (shown in Figure 7.9) via Equation (4.1), and again capture the true values of the missing data. Figure 7.10 goes to demonstrate that the paths generated by our scheme are plausible: visualising four randomly chosen flock configurations considered by our sampler.

In Figures 7.8 and 7.9 we only inspect a small amount of the total missingness targeted by our inference. Figure 7.11 summarises the posterior densities about all the missing observations of our simulated data. We see that in all cases the true values (green) are captured within our posterior densities (blue). From the central and rightmost panels of Figure 7.11 we see that posterior variance (uncertainty) increases the further into the future we extrapolate.

Figure 7.11 shows that our approach to missingness is able to capture data missing at the end of a recording event. We now wish to inspect how this approach has effected our posterior densities. To do so we perform parameter inference on the simulated data as if all the data had been contained within our field of vision. In addition to this we perform the same inference but discarding all the frames in which any agent was out of view. The model parameters inferred under these three scenarios are compared in Figure 7.12. We see that integrating out the missingness reveals results similar to as if we had observed all the data, and that the naive approach of discarding missingness resulted in a larger posterior variance.

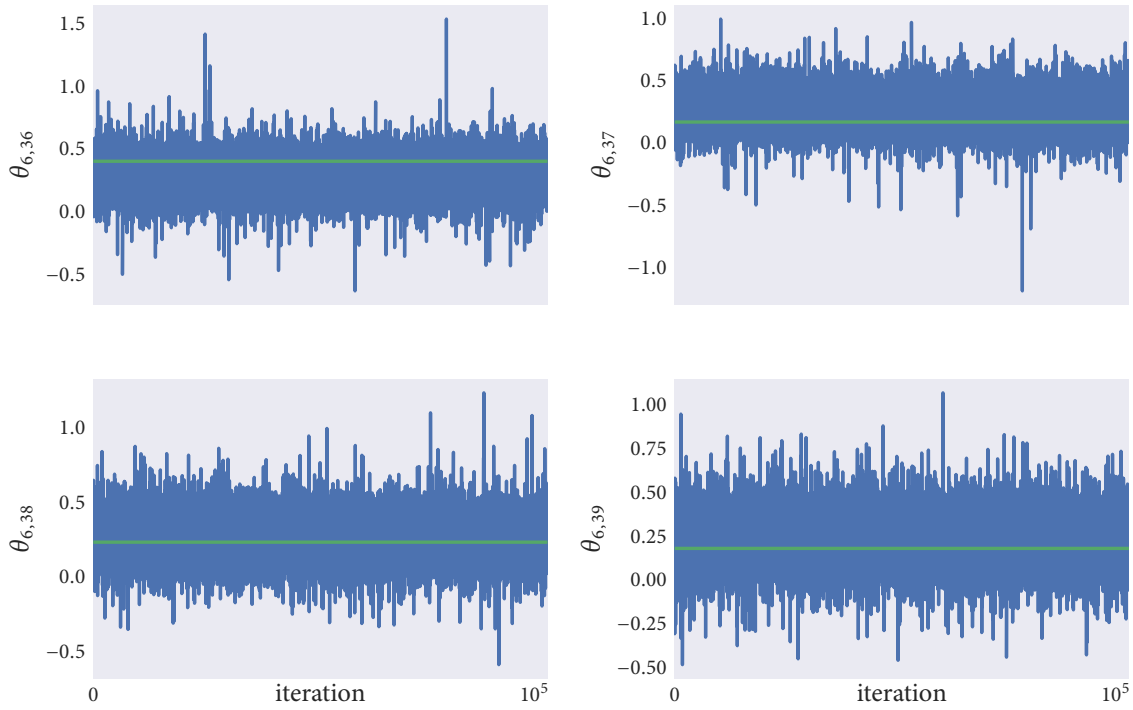



**Figure 7.8:** Trace plots targeting directions of motion missing at the end of the simulated flocking event. The chains give strong evidence of convergence and show that the true values of the missing directions (green) are captured within the posterior densities.

The Gibbs sampler explores the sample space more efficiently than the Metropolis–Hastings algorithm, an artefact of Gibbs' lack of an accept / reject step. As a result we can handle data missing at the end of a sequence with less computational expense than data missing at the beginning of a sequence. We conclude that although all effort should be made by the scientist to avoid recording sequences with any missing observations, where possible care should be taken to bias the occurrence of data missing at the end of a sequence over data missing at the beginning of a sequence.

### 7.1.3 *Missing in the beginning and end of a sequence*

Having demonstrated that we can handle sequences with observations missing at the *beginning* of an event and sequences with observations missing at the *end* of an event, we now seek to verify that we can handle sequences with data missing at the beginning *and* end. Having considered these two cases separately we now consider them jointly.

To handle sequences with data missing at the beginning and end, it is sufficient to use the Gibbs step as outlined in Section 7.1.2 to sample paths missing at the end of the sequence, and the Metropolis–Hastings algorithm as in Section 7.1.1 to propose paths missing at the

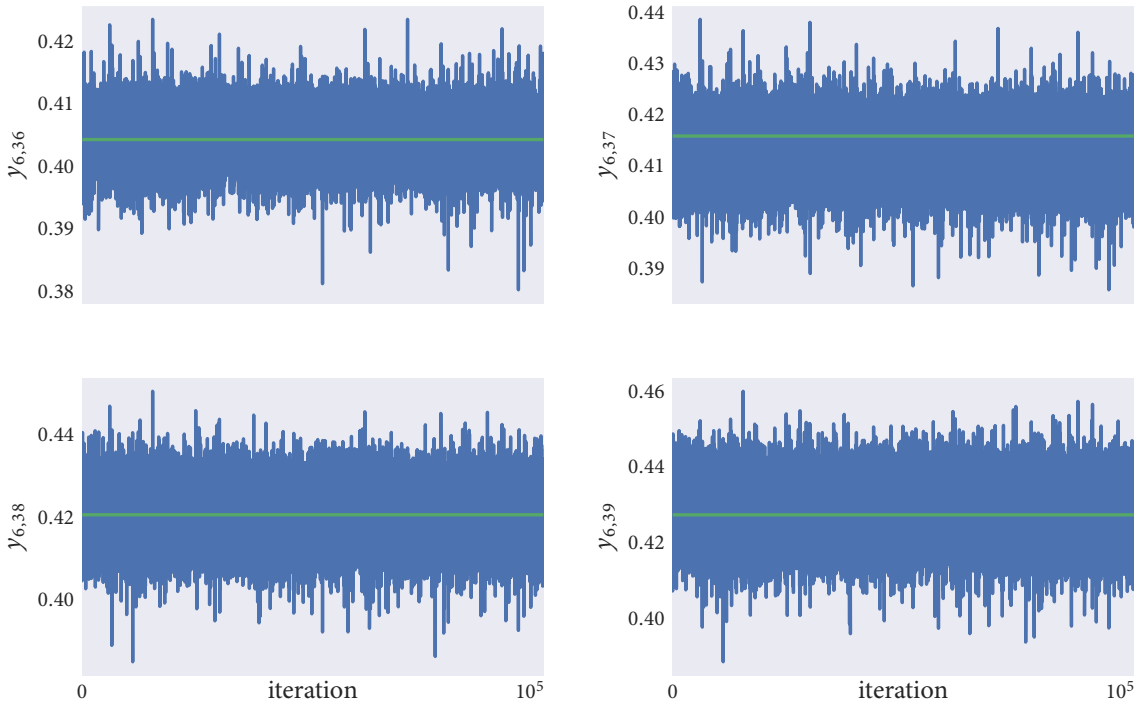



**Figure 7.9:** Trajectories targeting the $y$-component of the missing positions of agent $i = 6$. The chains show evidence of convergence: oscillating around a fixed point with constant variance. The true values (green) are captured within the oscillations of the chains.

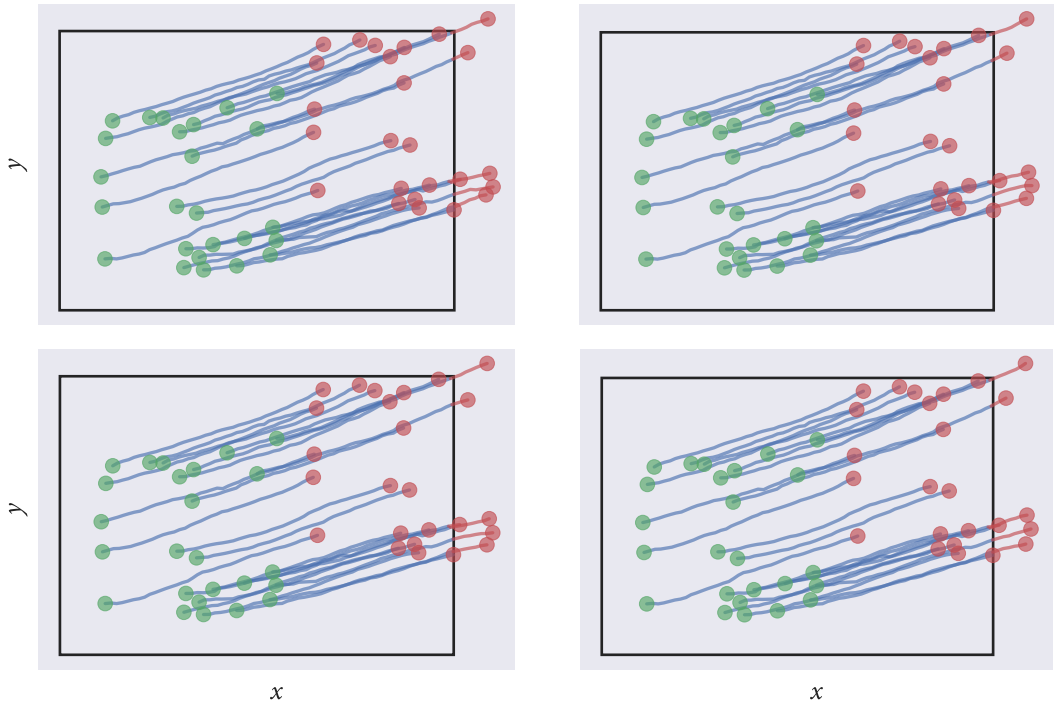



**Figure 7.10:** Visualising, at four randomly chosen iterations, missing trajectories that were completed by our inference scheme. The sampled paths look plausible.

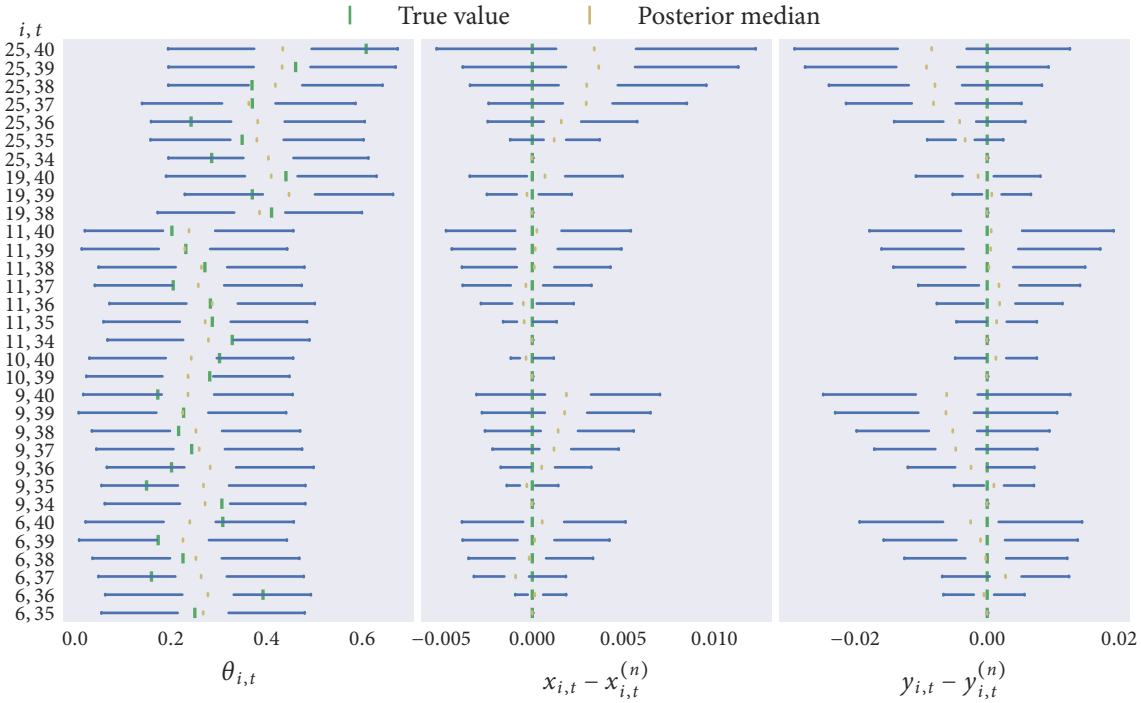



**Figure 7.11:** Summarising the posterior densities of the missing directions (left panel), $x$ co-ordinates (central panel), and $y$ co-ordinates (right panel). For comparative ease the samples of the missing positions are subtracted from the true missing positions. The true values being targeted are shown by the green markers, and all lie within the box and whisker plots representing our posterior densities. As we extrapolate further into the future we see that our posterior variance increases, as our uncertainty about the values of the missing observations increase.

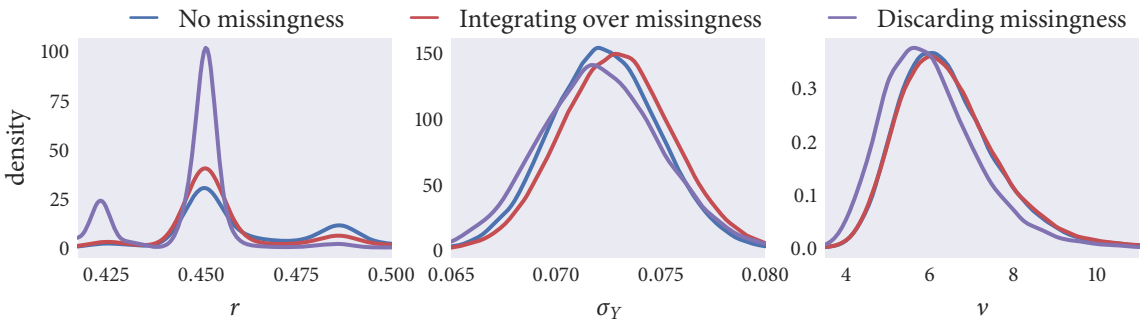



**Figure 7.12:** Kernel density estimates of the model parameters inferred from simulated data. The blue kernels represent posterior densities when the field of vision was widened to include all movements; the red kernels show posterior densities when the missingness was integrated out, and the purple kernels represent the posterior samples when all the frames containing missing data were discarded.

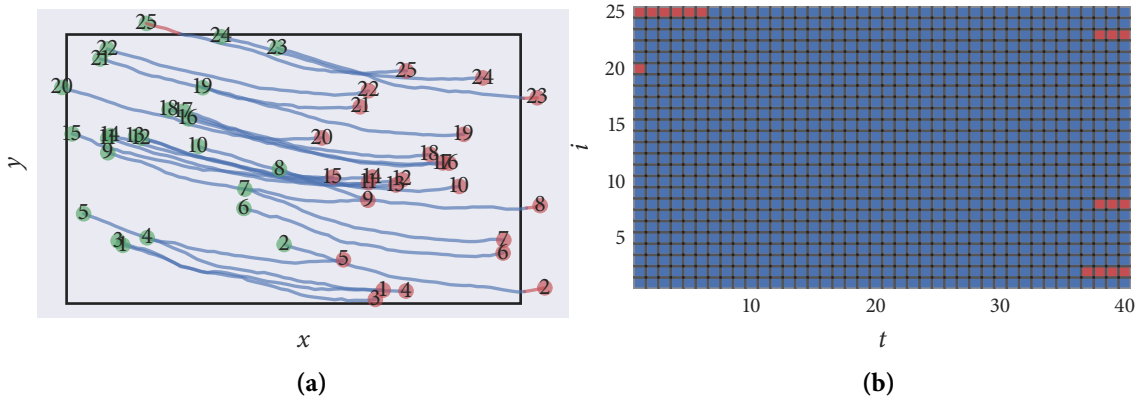



**Figure 7.13:** Simulated data with observations missing at the beginning and end of the sequence. (a) Showing the trajectories of the simulated agents. The rectangle overlain in black shows the simulated field of vision. As agents move from the green markers to the red markers we see that there are agents missing at the beginning and end of the sequence. (b) Schematic illustrating which data points were observed and which were classified as missing; a blue marker represents an observed data point, and a red marker represents a missing data point.

beginning of the sequence.

The data simulated for this study are shown in Figure 7.13. See that the simulated data shows data points missing at the beginning *and* end of the event. This is again seen in 7.13(b), which illustrates which data points were observed and which classified as missing.

We fit the Vicsek model to this simulated data using our implementation of the Metropolis–Hastings algorithm. Parameter inference is performed on (*i*) data with no missingness, (*ii*) data with missingness integrated out and (*iii*) data with missingness discarded. The posterior densities realised under these three regimes are compared in Figure 7.14. From this figure we see that integrating over the missingness resulted in posterior densities most similar to the densities with all the data observed.

## 7.2 CASE STUDY II: FORAGING SCOTERS

In Section 7.1 we discussed how the use of fixed-location recording equipment during data acquisition can result in individuals moving out-of-frame *during* recording. Such movements are then missing from the recorded flocking event. As out-of-frame movement may still be influencing the behaviour of the observed flock, this movement must be accounted for during model fitting.

In this section we consider a *real* flocking event with missing observations. As foreshadowed in the previous sections, the missing observations of this dataset will be realised as a

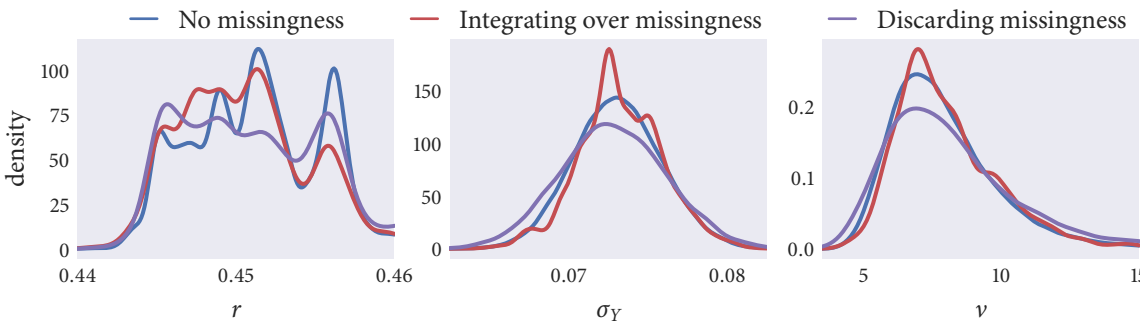



**Figure 7.14:** Comparing posterior densities about the parameters of the Vicsek model with and without missingness. The blue kernels show the posterior densities as if all the data was observed. The red kernels show the posteriors when the missing observations were integrated out. Finally, the purple kernels shows the posterior beliefs had the missing observations been discarded.

consequence of employing fixed-location recording equipment.

The dataset considered describes the movements of a large number of surf scoters, a migratory sea bird, which gather in groups to forage. The captured flocking events take place on the surface of a lake, and so movement is effectively restricted to a two-dimensional plane. This dataset was provided courtesy of work performed by Lukeman (2009) and Lukeman, Li and Edelstein-Keshet (2010). At the time of their publication the captured events represented a tenfold increase in the number of individuals which could be reliably linked *between* frames.

Unfortunately, owing to the fixed-location cameras used to record these flocking events, there were often scoters out of view in the recorded frames. Figure 2.6 illustrates the steps performed to extract the positions of individuals from raw footage; this figure also shows how the flock occupied an area larger than that which the camera captured—resulting in missing observations. As earlier, the missing data points can be partitioned into observations missing at the *beginning* of the recording event, and observations missing at the *end* of the recording event.

### 7.2.1 *Scoter data*

Lukeman, Li and Edelstein-Keshet (2010) recorded a number of flocking events from an aerial vantage point: an elevated promenade at the side of a large lake. From this position the authors were able to direct a camera toward an inlet where overwintering scoters had been observed foraging. Knowing the height of this camera above the water, and the angle of its approach with the horizontal, the authors were able to transform the captured camera

data back to "real-world" co-ordinates.

However, to reliably transform back from camera co-ordinates to real-world co-ordinates, the height of the camera above the lake and its angle with the horizontal had to be fixed. With the camera fixed in position, the authors then waited for flocking events to occur. Recording began when individuals entered the camera's field of vision, and ended when all the individuals left the frame, or the flock became stationary. Each recording event captured the movements of around 170 scoters, with each event consisting of approximately 30 frames. However, of the recorded events between 16% and 64% of each flock's movements took place out-of-frame.

Lukeman, Li and Edelstein-Keshet (2010) discounted the influence of individuals out-of-frame, and instead focused on reproducing radial and angular neighbour densities of *internal* group members (edge individuals were also discarded). This approach, although representing a significant step forward in the literature, came with its drawbacks. Most significantly, focusing the fitting on reproducing radial and angular neighbour distributions removed the *dynamic* component to the data which the authors worked so hard to achieve. Additionally, the authors determined the size of the agents' interaction radius by visual inspection of nearest-neighbour densities; the interaction radius used in their model being chosen *before* the fitting process began.

In this chapter we take a more holistic approach to model fitting; making sure to account for *all* individuals—both observed and missing, as well as internal and boundary individuals—and focusing the fitting on reproducing the *movements* of the flock, rather than some epiphenomena of their movements. Significant computation is necessary to account for the large amount of data missing from the recorded flocking events.

### 7.2.2 *Model fitting*

As the computational demands of missing data problems increase as the amount of missingness increases, we focus on the flocking event within the Lukeman dataset which displays the *least* amount of missingness. For the events captured by Lukeman, Li and Edelstein-Keshet (2010) this is represented by a sequence of 199 scoters moving over 23 frames. Of the $199 \times 23 = 4,577$ data points represented by this event, 680 ($\approx 16\%$) occur outside the visual field of the camera. This flock is visualised in Figure 7.15. From this figure we see that the flock was strongly polarised throughout the recording event; with this observation we consider an alignment-based model a plausible candidate model for fitting.

Figure 7.16 gives a visual representation of which data points were observed, and which

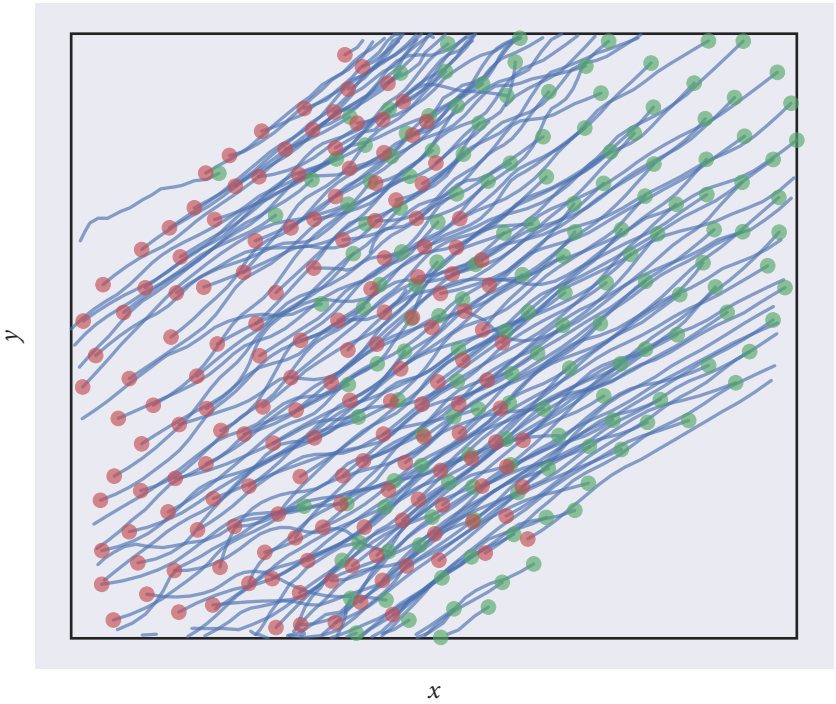



**Figure 7.15:** A visualisation of the trajectories of foraging surf scoters in an analysed sequence. In total, the sequence represents the movements of 199 scoters over 23 frames (4,577 data points). The scoters move along the blue trajectories, starting from the positions denoted by the green markers and finishing at the positions shown by the red markers. The black frame containing the trajectories represents an approximation of the fixed field of vision of the recording equipment.

data points were missing. In this visualisation blue markers represent observed data and red markers represent missing data. From Figure 7.16 we see that in *every frame* there is *at least one* missing scoter. As such, if we were to discard all the frames with as least one missing scoter (as in the naive approach), we would discard the *entire* flocking event. Therefore, to evaluate the likelihood of observing this flock we have no choice but to account for the missing observations.

The simulated flocks considered in Sections 7.1.1 and 7.1.2 had between 10 and 20 missing observations per sequence. With 680 data points missing from the scoter sequence with the *least* amount of missingness, this problem represents a considerable increase in complexity. With this amount of missingness, long and expensive simulations were necessary to realise a satisfactory number of samples from the posterior.

We fit a variation of the Vicsek model to this dataset where each agent interacts with neighbours within distance $r$ (Equation (4.4)), and experiences noise sampled from a generalised Student's $t$ distribution with $\nu$ degrees of freedom and scale $\sigma_Y$ (Equation (4.6)).

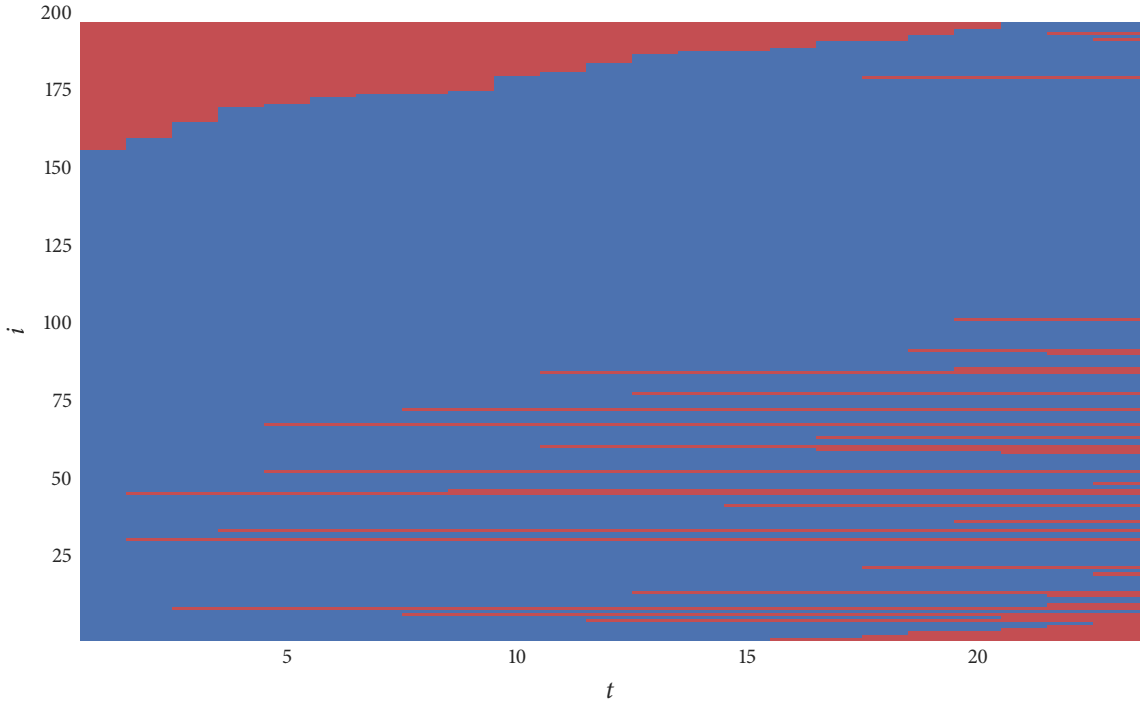



**Figure 7.16:** A representation of the missing and observed data points of the foraging event illustrated in Figure 7.15. The $x$-axis represents the frame of the sequence, and the $y$-axis corresponds to tracked individuals. A blue tile at location $(t, i)$ tells us that scoter $i$ was observed in frame $t$. A red tile at location $(t, i)$ indicates that scoter $i$ was missing at time $t$.

A random walk Metropolis–Hastings algorithm was implemented to infer the parameters of this model. In a similar manner, a random walk Metropolis–Hastings algorithm, with proposal scheme as outlined in Section 7.1.1, was used to infer the observations missing at the beginning of the sequence. Finally, using a Gibbs step allowed observations missing at the end of the sequence to be sampled. This sampler was implemented for $10^8$ iterations. To reduce the memory overhead of this sampler, the recorded samples were thinned by a factor of 1000; meaning that all but every 1000th iteration of the sampler was discarded, producing a total of $10^8/10^3 = 10^5$ samples.

Previously we used summary plots—such as those in Figures 7.5 and 7.11—to help visualise high-dimensional posterior distributions. However, with almost 700 dimensions, this posterior has too many dimensions for even these summary plots. In such a situation it can be informative to graphically assess the trace of the log-likelihood. If we can observe that the log-likelihood has converged, we have strong evidence that the sampler has converged. In Figure 7.17 we see a trace and histogram plot of evaluations of the log-likelihood. From this plot we see evidence of convergence, as the log-likelihood oscillates regularly around

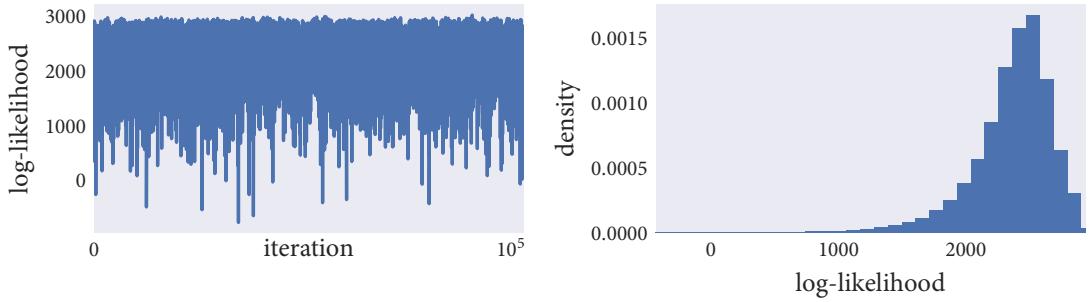



**Figure 7.17:** Trace and histogram plots of the log-likelihood, as computed at every iteration of the inference scheme. As it can be difficult to assess the convergence of all the parameters for high-dimensional problems, as we have here, it can instead be informative to assess the convergence of the log-likelihood. If we see that the value of the log-likelihood has converged, then we have evidence that the chains targeting the corresponding parameters also converged.

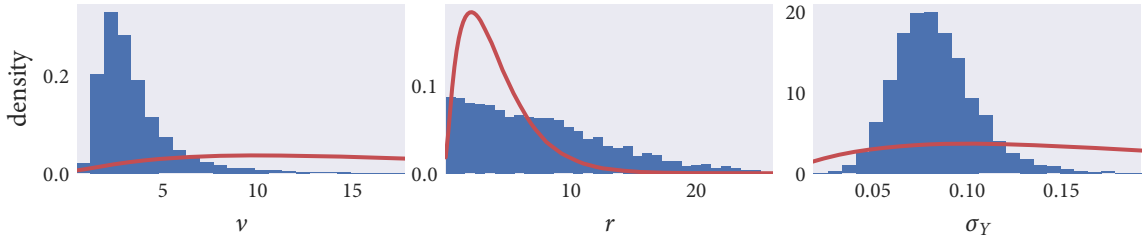



**Figure 7.18:** Samples drawn from the posterior distribution realised in fitting a sequence of the Lukeman data to a variation of the Vicsek model. The posterior distribution of $r$ is non-normal, and tends towards zero, representing evidence that there is no explicit alignment interaction between individuals.

some fixed location.

The parameters of the Vicsek model inferred in this fitting are shown in Figure 7.18. The tendency of the interaction radius $r$ towards zero suggests that there is no direct alignment interaction between individuals of this flock. The flock's highly polarised structure must then be the result of behaviours unaccounted for by our model: such as attraction or repulsion interactions. With no effective interaction term, all observed directional changes must be accounted for by the inferred noise parameters, as in the Null model (Section 4.2.1). The inferred degrees of freedom and noise-scale parameters represent a diffuse noise distribution, reflecting their need to capture all directional changes.

In Figure 7.19 we see histogram plots of samples realised from the posterior: here representing four directions missing at the end of the sequence. We see that the uncertainty in the possible directions of motion of the missing agent is large. This large posterior variance is a result of the large amounts of missingness. As no effective interaction term was realised,

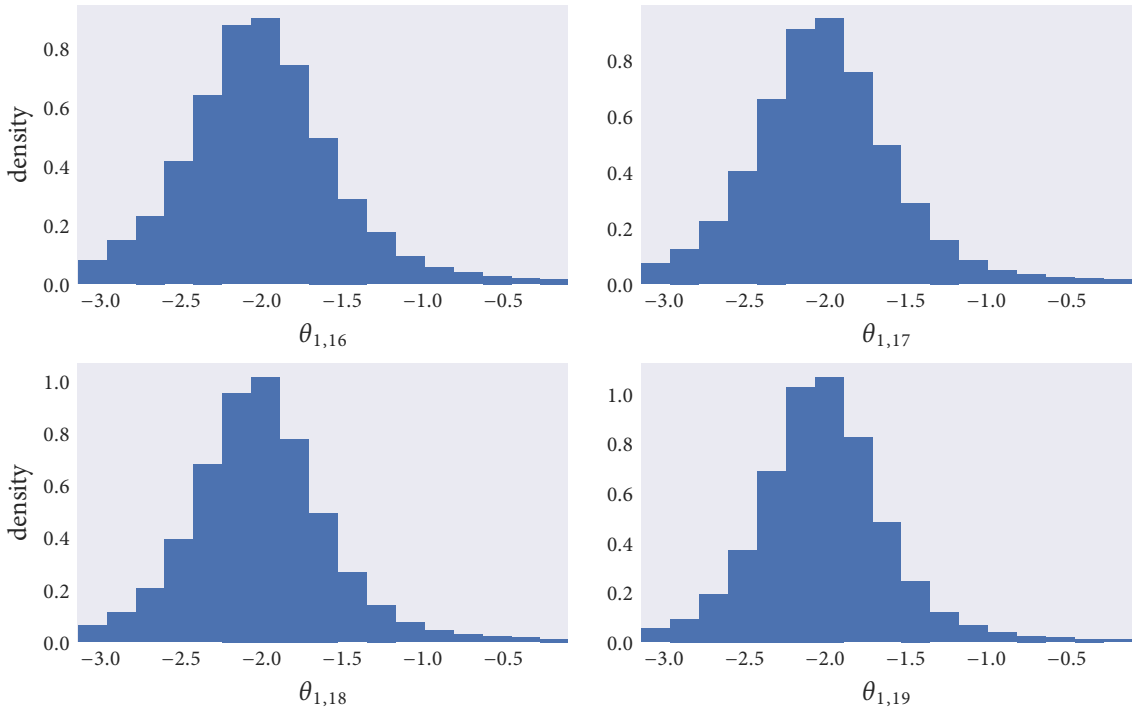



**Figure 7.19:** Histogram plots of posterior draws of 4 missing directions (out of a total of 680 missing data points). The posterior draws visualised correspond to data missing at the end of the sequence. We see large amounts of uncertainty in these posterior densities, reflecting the large amount of missingness represented by this sequence.

a diffuse noise structure was inferred to account for the observed directional changes. This diffuse noise structure further contributes to the large posterior variance inferred for the missing directions.

Although our model has been unable to capture the interactions between individuals present in this flock, it is only by handling the missing data that we have been able to fit this model and make this realisation.

## CONCLUSIONS

In this chapter we considered the experimental set-ups required to record flocking events in the wild. We saw missing data arises naturally as a consequence of recording data with fixed-location cameras. However, as individuals which move out of frame may still be influencing the movements of the observed flock, they cannot simply be ignored.

A naive approach to this missing data problem would be to discard *all* the frames in which *any single* agent moves out of frame. Although this approach allows us to evaluate the likelihood of the flock—and hence perform our desired fitting—it has the potential to

drastically reduce the amount of data available for analysis.

We outlined an alternative approach in which all possible out-of-frame movements are integrated over. For this approach we first partition missing data points into observations missing at the *beginning* of a recording event and observations missing at the *end* of a recording event. With this we outlined inference schemes which could infer data missing under these two regimes.

Simulation studies were performed in which an artificial fixed field of vision was overlain on the data. We demonstrated that our inference schemes were able to capture the true values of the missing observations. We discussed how increasing the amount of missing data increases the complexity, and hence computational demand, of the problem. Although all care should be taken to avoid recording events with missing observations, we advised that as data missing at the end of a sequence could be handled at less computational expense than data missing in the beginning, emphasis should be put on avoiding observations going missing at the beginning of a sequence.

To consider the efficacy of our approach to missing data we compared posterior densities generated when integrating over the missingness, to posterior densities on the same dataset without missingness, and to posterior densities realised with the missingness discarded (the naive approach). We found that our approach to missingness generated posterior densities closer to the "observed-all" densities than the naive approach was able to.

The developed methodology was then applied to a flocking event in which at least one individual was out of view during every frame of the recording. Of the 4,577 data points in the analysed sequence, 680 ($\approx$ 16%) of the observations were missing. To evaluate the likelihood of this data we had to be able to account for the position and direction of every individual in the analysed frames. As at least one individual is missing in every frame, discarding frames containing missingness is not a viable solution.

The large amount of missing data in the analysed sequence resulted in a high-dimensional posterior distribution, and hence a computationally intensive inference problem. Long simulations were necessary to allow the simulated Markov chains to achieve convergence. Although the inference scheme converged, our results showed evidence that there is no explicit alignment interaction between individuals, suggesting the presence of interactions not accounted for by our model. As our model couldn't capture these interactions, it instead attempted to account for directional changes through noise alone, which in turn resulted in large uncertainties about missing positions and directions of motion. Despite our candidate model being unable to capture the interactions between individuals, it was only by handling this missing data that we were able to come to this realisation.

# *Conclusions*

Mathematical modelling has become a cornerstone of the study of collective behaviour, as the formation of macro-level structure was realised possible by implementations of simple behavioural rules. Although model simulation can be informative in its own right, its real power comes in comparison with observation. Comparison of model prediction and real-world observation is an essential part of any model-fitting and model-verification process. However, until recently thorough comparison between collective behaviour model and data has been limited by the availability of data of real flocking events.

In this work we acknowledged the technical challenges presented by the recording of flocking events, and considered the affect of this on the resulting literature. A lack of data with which to perform model-verification has resulted in a plethora of competing models based on aprioristic assumptions about how individuals behave and respond to stimulus. However, more recently there have been notable attempts to bridge the gap between model and data. An overview of this recent work, and work prior to it, was presented in Chapter 2.

Backed by previous studies, we argued the importance of the quantification of uncertainty during model-fitting. Bayesian inference, representing a fully-probabilistic approach to parameter inference, is a natural approach to quantifying uncertainty. The underlying philosophy of the Bayesian paradigm was outlined in Chapter 3. Here we also introduced Markov chain Monte Carlo methods: a class of algorithms which allow a practitioner to draw samples from the posterior distribution, whilst only requiring knowledge of the posterior up to a constant of proportionality.

In fitting models to data it is advisable to first consider simple models, and to only increase model-complexity as necessary. The Vicsek model in Chapter 4 was introduced as one such simple model, accounting for directional changes as some combination of alignment interactions and noise. We questioned the biological-realism the zonal interaction rule implemented by Vicsek represents, instead advocating the merits of continuously differentiable interaction rules. As well as biological-realism, we acknowledged the difficulties which discontinuous interaction rules pose for parameter inference.

Despite understandings of the importance of biological variation in nature, the majority of models in the literature assume that all agents behave identically when presented with the same stimulus. With this shortcoming in mind we introduced models with hierarchical

structure: allowing intra-flock behavioural and biological variation.

Chapter 5 was a chapter of simulation studies. Simulation studies are a useful tool in assessing the viability of performing parameter inference on real data. The models introduced and developed in the preceding chapter were forward simulated for known parameter values. We then demonstrated that the Metropolis–Hastings and NUTS algorithms could be used to accurately infer the true parameter values from observations of the simulated flocks alone.

Having demonstrated the efficacy of our approach to inference on simulated data, we proceeded to re-use our statistical machinery to perform inference on real data. In Chapter 6 we introduced a dataset consisting of three recording events of flocking sheep. To this data we fit variations of the Vicsek model as introduced in Chapter 4. The within-sample predictive performance of these models was quantified with information criteria. Models were then ranked by their performance. We observed that continuous interaction rules consistently outranked their discrete counterpart, as well as observing evidence of non-normally distributed noise. In general the models provided a good fit to data, however in one of the sequences there was evidence of the existence of behaviours *not* accounted for in our models. We speculated that these behaviours could be the result of some unaccounted for interactions with the external environment, or additional inter-individual interactions such as attraction or repulsion.

As it is impossible to predict when and where a given flocking event may occur next, there can become a frustrating "right-place-right-time" component to data collection. Flocking events may be captured by establishing fixed-location recording equipment where a researcher believes a flocking event may occur in the future. However, the stationary nature of the recording equipment can result in capturing incomplete flocking events. To evaluate the likelihood of the data—a critical component of any model-fitting, Bayesian or otherwise—the observation of every individual in the flock, at every recorded frame, is required. A naive solution to this problem is to discard every frame in which the entire flock isn't visible. However, this approach has the potential to *drastically* reduce the amount of data available for analysis. In Chapter 7 we presented an alternative approach which incorporates the additional uncertainty introduced by missingness into our posterior distributions. Simulation studies were performed which showed that our approach consistently outperformed the naive approach of discarding data.

The methodology we developed in Chapter 7 was used on a real dataset exhibiting missingness. This dataset detailed the movements of ducks foraging on the surface of a lake, and was captured with fixed-location recording equipment. A large number of data

points were missing from this sequence, as much movement took place out-of-frame. This presented a computationally demanding, high-dimensional inference problem. In fitting a variation of Vicsek to this data, we observed evidence that there is no direct alignment interaction between individuals. We postulated the existence of repulsion or attraction behaviours, not accounted for by our model, which maintain the flocks highly polarised structure. Although our model could not explain the interactions between individuals, it was only by accounting for the missing observations that we were able to reach this conclusion.

We demonstrated to the reader that although fitting models of collective motion to real data is difficult, the process can be *greatly* eased by using existing, general purpose inference tools such as the Stan programming language. Such software will undoubtedly outperform the hand-coded MCMC algorithms of all but the most talented computer programmers. To the reader we also emphasise the utility of simulation studies. Simulation studies represent an invaluable opportunity to assess the feasibility of parameter inference on any model of interest, and give scope to develop *and* debug inference code in an environment where the answer is known. Although methodology was outlined which allowed the researcher to account for missing data within recordings of flocking events, the computational cost that this approach incurs is so large that in the first instance the greatest effort should be made to *avoid* missing data occurring during the recording process.

Throughout this thesis we demonstrated a fully-Bayesian approach to fitting models of collective behaviour to real and simulated data. We advocated consideration of simple models over the needlessly complex, and the addition of complexity only when simpler models perform poorly. Although this work only considered a small number of possible candidate models, the framework we outlined can be used to compare any number of competing models.

# 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.

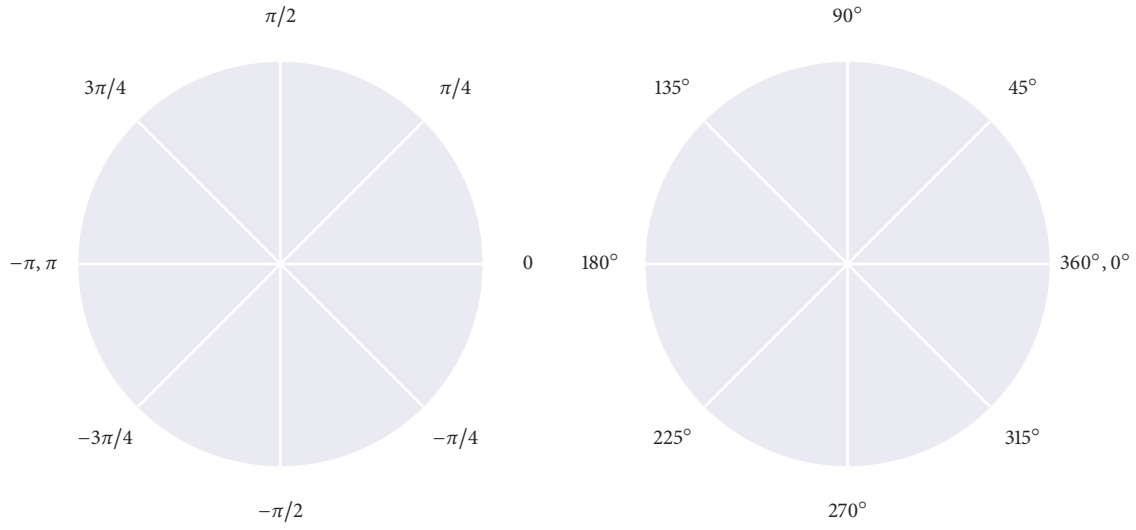
In this appendix we shall consider why standard techniques, methods and summaries are inappropriate to use with circular data. After this realisation, we proceed to introduce some useful techniques which can be used to handle and visualise directional data.

## 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 $\pi/180^\circ$.

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.

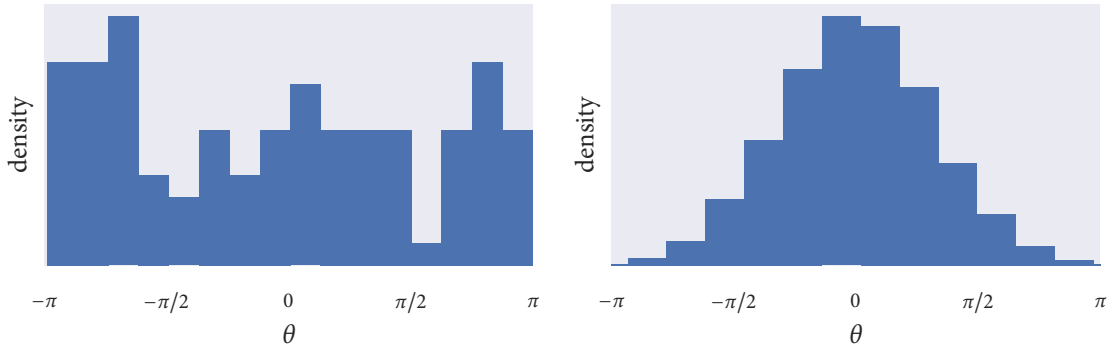



**Figure A.2:** Using histograms to visualise (a): one hundred samples drawn from $U(-\pi, \pi)$ and (b): ten thousand samples drawn from $N(0, 1)$.

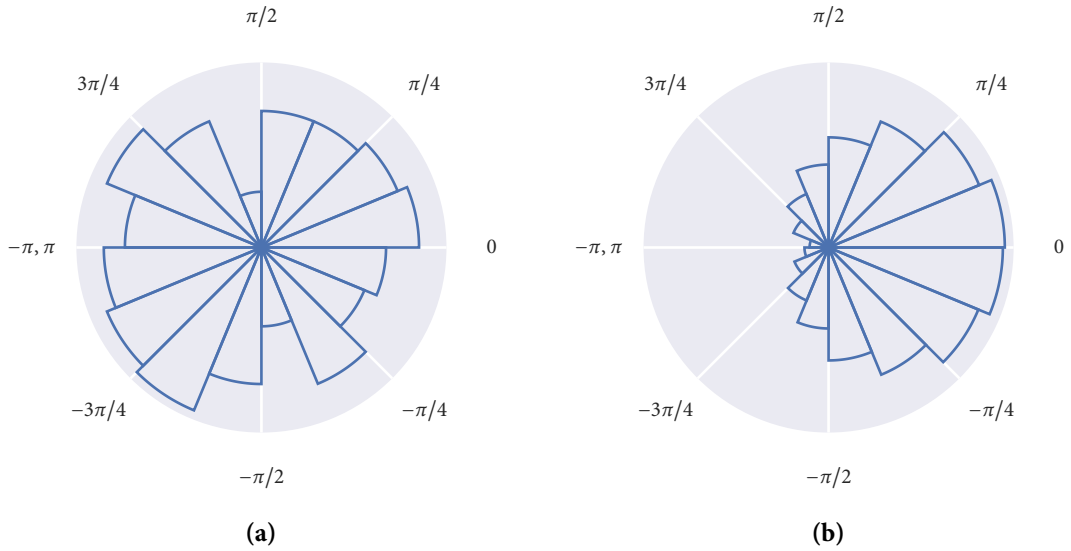


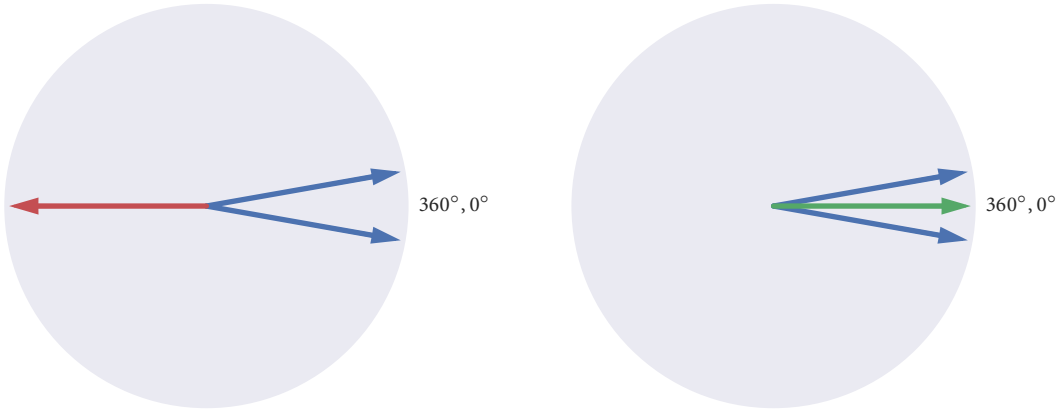

**Figure A.3:** Using polar histograms to visualise (a): one hundred samples drawn from $U(-\pi, \pi)$ and (b): ten thousand samples drawn from $N(0, 1)$.

## 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:

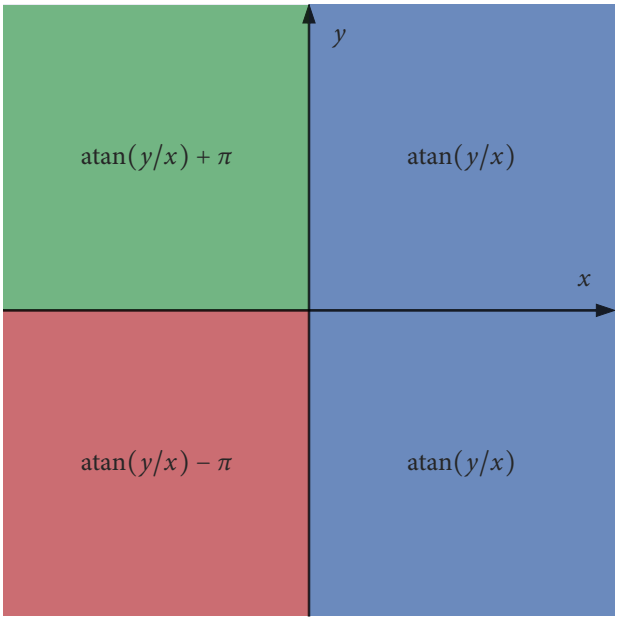



**Figure A.5:** An illustration of the quadrant corrections made by atan2.

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

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.

Camperi, Marcelo, Andrea Cavagna, Irene Giardina, Giorgio Parisi and Edmondo Silvestri (2012). 'Spatially balanced topological interaction grants optimal cohesion in flocking models'. *Interface focus* 2.6, 715–725.

Cavagna, Andrea, Irene Giardina, Alberto Orlandi, Giorgio Parisi, Andrea Procaccini, Massimiliano Viale and Vladimir Zdravković (2008). 'The STARFLAG handbook on collective animal behaviour: 1. Empirical methods'. *Animal Behaviour* 76.1, 217–236.

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

Couzin, Iain 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.

Couzin, Iain D., Jens Krause, Richard James, Graeme D. Ruxton and Nigel R. Franks (2002). 'Collective memory and spatial sorting in animal groups'. *Journal of Theoretical Biology* 218, 1–11.

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.

Fong, Edwin and CC Holmes (2020). 'On the marginal likelihood and cross-validation'. *Biometrika* 107.2, 489–496.

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

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.

Geman, Stuart and Donald Geman (1984). 'Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images'. *Ieee transactions on pattern analysis and machine intelligence* 6, 721–741.

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

Ginelli, Francesco (2016). 'The physics of the Vicsek model'. *The European journal of physics special topics* 225, 2099–2117.

Ginelli, Francesco and Hugues Chaté (2010). 'Relevance of metric-free interactions in flocking phenomena'. *Physical Review Letters* 105.16, 168103.

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.

Jhawar, Jitesh, Richard G. Morris, UR Amith-Kumar, M D. Raj, Tim Rogers, Harikrishnan Rajendran and Vishwesha Guttal (2020). 'Noise-induced schooling of fish'. *Nature physics* 16.4, 488–493.

Juárez, Miguel A. and Mark F. J. Steel (2010). 'Model-based clustering of non-gaussian panel data based on skew-t distributions'. *Journal of business & economic statistics* 28.1, 52–66.

Kalman, Rudolph E. (1960). 'A new approach to linear filtering and prediction problems'. *Transactions of the ASME–journal of basic engineering* 82.Series D, 35–45.

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.

Kuhn, Harold W. (1955). 'The Hungarian method for the assignment problem'. *Naval research logistics quarterly* 2.1-2, 83–97.

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.

Mann, Richard P., Andrea Perna, Daniel Strömbom, Roman Garnett, James E. Herbert-Read, David J. T. Sumpter and Ashley J. W. Ward (2013). 'Multi-scale inference of interaction rules in animal groups using Bayesian model selection'. *PLOS computational Biology* 9.3, e1002961.

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 and H. Chap. 5.

Newman, Mark E. (2005). 'Power laws, Pareto distributions and Zipf's law'. *Contemporary physics* 46.5, 323–351.

Nieder, Andreas (2005). 'Counting on neurons: the neurobiology of numerical competence'. *Nature reviews neuroscience* 6.3, 177–190.

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 H. 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*.

Strandburg-Peshkin, Ariana, Colin R. Twomey, Nikolai W. F. Bode, Albert B. Kao, Yael Katz, Christos C. Ioannou, Sara B. Rosenthal, Colin J. Torney, Hai S. Wu, Simon A. Levin and Ian D. Couzin (2013). 'Visual sensory networks and effective information transfer in animal groups'. *Current biology* 23.17, R709–R711.

Strömbom, Daniel (2011). 'Collective motion from local attraction'. *Journal of theoretical biology* 283.1, 145–151.

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

Thorup, Kasper, Thomas Alerstam, Mikael Hake and Nils Kjellén (2003). 'Bird orientation: compensation for wind drift in migrating raptors is age dependent'. *Proceedings of the Royal Society of London B* 270, S8–S11.

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.

Vicsek, Tamás and Anna Zafeiris (2012). 'Collective motion'. *Physics Reports* 517, 71–140.

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.